

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΥΠΟΛΟΓΙΣΤΩΝ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Σημειώσεις

Χαράλαμπος Π. Στρουθόπουλος
Καθηγητής

ΣΕΡΡΕΣ ΝΟΕΜΒΡΙΟΣ 2026

(1)88811114125

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1 3
ΕΙΣΑΓΩΓΗ 3
ΚΕΦΑΛΑΙΟ 2 6
ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΠΡΟΤΥΠΩΝ ΚΑΙ ΟΡΙΣΜΟΙ 6
2.2 Εσωτερικό γινόμενο 9
2.3 Αποστάσεις 10
ΚΕΦΑΛΑΙΟ 3 12
ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ – ΕΚΠΑΙΔΕΥΣΗ ΜΕ ΕΠΟΠΤΗ 12
3.1. Ταξινόμηση με βάση την απόσταση από τους Κ-γείτονες 12
3.1.1 Αναγνώριση με βάση τα κέντρα των τάξεων 13
3.2 Ταξινόμηση σε δύο κλάσεις με γραμμική διακριτική συνάρτηση, ο νευρώνας Perceptron. 14
3.2.1. Η περίπτωση πολλών κλάσεων. 19
3.2.2. Η περίπτωση της XOR, ταξινομητές πολλών επιπέδων 21
3.3. Πολυεπίπεδοι ταξινομητές – Διόρθωση σφάλματος με οπισθοδιάδοση (Back Error Propagation) 23
Εκπαίδευση με οπισθοδιάδοση του σφάλματος (back-error propagation) 27
3.4. Συνελικτικά Νευρωνικά Δίκτυα 36
Επίπεδο Συνέλιξης (Convolution) 38
Επίπεδο δειγματολειψίας (Pooling Layer) 41
Επίπεδο Ταξινόμησης 42
Υπολογισμός του πλήθους βαρών και εκπαίδευση στα ΣΝΔ 42
Εκπαίδευση των Συνελικτικών Νευρωνικών Δικτύων 43
Εκπαίδευση στο επίπεδο πλήρους διασύνδεσης 44
Διόρθωση σφάλματος με οπισθοδιάδοση στο επίπεδο δειγματοληψίας. 46
Διόρθωση σφάλματος με οπισθοδιάδοση στο επίπεδο συνέλιξης 46
3.4.1 Συνελικτικά νευρωνικά δίκτυα για σημασιολογική κατάτμηση εικόνων 49
3.6 Δένδρα απόφασης 77
ΚΕΦΑΛΑΙΟ 4 59
ΕΚΠΑΙΔΕΥΣΗ ΧΩΡΙΣ ΕΠΟΠΤΗ 59
4.1 Ο Αλγόριθμος ISODATA ή Κ-Μέσων (k-means ή c-means) 59
4.2 Απεικόνιση αλυσίδας κοντινών γειτόνων. 62
4.3 Χάρτης απεικόνισης χαρακτηριστικών με αυτό-οργάνωση 64
ΚΕΦΑΛΑΙΟ 5 69

ΑΝΑΛΥΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ 69
5.1 Ανάλυση χαρακτηριστικών στην εκπαίδευση με επόπτη 69
5.2 Ανάλυση χαρακτηριστικών στην εκπαίδευση χωρίς επόπτη 71
3.7 Ταξινομητές με βάση τον κανόνα πιθανοτήτων του Bayes 80
ΠΑΡΑΡΤΗΜΑ Α: ΒΑΣΙΚΕς ΠΡΑΞΕΙΣ ΓΡΑΜΜΙΚΗΣ ΑΛΓΕΒΡΑΣ 85
ΠΑΡΑΡΤΗΜΑ Γ: Στατιστικά χαρακτηριστικά του χώρου των προτύπων. 85
ΠΑΡΑΡΤΗΜΑ Γ:ΤΑ ΙΔΙΟΔΙΑΝΥΣΜΑΤΑ ΤΟΥ ΠΙΝΑΚΑ ΣΥΝΔΙΑΣΠΟΡΑΣ 86

888111114125

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η Υπολογιστική Νοημοσύνη είναι ένα υποσύνολο της τεχνητής νοημοσύνης (AI) που δομεί συστήματα ικανά να μαθαίνουν και να βελτιώνονται από την εμπειρία δηλαδή από τα δεδομένα εισόδου εξόδου, χωρίς να είναι ρητά δηλωμένος ο μηχανισμός που διέπει την σχέση τους. Για παράδειγμα μία λογική πύλη (π.χ. της διάζευξης, XOR) περιγράφεται ρητά με την άλγεβρα του *Boole*, εν τούτοις είναι δυνατόν μόνο από τα δεδομένα του πίνακα αληθείας της να δημιουργήσουμε ένα σύστημα Μηχανικής Μάθησης που θα προσομοιώνει την λειτουργία της. Αναφέρεται το σύστημα αυτό και ως εμπειρικό μοντέλο. Η αδυναμία ρητής περιγραφής ενός συστήματος και η ανάγκη επιστράτευσης της μηχανικής μάθησης μπορεί να προέρχεται είτε από την πολυπλοκότητά του είτε από την αδυναμία γλωσσικής περιγραφής. Για παράδειγμα δεν μπορούμε να διδάξουμε σε κάποιον χρησιμοποιώντας την φυσική γλώσσα πως να κάνει ποδήλατο, πρέπει να δοκιμάσει και να μάθει εμπειρικά.

Βασικές έννοιες και όροι που εμπλέκονται συχνά στην Υπολογιστική Νοημοσύνη είναι επιγραμματικά:

- Τα δεδομένα
- Οι αλγόριθμοι ταξινόμησης, παλινδρόμησης, ομαδοποίησης, ενισχυτικής μάθησης.
- Η Εκπαίδευση και ή δοκιμή
- Η Μάθηση, εποπτευόμενη, μη εποπτευόμενη και ενισχυτική
- Η Ανάλυση Χαρακτηριστικών

Εφαρμογές Μηχανικής Μάθησης:

Η Υπολογιστική Νοημοσύνη εφαρμόζεται σε διάφορους τομείς που αναφέρονται ενδεικτικά παρακάτω:

- Επεξεργασία Φυσικής Γλώσσας (NLP)
- Αναγνώριση εικόνας και ομιλίας
- Προβλεπτική Ανάλυση
- Συστήματα συστάσεων
- Υγεία και Ιατρική
- Χρηματοδότηση
- Αυτόνομα Οχήματα

Ένα σύνολο όμοιων προτύπων, όχι αναγκαστικά πανομοιότυπων, ορίζει μία σύλληψη (concept). Τα νοήμονα έμβια παρατηρούν και μαθαίνουν τα όντα (πρότυπα) του περιβάλλοντος κόσμου και δημιουργούν γι' αυτά συλλήψεις. Για παράδειγμα το σύνολο των αλόγων που γνωρίσατε σας έμαθε την γενική σύλληψη (ιδέα) άλογο (κάντε συσχετισμό με τον κόσμο των ιδεών του Πλάτωνα). Στον χώρο των γραμμάτων της Ελληνικής γλώσσας υπάρχουν 24 συλλήψεις που είναι τα 24 γράμματα της Ελληνικής αλφαβήτου. Κάθε γράμμα της Ελληνικής γλώσσας που γράφτηκε αποτελεί μία πραγματοποίηση ή συμβάν (event) μιας από τις 24 συλλήψεις. Για παράδειγμα οι μορφές:

β β **β** **β** **Β** β β **β** **Β**

δ ∆ **δ** **∆** **δ** δ ∂ **δ**

αποτελούν συμβάντα (πρότυπα) των συλλήψεων των γραμμάτων ΒΗΤΑ και ΔΕΛΤΑ. Τα συγκεκριμένα πρότυπα μπορεί να διαφέρουν στο μέγεθος, την γραμματοσειρά, να είναι κεφαλαία ή μικρά, ή να έχουν μικροδιαφορές που δημιουργήθηκαν κατά τη στιγμή της εκτύπωσης (θόρυβος). Αντικείμενα (πρότυπα) που ανήκουν στην ίδια σύλληψη αποτελούν μία *κλάση ή τάξη* (*class*). Είναι προφανές πως οι συλλήψεις και τα οικεία πρότυπά τους πρέπει να διαθέτουν χαρακτηριστικά που να τις διακρίνουν από άλλες συλλήψεις και πρότυπα. Η εύρεση και μέτρηση τέτοιων χαρακτηριστικών (features) των προτύπων είναι πρωταρχικής σημασίας για την περιγραφή και την αναγνώρισή τους.

Η ταξινόμηση των αντικειμένων σε κλάσεις λέγεται *αναγνώριση προτύπων* (*pattern recognition*) και μπορεί να αντικαταστήσει πληθώρα επίπονων ανθρώπινων εργασιών, να βελτιώσει το εργασιακό περιβάλλον και να αυξήσει την παραγωγικότητα. Για παράδειγμα η μετατροπή τυπωμένων εγγράφων σε ηλεκτρονική μορφή με κωδικοποίηση κατά ASCII, μπορεί να γίνει από τον άνθρωπο δακτυλογράφο που διαβάζει το κείμενο, αναγνωρίζει τους χαρακτήρες και τους πληκτρολογεί, μπορεί όμως ακόμη να γίνει με την σάρωση και ψηφιοποίηση (scanning) του εγγράφου από κατάλληλο πρόγραμμα οπτικής αναγνώρισης χαρακτήρων (*OCR: Optical Character Recognition*) που θα αναγνωρίζει την ψηφιακή μορφή κάθε χαρακτήρα και θα τον κωδικοποιεί με το αντίστοιχο κώδικα ASCII.

Ένα σύστημα (αλγόριθμος) που αναγνωρίζει πρότυπα λέγεται *ταξινομητής* (*classifier*). Σε ένα ταξινομητή επιτελούνται δύο κύριες εργασίες που είναι η *εκπαίδευση* (*training*) και η *ταξινόμηση*. Κατά την εκπαίδευση διαχωρίζονται κατάλληλα οι κλάσεις ή προσδιορίζονται οι συγκεντρώσεις των προτύπων και ρυθμίζονται οι παράμετροι του συστήματος ταξινόμησης με βάση τις οποίες θα είναι αυτό ικανό να ταξινομήσει τα πρότυπα σωστά στις κλάσεις. Με την ταξινόμηση αποδίδεται ένα πρότυπο σε μία κλάση ή σε μία συγκέντρωση. Οι μέθοδοι εκπαίδευσης χωρίζονται σε δύο βασικές κατηγορίες:

Α) Εκπαίδευση με επόπτη (supervised learning) κατά την οποία είναι γνωστή η κλάση στην οποία ανήκει κάθε πρότυπο του συνόλου εκπαίδευσης και επιδιώκεται η ορθή απόδοση των προτύπων που δεν ανήκουν στο σύνολο εκπαίδευσης, σε μία από τις προκαθορισμένες κλάσεις. Σε κάθε πρότυπο του συνόλου εκπαίδευσης αντιστοιχίζουμε το όνομα της κλάσης του με μία ετικέτα (label).

Β) Εκπαίδευση χωρίς επόπτη (unsupervised learning) κατά την οποία δεν είναι γνωστή η κλάση στην οποία ανήκει κάθε πρότυπο του συνόλου εκπαίδευσης και αναζητούνται βασικά οι συγκεντρώσεις των προτύπων (clustering).

Τα πρότυπα που χρησιμοποιούνται για την εκπαίδευση του ταξινομητή αποτελούν το *σύνολο εκπαίδευσης* (*training set*). Συχνά χρησιμοποιούμε και ένα σύνολο προτύπων για την επικύρωση των αποτελεσμάτων το οποίο ονομάζουμε *σύνολο επικύρωσης* (*validation set*).

Η εκπαίδευση ενός ταξινομητή επαναλαμβάνεται ώστε η μηχανή να προσαρμόζεται στις αλλαγές των προτύπων που ταξινομεί. Οι ταξινομητές είναι ουσιαστικά το αντικείμενο με το οποίο θα ασχοληθούμε περαιτέρω και θα αναλυθούν στα επόμενα κεφάλαια.

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΠΡΟΤΥΠΩΝ ΚΑΙ ΟΡΙΣΜΟΙ

Για κάθε πρότυπο μετρούμε ένα πεπερασμένο πλήθος χαρακτηριστικών του (π.χ. βάρος, χρώμα, ύψος, πλάτος, μήκος). Για ένα ποιοτικό χαρακτηριστικό που δεν μετριέται άμεσα με αριθμητικές τιμές όπως για παράδειγμα το χρώμα, χρησιμοποιούμε κάποια κωδικοποίηση που αντιστοιχεί αριθμούς στις διάφορες μορφές του (π.χ. 0 για το μαύρο, 1 για το μπλε, 2 για το κόκκινο κ.λ.π.) Αν Ω είναι ένα σύνολο που αποτελείται από Κ το πλήθος πρότυπα Π_κ, κ=1,2,3,...,Κ, δηλαδή $\Omega=\{\Pi_1,\Pi_2,...,\Pi_\kappa,...\Pi_K\}$ και μετρούμε Ν το πλήθος κοινά χαρακτηριστικά για κάθε πρότυπο, τότε το Π_κ πρότυπο μπορεί να περιγραφεί από ένα διάνυσμα (πίνακα στήλης) $\mathbf{x}_\kappa$ σε ένα χώρο Ν διαστάσεων σύμφωνα με την σχέση

$$\boldsymbol{x_\kappa} = \begin{vmatrix} x_{1\kappa} \\ x_{2\kappa} \\ \vdots \\ x_{\nu\kappa} \\ \vdots \\ x_{N\kappa} \end{vmatrix} = |x_{1\kappa} \quad x_{2\kappa} \quad \dots \quad x_{\nu\kappa} \quad \dots \quad x_{N\kappa}|^T \qquad (2.1.1)$$

Κάθε πρότυπο θεωρείται ξεχωριστό από κάθε άλλο ακόμη και όταν όλες οι τιμές των αντίστοιχων χαρακτηριστικών τους είναι ίσες. Δηλαδή το πρότυπο Π_μ είναι άλλο στοιχείο του συνόλου Ω, από το πρότυπο Π_λ αν και $\mathbf{x_\mu=x_\lambda}$ (μ, λ ∈ {1,2,...,Κ}, και Π_μ, $\Pi_\lambda \in \Omega$). Ως εκ τούτου το Ω είναι σύνολο με την μαθηματική έννοια. Όλους τους πίνακες θα τους συμβολίζουμε με παχιά γράμματα ή με υπογράμμιση.

Αν Κ το πλήθος των προτύπων, Ν το πλήθος των χαρακτηριστικών που μετράμε σε κάθε πρότυπο και Τ το πλήθος των τάξεων, μπορούμε να χρησιμοποιήσουμε ένα δείκτη κ=1,2,...,Κ για κάθε πρότυπο, ένα δείκτη ν=1,2,...,Ν για κάθε χαρακτηριστικό και ένα δείκτη τ=1,2,...,Τ για κάθε κλάση. Ο διανυσματικός χώρος των προτύπων έχει διάσταση Ν, (R^N). Για την τιμή του *ν-οστού* χαρακτηριστικού του *κ* προτύπου που ανήκει στην κλάση *τ* γράφουμε

$$x_{\nu\kappa}^{\tau}.$$

Συχνά κάποιοι δείκτες δεν χρησιμοποιούνται όταν έχουν προφανή τιμή ή όταν η αναφορά τους δεν είναι απαραίτητη. Η μαθηματική περιγραφή των πραγμάτων στην αναγνώριση προτύπων γίνεται με την χρήση όρων της γραμμικής άλγεβρας. Το διάνυσμα $\vec{x}$ που περιέχει Ν χαρακτηριστικά ενός προτύπου θα θεωρείται πίνακας με Ν γραμμές και μία στήλη και θα συμβολίζεται $\mathbf{x}_{Nx1}$ ή χάριν συντομίας **x** ή $\underline{x}$. Συχνά θα χρησιμοποιούμε τον ανάστροφο πίνακα $\mathbf{A}^T$ ενός πίνακα **A** (Παράρτημα Α). Για εξοικονόμηση χώρου στην σελίδα, τον πίνακα στήλης **x** ενός προτύπου θα τον γράφουμε ανεστραμμένο $\mathbf{x}^T=[\ x_1,\ x_2,\ ...,\ x_v,\ ...,x_N]$. Με βάση τον ορισμό του πολλαπλασιασμού δύο πινάκων (Παράρτημα Α) η Ευκλείδεια απόσταση $d_E(\mathbf{x},\mathbf{y})$ των πινάκων στήλης **x, y** δύο προτύπων Π_x, Π_u δίνεται από την σχέση

$$d_E(\mathbf{x},\mathbf{y}) = [(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})]^{1/2} \qquad (2.1.2)$$

Αν Ν≤3 τότε μπορούμε να παραστήσουμε τα διανύσματα των προτύπων στον φυσικό χώρο. Ακολούθως θα δούμε ένα παράδειγμα προτύπων που είναι άτομα για τα οποία μετράμε δύο χαρακτηριστικά που είναι το βάρος και το ύψος. Οι τιμές των χαρακτηριστικών για όλα τα πρότυπα που μετρήθηκαν δείχνονται στο ακόλουθο πίνακα 2.1.

Π (πρότυπο)	**Υ** (ύψος σε εκ.)	**Β** (βάρος σε κιλά)
Π1	185	85
Π2	185	82
Π3	182	115
Π4	180	80
Π5	178	78
Π6	175	112
Π7	171	114
Π8	160	58
Π9	160	55
Π10	155	62
Π11	150	85
Π12	150	82
Π13	145	84
Π14	143	80

Πίνακας 2.1

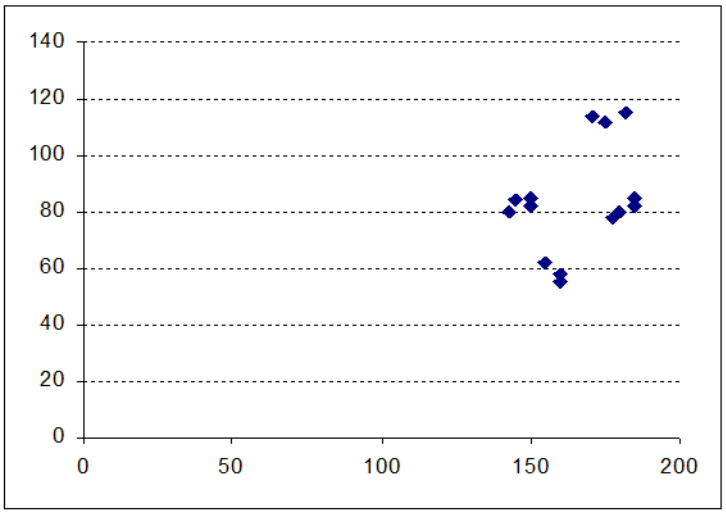
Τα αντίστοιχα διανύσματα των προτύπων γράφονται ως πίνακες στήλης όπως παρακάτω:

$$\boldsymbol{x}_1 = \begin{bmatrix}185\\85\end{bmatrix},\ \boldsymbol{x}_2 = \begin{bmatrix}185\\82\end{bmatrix},\ \boldsymbol{x}_3 = \begin{bmatrix}182\\115\end{bmatrix},\ \boldsymbol{x}_4 = \begin{bmatrix}180\\80\end{bmatrix},$$

$$\boldsymbol{x}_5 = \begin{bmatrix}178\\78\end{bmatrix},\ \boldsymbol{x}_6 = \begin{bmatrix}175\\112\end{bmatrix},\quad \boldsymbol{x}_7 = \begin{bmatrix}171\\114\end{bmatrix},\quad \boldsymbol{x}_8 = \begin{bmatrix}160\\58\end{bmatrix}, \qquad (2.1.3)$$

$$\boldsymbol{x}_9 = \begin{bmatrix}160\\55\end{bmatrix}, \ldots,\ \boldsymbol{x}_{14} = \begin{bmatrix}143\\80\end{bmatrix}$$

Στο Σχ.2.1 που ακολουθεί φαίνονται τα σημεία τέλους των διανυσμάτων που αντιστοιχούν σε κάθε πρότυπο.



Σχήμα.2.1.

Έστω ότι επιθυμούμε να κατατάξουμε τα πρότυπα μας σε τάξεις ως εξής: Βαρείς αν Β>80, ελαφρούς αν Β≤80, ψηλούς αν Υ>170 και κοντούς αν Υ≤170. Παρατηρούμε ότι τα άκρα των διανυσμάτων των προτύπων συγκεντρώνονται σε ομάδες. Το γεγονός αυτό οφείλεται στο ότι τα όμοια πρότυπα γειτνιάζουν επειδή τα αντίστοιχα χαρακτηριστικά τους έχουν κοντινότερες τιμές (ομόλογες συντεταγμένες των ανυσμάτων τους) σε σχέση με τις τιμές των χαρακτηριστικών των άλλων προτύπων. Για παράδειγμα η ευκλείδεια απόσταση D_{12} των προτύπων $Π_1$, $Π_2$ είναι

$$D_{12} = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2} = \sqrt{(185 - 185)^2 + (85 - 82)^2} = 3 \qquad (2.1.4)$$

και η ευκλείδεια απόσταση των προτύπων Π1, Π5 είναι

$$D_{15} = \sqrt{(x_{11} - x_{15})^2 + (x_{21} - x_{25})^2} = \sqrt{(185 - 178)^2 + (85 - 78)^2} = 9\sqrt{2} \qquad (2.1.5)$$

Άρα η απόσταση των ανυσμάτων των προτύπων είναι ένα μαθηματικό κριτήριο της ομοιότητα; δύο προτύπων. Οι ομάδες που σχηματίζουν τα πρότυπα λέγονται συγκεντρώσεις (clusters). Οι συγκεντρώσεις δημιουργούνται από πρότυπα με μορφολογική ομοιότητα. Στο παράδειγμα μας οι τέσσερις συγκεντρώσεις δημιουργήθηκαν από άτομα α) με Β>100 κ. β) με Β<70 κ. γ) με Υ<1.80μ και 70<Β<100 κ. και δ) με Υ>1.80μ και 70κ<Β<100κ.

Ο εντοπισμός των συγκεντρώσεων (clusters) προτύπων που δημιουργούνται στον χώρο των προτύπων είναι σημαντικός για την αναγνώριση προτύπων και αναφέρεται ως πρόβλημα εύρεσης των συγκεντρώσεων (clustering). Η σωστή επιλογή των χαρακτηριστικών που μετράμε για τα πρότυπα (feature selection) είναι καθοριστικής σημασίας για την εύρεση των ομάδων των προτύπων. Συχνά ενδιαφερόμαστε για πρότυπα που έχουν κάποια ιδιότητα και βρίσκονται σε περισσότερες από μία ομάδες. Για παράδειγμα τα πρότυπα των ατόμων με ύψος μεγαλύτερο από 1,70 μ ανήκουν σε δύο συγκεντρώσεις. Ανάλογα συμβαίνει και με τα πρότυπα των πεζών και κεφαλαίων γραμμάτων όπως των 'Α', 'α' ή 'Ω', 'ω' κ.α. Η εύρεση της ομάδας ή των ομάδων των προτύπων που έχουν μία τελική κοινή χαρακτηριστική ιδιότητα (ταυτότητα: label) αποτελεί την ταξινόμηση των προτύπων και οδηγεί στην αναγνώρισή τους. Στην γενική περίπτωση το πρόβλημα της ταξινόμησης είναι δύσκολο κυρίως λόγω α)κακής επιλογής των μετρούμενων χαρακτηριστικών, β) μεγάλης ποικιλομορφίας των προτύπων, γ)θορύβου, δ)της ασάφειας των ορίων των κλάσεων στον χώρο των προτύπων.

2.2 Εσωτερικό γινόμενο

Για δύο πίνακες $\mathbf{x} = [x_1, x_2, \ldots, x_\nu, \ldots, x_N]^T$, και $\mathbf{y} = [y_1, y_2, \ldots, y_\nu, \ldots, y_N]^T$ το εσωτερικό γινόμενό τους ορίζεται από τη σχέση

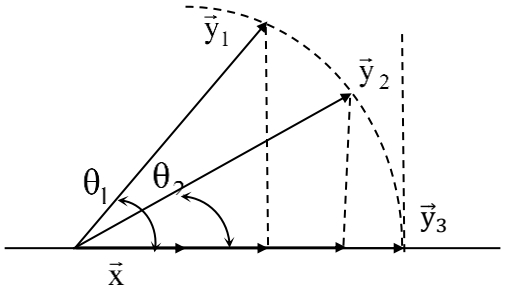
$$\mathbf{x}^T \cdot \mathbf{y} = \sum_{\nu=1}^{N} x_\nu \cdot y_\nu = x_1 \cdot y_1 + x_2 \cdot y_2 + \cdots + x_\nu \cdot y_\nu + \cdots x_N \cdot y_N = |\mathbf{x}||\mathbf{y}|\cos(\theta) \quad (2.2.1)$$

Αν N=2 ή N=3, θ είναι η κυρτή γωνία που σχηματίζουν τα δύο διανύσματα με την γεωμετρική έννοια. Αν N>3, $\theta = \cos^{-1}\frac{\mathbf{x}^T \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|}$ με $|\mathbf{x}| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}}$ το μέτρο του $\mathbf{x}$.

Γεωμετρικά αν το διάνυσμα $\vec{x}$ έχει μέτρο (μήκος) ίσο με τη μονάδα, το γινόμενο $\vec{x} \cdot \vec{y}$ είναι η αλγεβρική τιμή της προβολής του $\vec{y}$ επάνω στην ευθεία που διέρχεται από το διάνυσμα $\vec{x}$ (φορέας του $\vec{x}$) και λέγεται προσημασμένη απόσταση (Σχ.2.2). Όταν η γωνία θ των $\vec{x}$ και $\vec{y}$ μικραίνει το εσωτερικό τους γινόμενο $\vec{x} \cdot \vec{y}$ αυξάνει και μεγιστοποιείται όταν τα $\vec{x}$ και $\vec{y}$ βρίσκονται στην ίδια ευθεία. Αυτό σημαίνει ότι τότε ισχύει η σχέση

$$\vec{y} = \lambda \cdot \vec{x}, \quad \mathbf{y} = \lambda \cdot \boldsymbol{x} \text{ όπου } \lambda \in R \text{ και } \frac{y_1}{x_1} = \frac{y_2}{x_2} = \cdots \frac{y_\nu}{x_\nu} = \cdots \frac{y_N}{x_N} = \lambda \quad (2.2.2)$$

Η σχέση (2.2.2) είναι σχέση αναλογίας και το εσωτερικό γινόμενο μπορεί να χρησιμοποιηθεί ως κριτήριο αναλογίας (όχι υποχρεωτικά ισότητας) δύο προτύπων.



Σχήμα 2.2.

2.3 Αποστάσεις

Εκτός από την Ευκλείδεια απόσταση υπάρχουν και άλλοι τύποι αποστάσεων σε διανυσματικούς χώρους. Αν $d(\vec{x},\vec{y})$ είναι απόσταση θα πρέπει για οποιοδήποτε διάνυσμα **x, y, z** να ικανοποιούνται οι σχέσεις:

$d(\vec{x},\vec{y}) = d(\vec{y},\vec{x})$

$d(\vec{x},\vec{y}) \leq d(\vec{x},\vec{z}) + d(\vec{z},\vec{y})$

$d(\vec{x},\vec{y}) \geq 0$

Αν $\mathrm{d}(\vec{x},\vec{y}) = 0 \Leftrightarrow \vec{x} = \vec{y}$ (2.3.4)

Συνήθεις τύποι αποστάσεων είναι οι παρακάτω:

α) Minkowski τάξης s

$$d_\mu(\vec{x},\vec{y}) = [\sum_{\nu=1}^{N} |x_\nu - y_\nu|^s]^{1/s} \quad (2.3.5)$$

β) City Block

Είναι ειδική περίπτωση της Minkowski με s=1

$$d_c(\vec{x},\vec{y}) = \sum_{\nu=1}^{N} |x_\nu - y_\nu| \quad (2.3.6)$$

γ) Ευκλείδεια

Είναι ειδική περίπτωση της Minkowski για s=2

$$d_\varepsilon(\vec{x},\vec{y}) = \left[\sum_{\nu=1}^{N} (x_\nu - y_\nu)^2\right]^{1/2} = [(\boldsymbol{x}-\boldsymbol{y})^T \cdot (\boldsymbol{x}-\boldsymbol{y})]^{1/2} \quad (2.3.7)$$

δ) Chebychev

$$d_T(\vec{x},\vec{y}) = \max(|x_\nu - y_\nu|) \quad (2.3.8)$$

ε) Mahalanobis

$$d_R(\vec{x},\vec{y}) = d_R(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^T \cdot \mathrm{Cov}^{-1} \cdot (\boldsymbol{x}-\boldsymbol{y}) \quad (2.3.9)$$

όπου Cov ο πίνακας συμμεταβλητότητάς των **x** και **y**.

στ) Μη γραμμική (Non Linear)

$$d_{NL}(\vec{x},\vec{y}) = \begin{cases} 0 \text{ αν } d(x,y) \leq T \\ H \text{ αν } d(x,y) > T \end{cases} \qquad (2.3.10)$$

όπου $H,T \in R$ παράμετροι της απόστασης και $d(\vec{x},\vec{y})$ μια άλλη απόσταση του χώρου.

ΚΕΦΑΛΑΙΟ 3

ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ – ΕΚΠΑΙΔΕΥΣΗ ΜΕ ΕΠΟΠΤΗ

3.1. Ταξινόμηση με βάση την απόσταση από τους Κ-γείτονες

Μία απλή και μέθοδος ταξινόμησης ενός προτύπου **x** βασίζεται στην εύρεση του κοντινότερου γείτονά του που ανήκει στο αρχικό σύνολο προτύπων γνωστής κλάσης. Σύμφωνα με την μέθοδο αν $\mathbf{x}_r^c$ είναι ο γείτονας αυτός και ανήκει στην κλάση C, τότε το **x** ταξινομείται στην κλάση C. Η μέθοδος ισχυροποιείται αν λάβουμε υπόψη περισσότερους από έναν γείτονες σε περιττό κατά προτίμηση αριθμό Κ και αποδώσουμε στο **x** την κλάση της πλειοψηφίας αυτών.

```
% K nearest neighbor's classification example in Matlab R2018a
load fisheriris
x = [4.8 3.5 1.5 0.2];
[idx D] = knnsearch(meas,x,'k',3);
species(idx)
D
% Result Presentation
plot(meas(  1: 50,1),meas(  1: 50,2),'ro');hold on;
plot(meas( 51:100,1),meas( 51:100,2),'go')
plot(meas(101:150,1),meas(101:150,2),'bo')
plot(x(1),x(2),'kx')
plot(meas( idx,1),meas(idx,2),'k+')
```

Ο αλγόριθμος των *K* πλησιέστερων γειτόνων (*KNN*: K-Nearest Neighbors) είναι ακριβής και απλός στην υλοποίηση του. Σημαντικά μειονεκτήματά του είναι η διατήρηση στην μνήμη των προτύπων του συνόλου εκπαίδευσης και ο υπολογισμός των αποστάσεων του υπό ταξινόμηση προτύπου από αυτά του συνόλου εκπαίδευσης.

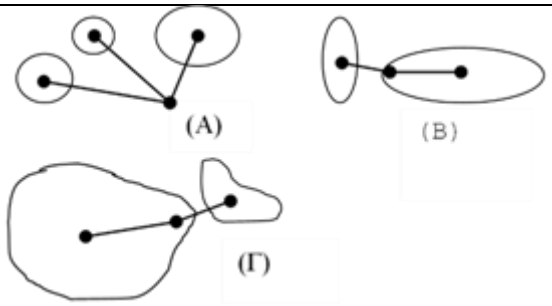
Υπάρχουν προτάσεις για την βελτίωση των δύο αυτών μειονεκτημάτων. Ειδικά για τον υπολογισμό της εύρεσης των κοντινότερων γειτόνων επιστρατεύεται ο αλγόριθμος *k-d tree* .

3.1.1 Αναγνώριση με βάση τα κέντρα των τάξεων

Είναι μια απλοϊκή μέθοδος αναγνώρισης με επόπτη σύμφωνα με την οποία κατά την εκμάθηση υπολογίζεται η μέση τιμή της κάθε κλάσης (κέντρο της κλάσης). Κάθε νέο πρότυπο ταξινομείται στην κλάση που το κέντρο της απέχει λιγότερο από το πρότυπο. Αν T το πλήθος των τάξεων με κέντρα μ_t, t=1,2,…,T το πρότυπο με πίνακα $\mathbf{x}$ αποδίδεται στην κλάση $\min_t(d(\boldsymbol{x},\boldsymbol{\mu}_t))$. Η ευκλείδεια απόσταση $d_E(\mathbf{x},\boldsymbol{\mu}_i)^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)=\mathbf{x}^T\mathbf{x}-2\mathbf{x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu}$. Για την σύγκριση των αποστάσεων του $\mathbf{x}$ από τις κλάσεις *i* και *j* αρκεί ο υπολογισμός της διαφοράς τους

$$d_{ij}=d_E(\boldsymbol{x},\boldsymbol{\mu}_i)^2-d_E(\boldsymbol{x},\boldsymbol{\mu}_j)^2 = 2\boldsymbol{x}^T(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i)+\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i-\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j . \quad (\alpha)$$

Έτσι η σύγκριση απαιτεί κάθε φορά κυρίως τον υπολογισμό του εσωτερικού γινομένου $\mathbf{x}^T(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i)$. Αν $d_{ij}=0$ τότε η σχέση (α) περιγράφει το υπερ-επίπεδο που είναι μεσοκάθετο στο ευθύγραμμο τμήμα με άκρα τα κέντρα των κλάσεων. Η μέθοδος αυτή μπορεί να είναι αποδοτική όταν οι κλάσεις έχουν τη μορφή υπερ-σφαιρών με ακτίνες μικρότερες της ημιαπόστασης των κέντρων τους (Σχ.3.2).



Σχήμα 3.1.1.

Α) Ορθή ταξινόμηση

Β) Εσφαλμένη Ταξινόμηση λόγω της συμμεταβλητότητας, προτείνεται η χρήση της απόστασης Mahalanobis

Γ) Εσφαλμένη ταξισνόμηση λόγω της μορφής των κλάσεων

3.2 Ταξινόμηση σε δύο κλάσεις με γραμμική διακριτική συνάρτηση, ο νευρώνας Perceptron.

Ο υπολογισμός των αποστάσεων είναι υπολογιστικά δαπανηρός και στην περίπτωση του KNN μνημοβόρος. Μία άλλη προσέγγιση υπολογιστικά ταχύτερη με μικρότερες απαιτήσεις μνήμης για την διαδικασία ταξινόμησης μπορεί να βασισθεί στον διαχωρισμό των κλάσεων με γραμμικές διακριτικές συναρτήσεις. Αν x το διάνυσμα στήλης που περιγράφει ένα πρότυπο και w διάνυσμα παραμέτρων, $\boldsymbol{x}, \boldsymbol{w} \in \mathcal{R}^N, w_0 \in \mathcal{R}$ με N το πλήθος των χαρακτηριστικών, η σχέση

$$d(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

αποτελεί μία γραμμική διακριτική συνάρτηση. Αν d(**x**)=0 τότε **x** βρίσκεται σε μία ευθεία αν N=2, σε ένα επίπεδο εάν N=3 ή ένα υπερεπίπεδο για N>3. Η d(**x**)=0 χωρίζει τον $\mathcal{R}^N$ σε δύο μέρη. Το H^+={**x**/**x**, d(**x**)>0} που περιλαμβάνει όλα τα **x** που καθιστούν την d(**x**)>0 και το H^- που περιλαμβάνει όλα τα **x** που καθιστούν την d(**x**)<0. Αν το άνυσμα **w** των παραμέτρων έχει κατάλληλες τιμές έτσι ώστε τα πρότυπα της μίας κλάσης να βρίσκονται στον ένα μέρος (πχ H^+) και της άλλης στο άλλο, τότε ένα άγνωστο πρότυπο **x'** ταξινομείται με κριτήριο την τιμή d(**x'**), αν δηλαδή d(**x'**)>0 ή d(**x'**)<0.

Ακολούθως θα παρουσιάσουμε τα παραπάνω και ακολούθως τον αλγόριθμο εύρεσης των **w** και w_0 για N=2 ώστε να είναι δυνατή η σχηματική και γεωμετρική αναπαράσταση στο επίπεδο. Αυτό δεν καταργεί την γενικότητα για N>2 και βοηθά στην κατανόηση και θεώρηση του πράγματος γεωμετρικά.

Η ευθεία (ε) στο Σχ.3.2.1 χωρίζει το επίπεδο σε δύο ημιεπίπεδα κάθε ένα των οποίων περιέχει και μια κλάση. Κάθε σημείο της ευθείας είναι το πέρας ενός διανύσματος $\vec{\boldsymbol{x}}$ (x_1, x_2) που ικανοποιεί την εξίσωση της ευθείας που είναι

$$w_1 \cdot x_1 + w_2 \cdot x_2 + w_0 = 0 \qquad (3.2.1)$$

ή

$$d(\mathbf{x}) = 0 \text{ όπου}$$

$$d(\mathbf{x}) = w_1 \cdot x_1 + w_2 \cdot x_2 + w_0 \ = \mathbf{w}^T \mathbf{x} + w_0 \quad (3.2.2)$$

με $\mathbf{x} = [x_1, x_2]^T$ και $\mathbf{w} = [w_1, w_2]^T$

Τα διανύσματα είναι ελεύθερα και ανήκουν σε κλάσεις ισοδυναμίας. Αν $\mathbf{x}_A$ και $\mathbf{x}_B$ ανήκουν στην ευθεία θα ισχύουν οι σχέσεις

$\mathbf{w}^T\mathbf{x}_A + w_0 = 0$ και (3.2.3)

$\mathbf{w}^T\mathbf{x}_B + w_0 = 0$ (3.2.4)

από όπου με αφαίρεση κατά μέλη προκύπτει ότι

$\mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$ (3.2.5)

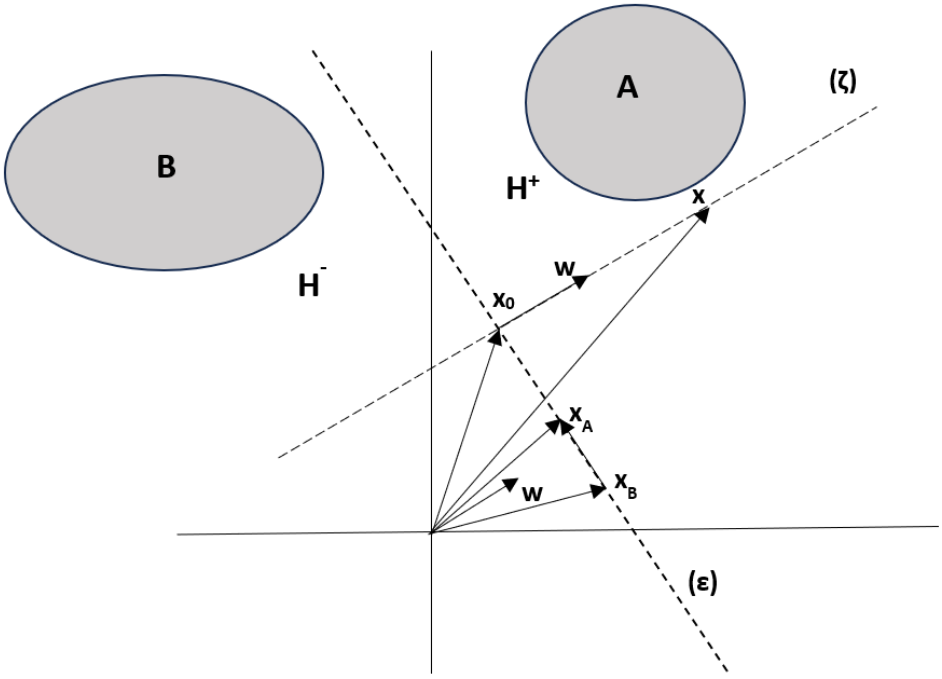
Η σχέση αυτή δείχνει ότι το άνυσμα **w** είναι κάθετο στην ευθεία (ή το επίπεδο αν το **x** ανήκει στον τρισδιάστατο χώρο). Όλα τα σημεία του ενός ημιεπιπέδου καθιστούν την d(**x**)>0 και του άλλου ημιεπιπέδου την d(**x**)<0. Πράγματι, αν $\mathbf{x}_0$ το σημείο που τέμνει η κάθετος (ζ) που άγεται από το σημείο **x** στην ευθεία (ε) θα ισχύει:

$$d(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0 + \mathbf{x}_0) + w_0 = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) + \mathbf{w}^T\mathbf{x}_0 + w_0 = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \Rightarrow$$

$$d(\mathbf{x}) = |\mathbf{w}||\mathbf{x} - \mathbf{x}_0|\cos(\theta),$$

αν $\mathbf{x}-\mathbf{x_0}$ είναι ομόρροπο του $\mathbf{w}$ τότε θ=0, cos(θ)=+1 και d(**x**)>0,

αν $\mathbf{x}-\mathbf{x_0}$ είναι αντίρροπο του $\mathbf{w}$ τότε θ=π, cos(θ)=-1 και d(**x**)>0



Σχήμα 3.2.1

Συνεπώς αν d(**x**)>0 το πέρας του **x** βρίσκεται στο ημιεπίπεδο που δείχνει το άνυσμα **w** (H^+) ενώ αν d(**x**)<0 το πέρας του **x** βρίσκεται στο άλλο ημιεπίπεδο (H^-).

Είναι επίσης προφανές πως η ποσότητα $\frac{\mathrm{d}(\mathbf{x})}{|\mathbf{w}|}$ είναι η προσημασμένη απόσταση του x από την ευθεία (ε).

Ζητούμε συνεπώς να βρούμε τα κατάλληλα **w** και w_0 για τα δοθέντα γνωστά **x** που ανήκουν στις δύο κλάσεις. Ένας τρόπος να γίνει αυτό είναι μία επαναληπτική διαδικασία διόρθωσης σφάλματος που θα περιγράψουμε ακολούθως και λέγεται εκπαίδευση του ταξινομητή.

Για την απλούστερη μαθηματική διατύπωση του μηχανισμού διόρθωσης θα θεωρήσουμε τους επαυξημένους πίνακες των προτύπων. Η γραμμική διακριτική συνάρτηση θα γράφεται ως ακολούθως

$$\mathrm{d}(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + w_0 = [\mathbf{w}^{\mathbf{T}}, w_0] \cdot \begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix}$$

$$\mathrm{d}(\mathbf{x}) = \tilde{\mathbf{w}}^{\mathrm{T}} \cdot \tilde{\mathbf{x}}$$

$$\tilde{\mathbf{w}} = \begin{bmatrix}\mathbf{w}\\ w_0\end{bmatrix}, \qquad \tilde{\mathbf{x}} = \begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix}$$

Αν δηλαδή ο χώρος των πινάκων των προτύπων έχει N διαστάσεις, οι σχέσεις θα γραφούν στον χώρο R^{N+1} που έχει N+1 διαστάσεις.

Έστω ότι ο επόπτης επιθυμεί d($\mathbf{x}^A$)>0 και d($\mathbf{x}^B$)<0. Αρχικά στο **w** αποδίδονται τυχαίες τιμές και πιθανότατα η d(**x**) δεν διαχωρίζει τις κλάσεις. Η βασική ιδέα

της εκμάθησης είναι η διόρθωση του υπερεπιπέδου $\tilde{\mathbf{w}}^{\mathrm{T}} \cdot \tilde{\mathbf{x}} = 0$ κάθε φορά που κάποιο πρότυπο ταξινομείται εσφαλμένα ώστε τελικά να αφήνει το πρότυπο προς την πλευρά που βρίσκεται η κλάση του. Η διόρθωση αυτή γίνεται σύμφωνα με την σχέση

$$\tilde{\mathbf{w}}(\mathrm{t}+1) = \tilde{\mathbf{w}}(\mathrm{t}) \pm \rho \cdot \tilde{\mathbf{x}} \qquad (3.2.12)$$

Όπου t ένας μετρητής επανάληψης, t=0,1,2… .

Το πρόσημο (+) χρησιμοποιείται όταν ταξινομείται λάθος ένα πρότυπο της κλάσης Α και (-) της κλάσης Β. Η ποσότητα ρ λέγεται παράμετρος ή ρυθμός εκμάθησης και παίρνει μικρές τιμές στο διάστημα (0,1), π.χ. ρ=0.2.

Ο μηχανισμός βασίζεται στο γεγονός ότι η ποσότητα $\rho \cdot \mathbf{x}^{\mathrm{T}}\mathbf{x}$ είναι πάντα θετική

$$\mathrm{d}_{\mathrm{t}+1}(\mathbf{x}) = \tilde{\mathbf{w}}(\mathrm{t}+1)^{\mathrm{T}} \cdot \tilde{\mathbf{x}} = (\tilde{\mathbf{w}}(\mathrm{t}) \pm \rho \cdot \tilde{\mathbf{x}})^{\mathrm{T}} \cdot \tilde{\mathbf{x}} = \mathrm{d}_{\mathrm{t}}(\mathbf{x}) \pm \rho \cdot \tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{x}) \qquad (3.2.12)$$

η διόρθωση μεταβάλει την εσφαλμένη απόσταση του προτύπου προς την πλευρά του ημιεπιπέδου της κλάσης και οδηγεί επαναληπτικά στο επιθυμητό αποτέλεσμα. Η παράμετρος ρ ρυθμίζει το βήμα διόρθωσης και την ταχύτητα σύγκλισης (Σχ. 3.2.2).

Τα παραπάνω παρουσιάζονται ακολούθως βηματικά.

Βήμα 1ο: Αρχικοποιήσεις:

Ορίζουμε ένα μετρητή επανάληψης t=0,1,2,3,... και του αποδίδουμε αρχικά την τιμή μηδέν. Αποδίδουμε τυχαίες (συνήθως μικρές θετικές) τιμές στο άνυσμα $\tilde{\boldsymbol{w}}(0) = [w_1, w_2, \dots, w_{N,} w_0]$, δημιουργούμε τα επαυξημένα διανύσματα του συνόλου εκπαίδευσης $\tilde{\boldsymbol{x}} = [\boldsymbol{x}, 1]$ και θεωρούμε την γραμμική διακριτική συνάρτηση $\mathrm{D_t}(\tilde{\mathbf{x}}) = \mathbf{w}^{\mathrm{T}}(\mathrm{t}) \cdot \tilde{\boldsymbol{x}}$. Ζητούμε κατά σύμβαση $\mathrm{D_t}(\tilde{\boldsymbol{x}}^A) > 0$ και $\mathrm{D_t}(\tilde{\boldsymbol{x}}^B) < 0$ (μπορούμε και το αντίθετο).

Βήμα 2ο: Αυξάνουμε τον μετρητή επανάληψης και επιλέγουμε ένα τυχαίο πρότυπο με πίνακα $\tilde{\boldsymbol{x}}$, από το σύνολο εκπαίδευσης και υπολογίζουμε την ποσότητα $\mathrm{D_t}(\tilde{\mathbf{x}})$. Αν το πρότυπο είναι σωστά ταξινομημένο επαναλαμβάνουμε το Βήμα 2 αλλιώς διορθώνουμε τα βάρη σύμφωνα με την σχέση

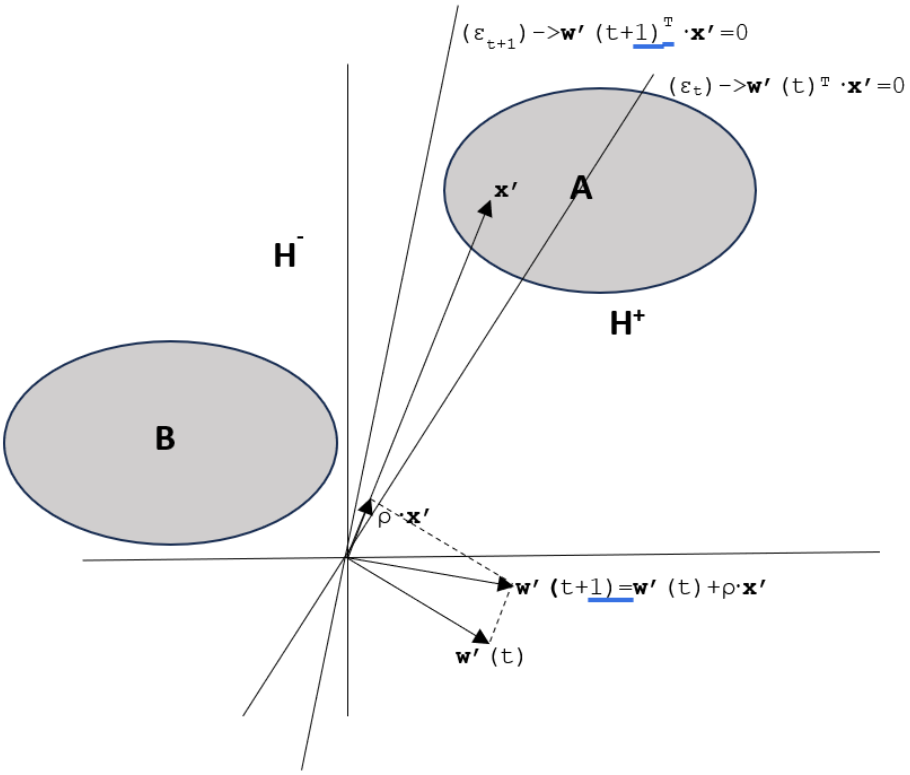
$$\tilde{\mathbf{w}}(\mathrm{t}+1) = \tilde{\mathbf{w}}(\mathrm{t}) \pm \rho \cdot \tilde{\mathbf{x}} \qquad (3.2.13)$$

Ή χωρίς την χρήση επαυξημένων διανυσμάτων

$$\mathbf{w}(\mathrm{t}+1) = \mathbf{w}(\mathrm{t}) \pm \rho \cdot \mathbf{x} \quad \text{και} \quad \mathrm{w_0}(\mathrm{t}+1) = \mathrm{w_0}(\mathrm{t}) \pm \rho$$

Με (+) αν το πρότυπο ανήκει στην κλάση Α, με (-) στην κλάση Β.

Βήμα 3ο: Αυξάνεται η τιμή του t κατά ένα και η διαδικασία επαναλαμβάνεται από το Βήμα 2 εωσότου όλα τα πρότυπα να ταξινομηθούν σωστά.



Σχήμα 3.2.2

(Το σύμβολο (΄) έχει αντικαταστήσει το σύμβολο (~) στα **x** και **w)**

Στον αλγόριθμο το άνυσμα **x** επιλέγεται τυχαία κάθε φορά και γι' αυτό θεωρείται στοχαστικός (stochastic). Αποδεικνύεται ότι για δύο γραμμικά διαχωρίσιμες κλάσεις ο αλγόριθμος οδηγεί πάντοτε σε λύση για κατάλληλα μικρή τιμή του ρ.

Υλοποίηση με MatLab.

```
function [ weights ] =...
perceptron_stochastic( trset,class_labels,...% +1 or -1
                       learning_rate,error_tolerance )
 ftrnum=size(trset,1); % feature number
 N=size(trset,2);      % vector number

 w=rand(ftrnum,1); w0=rand(1); % Weights initilization
 while sum( sign(w'*trset+w0)==class_labels ) < N-error_tolerance
    i=randi(N)
    x = trset(:,i);
    d = w'*x + w0;
    if( d*class_labels(i) < 0);
       w = w + class_labels(i)*learning_rate*x;
       w0= w0+ learning_rate*class_labels(i);
    end
 end
    weights = [w; w0];
end
```

Σε παρόμοιο αποτέλεσμα οδηγούμαστε εάν θεωρήσουμε το πρόβλημα εύρεσης των βαρών ως πρόβλημα βελτιστοποίησης (*optimization*) μιας συνάρτησης κόστους $K(\tilde{w})$. Μία τέτοια συνάρτηση είναι η

$$K(\tilde{\mathbf{w}}): R^{N+1} \to R^{+} \text{ με } \quad K(\tilde{\mathbf{w}}) = \sum_{x \in S} \delta(\tilde{\mathbf{x}}) D(\tilde{\mathbf{x}}) = \sum_{x \in S} \delta(\tilde{\mathbf{x}}) \tilde{\mathbf{w}}^{T} \tilde{\mathbf{x}}$$

όπου *S* το σύνολο των διανυσμάτων που ταξινομήθηκαν λάθος και

$$\text{Για } \tilde{x} \in S,\ \delta(\tilde{\mathbf{x}}) = sign\big(D(\tilde{\mathbf{x}})\big) = \begin{cases} -1 & \text{αν } \tilde{\mathbf{x}} \in A^{+} \\ +1 & \text{αν } \tilde{\mathbf{x}} \in B^{-} \end{cases} \qquad (3.2.14)$$

Η συνάρτηση $K(\tilde{\mathbf{w}})$ εκφράζει ουσιαστικά το συνολικό σφάλμα ταξινόμησης και είναι κατά τμήματα γραμμική συνάρτηση. Η βέλτιστη λύση του προβλήματος είναι η εύρεση ανύσματος βαρών $\tilde{\mathbf{w}}$ ώστε $K(\tilde{\mathbf{w}}) = 0$. Αν t=0,1,2,… και $\tilde{\mathbf{w}}(0)$ είναι μία αρχική τιμή, η τιμή του $\tilde{\mathbf{w}}$ μπορεί να προσεγγισθεί επαναληπτικά με την με την μέθοδο της *επικλινούς κατάβασης* (*gradient descent*) σύμφωνα με την σχέση

$$\tilde{\boldsymbol{w}}(t+1) = \tilde{\boldsymbol{w}}(t) - \rho \frac{\partial K}{\partial \tilde{\boldsymbol{w}}}\Big|_{\tilde{\boldsymbol{w}}(t)}^{T} = \tilde{\boldsymbol{w}}(t) - \rho \sum_{x \in S} \delta(\tilde{\boldsymbol{x}}) \tilde{\boldsymbol{x}} \qquad (3.2.15)$$

Η διόρθωση γίνεται αφού για το σύνολο *S* και η διαδικασία εκπαίδευσης χαρακτηρίζεται ως εκπαίδευση *δέσμης* (*batch*).

Υλοποίηση του Perceptron με εκπαίδευση κατά δέσμες

```
function [ w ] = perceptron_batch( trset,...% Το σύνολο εκπαίδευσης
  class_labels, ...   % η κλάση που ανήκει το κάθε πρότυπο, -1 ή +1
  learning_rate,...   % ο ρυθμός εκμάθησης ρ
  error_tolerance ... % Πλήθος προτύπων που μπορεί να ταξινομηθούν λάθος
)
vector_num = size(trset,2);  % Το πλήθος των διανυσμάτων (trset columns)
feature_num = size(trset,1); % Το πλήθος των χαρακτηριστικών (trset rows)
X=[ trset; ones(1, vector_num) ]; % Επαύξηση διανυσμάτων
w = rand(feature_num+1,1); % Αρχικοποίηση των βαρών με τυχαίες μικρές τιμές

 while true
  Dx=w'*X;
  flag_mscl_vector = sign(Dx) ~= class_labels;

  if sum( flag_mscl_vector ) <= error_tolerance; break; end

  msclSet = flag_mscl_vector.*X;

  deltaX = -class_labels;
  %loss= sum(deltaX.*Dx.*flag_mscl_vector);disp(loss)
  S = sum( msclSet .* deltaX, 2);
  w = w - learning_rate * S;
  % ---- Σχεδιασμός της διαχωριστική γραμμής,
  %      εφαρμόζεται μόνο για δύο χαρακτηριστικά
  %      άλλως πρέπει να αφαιρεθεί (ως σχόλιο)
  t=[-2 8]; l=-w(1)/w(2)*t-w(3)/w(2); h=plot(t,l,'k--');
  pause(.1);  delete(h);
    %----------
 end % of while loop
end % of function
```

Είναι χρήσιμο να εφαρμόσουμε στην τιμή του αθροιστή σ=D(.) την συνάρτηση *προσήμου* ή την *βηματική*

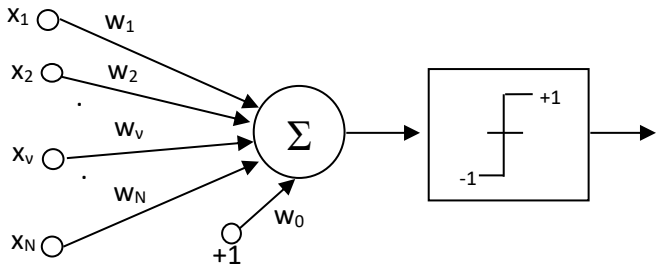
Βηματική 0/1 (step function 0/1):

$$f(\sigma) = \begin{cases} 0 & \alpha\nu \quad \sigma < 0 \\ 1 & \alpha\nu \quad \sigma \geq 0 \end{cases}$$

Προσήμου -1/1 (sign function -1/1):

$$f(\sigma) = \begin{cases} -1 & \alpha\nu \quad \sigma < 0 \\ +1 & \alpha\nu \quad \sigma \geq 0 \end{cases}$$

ώστε να κωδικοποιείται κάθε κλάση π.χ. με +1 αν $d(\mathbf{x}^A) \geq 0$ και το -1 ή το 0 αν $d(\mathbf{x}^B) < 0$. Η συναρτήσεις αυτές λέγονται και *συναρτήσεις ενεργοποίησης* (*activation functions*). Κατόπιν τούτων μπορούμε να παραστήσουμε διαγραμματικά τον ταξινομητή όπως δείχνεται στο Σχ.3.2.2.



Σχ.3.3.2

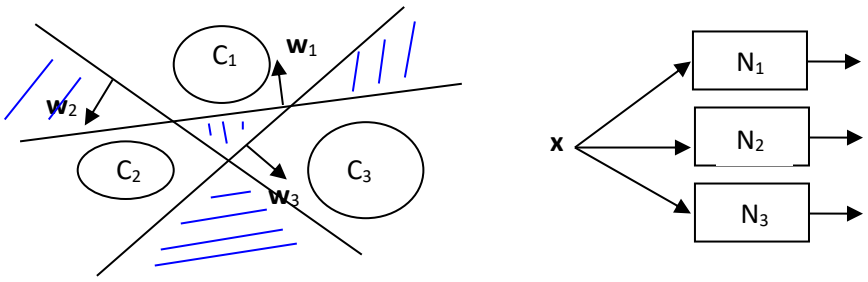
Ο γραμμικός ταξινομητής είναι ένα απλό μοντέλο του βιολογικού νευρώνα και ονομάσθηκε νευρώνας perceptron. Προτάθηκε αρχικά από τους McCulloch και Pitts για δυαδικές τιμές (0 ή 1) στην είσοδο και την έξοδο. Ολοκληρώθηκε από τον ψυχολόγο F. Rosenblatt, και αποτελεί το δομικό στοιχείο των τεχνητών νευρωνικών δικτύων (ΤΝΔ) που μπορεί να αποτελούνται από εκατομμύρια νευρώνες.

[1] McCulloch, W; Pitts, W (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics*. **5** (4): 115–133.

[2] Rosenblatt, Frank (1957). "The Perceptron—a perceiving and recognizing automaton". *Report 85-460-1*. Cornell Aeronautical Laboratory.

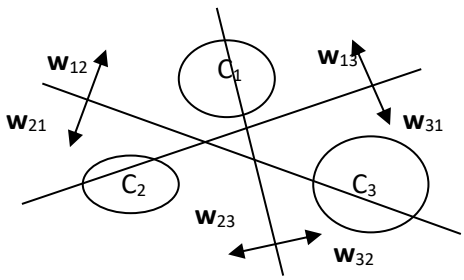
3.2.1. Η περίπτωση πολλών κλάσεων.

Στην περίπτωση που το πλήθος των κλάσεων είναι M>2 τότε θα πρέπει να χρησιμοποιήσουμε περισσότερους του ενός γραμμικούς ταξινομητές. Μία περίπτωση είναι να διαχωρίσουμε γραμμικά κάθε τάξη από τις υπόλοιπες (εάν αυτό είναι εφικτό) με M το πλήθος γραμμικές διακριτικές συναρτήσεις. Αν $d_\mu(\mathbf{x}) = \mathbf{w}_\mu^T \cdot \mathbf{x} + c > 0$, $\mu = 1 \ldots M$, όταν το $\mathbf{x}$ ανήκει στη κλάση C_μ (Σχ. 3.2.1.1), τότε ένα πρότυπο ανήκει στην κλάση της οποίας η γραμμική διακριτική συνάρτηση $d_\mu(\mathbf{x})$ είναι θετική. Στην περίπτωση υλοποίησης με ΝΔ και βηματική έξοδο 0/1 το άγνωστο πρότυπο ταξινομείται στην κλάση που ο αντίστοιχος νευρώνας έχει έξοδο 1. Μια τέτοια προσέγγιση όπως δείχνεται στο σχήμα. Σχ. 3.2.1.1 είναι δυνατόν να αποδώσει ένα πρότυπο σε καμία ή σε πολλές τάξεις.



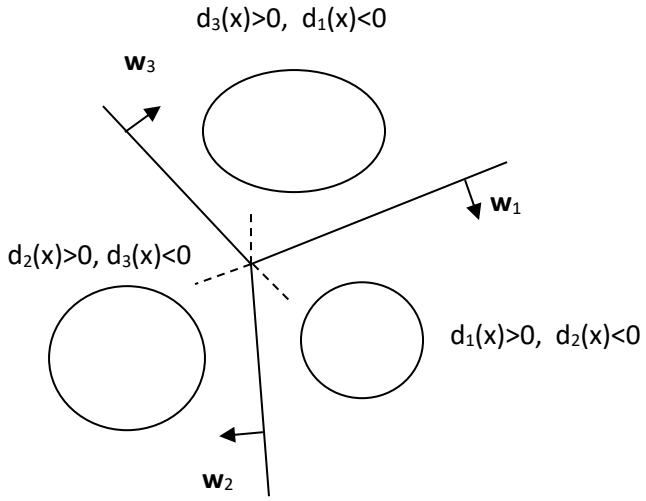
Σχ. 3.2.1.1

Ένας άλλος τρόπος αντιμετώπισης του προβλήματος είναι ο διαχωρισμός των κλάσεων ανά δύο αδιαφορώντας για τις υπόλοιπες. Το πρότυπο αποδίδεται σε εκείνη την κλάση που έχει M-1 θετικές τιμές στις διακριτικές συναρτήσεις που την χωρίζουν από τις υπόλοιπες. Η μέθοδος αυτή είναι υπολογιστικά πολύπλοκη στην περίπτωση των πολλών κλάσεων σχήμα 3.2.1.2. Τέλος αν οι διαχωριστικές επιφάνειες είναι όπως στο σχήμα 3.2.1.3 αποφεύγεται η ύπαρξη διφορούμενων περιοχών στο χώρο των προτύπων. Συνδυασμοί γραμμικών ή κατά τμήματα γραμμικών ταξινομητών με την χρήση διαφόρων συναρτήσεων στην έξοδο έχουν προταθεί όπως είναι τα ΝΔ ADALINE, MADALINE οι μηχανές επιτροπής (COMMITTEE MACHINES) κ.α. Στα συστήματα αυτά δεν θα αναφερθούμε ενδελεχώς δεδομένου ότι η κεντρική ιδέα της λειτουργίας του καλύπτεται από όσα αναφέρθηκαν μέχρι τώρα. Σε επόμενο κεφάλαιο θα περιγράψουμε συστήματα που αποτελούνται από διαδοχικά επίπεδα ταξινομητών (MLP: multilayer perceptron) με παραγωγίσιμες συναρτήσεις ενεργοποίησης για την επίλυση μη γραμμικών προβλημάτων.



Σχ. 3.2.1.2

,



Σχ. 3.2.1.3

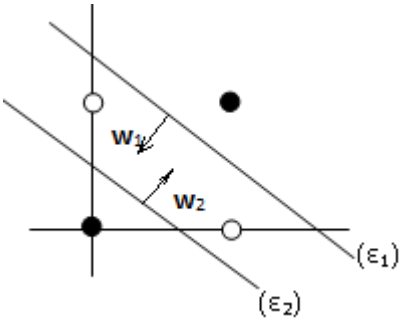
3.2.2. Η περίπτωση της XOR, ταξινομητές πολλών επιπέδων

Όταν το πρόβλημα της ταξινόμησης αφορά περισσότερες από δύο κλάσεις ή όταν οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες, είναι δυνατόν να επιτευχθούν λύσεις με κατάλληλους συνδυασμούς γραμμικών ταξινομητών. Μια τέτοια χαρακτηριστική περίπτωση είναι αυτή της λογικής πύλης XOR της οποίας ο πίνακας αληθείας δείχνεται στον Πιν.3.2.2.1.

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

Πινακας 3.2.2.1.

Σύμφωνα με αυτόν οι συνδυασμοί των τιμών των λογικών μεταβλητών α, β αποτελούν τέσσερα πρότυπα που περιγράφονται από τα διανύσματα στοιχείων συνόλου **Ω**={(0,0), (0,1), (1,0), (1,1)} και η πράξη α XOR β ορίζει τις κλάσεις C_0={(0,0),(1,1)} και C_1={(0,1),(1,0)}. Στο Σχ.3.4.2.1 φαίνονται τα άκρα των διανυσμάτων στο E^2. Είναι προφανές ότι οι κλάσεις C_0, C_0 δεν διαχωρίζονται με μια ευθεία. Ο διαχωρισμός των κλάσεων μπορεί να γίνει με δύο ευθείες (Σχ.3.2.2.1) που ορίζουν μία ζώνη στο εσωτερικό της οποίας βρίσκονται τα πρότυπα της κλάσης C_1. Η ευθεία (ε_1) μπορεί να προσδιοριστεί από έναν γραμμικό ταξινομητή T_1 που θα διαχωρίζει το πρότυπο (1,1) από τα υπόλοιπα. Η ευθεία (ε_2) μπορεί να προσδιορισθεί από έναν γραμμικό ταξινομητή T_2 που θα διαχωρίζει το πρότυπο (0,0) από τα υπόλοιπα.



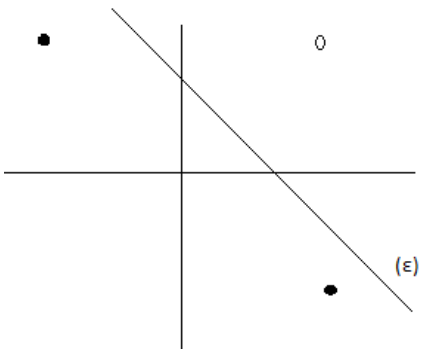
Σχήμα 3.2.2.1

Οι έξοδοι των T_1, T_2 θα είναι οι τιμές των συναρτήσεων $\sigma_1=f(\mathbf{x})$, $\sigma_2=f(\mathbf{x})$ για $x \in \Omega$ και f την συνάρτηση προσήμου *όπως* φαίνονται στον ακόλουθο πίνακα 3.2.2.2 :

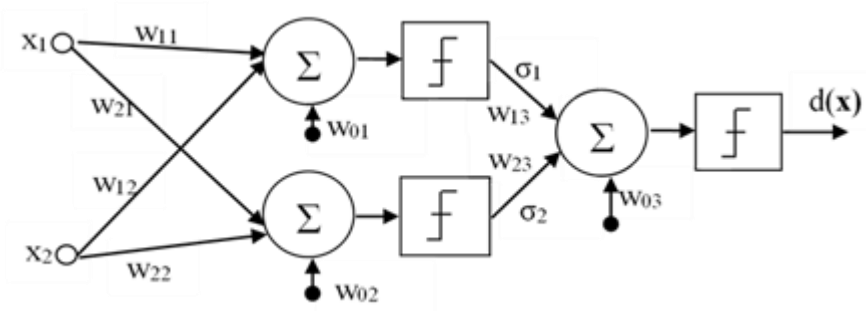
x_1	x_2	σ_1	σ_2	**ΚΛΑΣΗ**
0	0	+1	-1	C_0
0	1	+1	+1	C_1
1	0	+1	+1	C_1
1	1	-1	+1	C_0

Πίνακας 3.2.2.2

Οι τιμές των $\sigma_1(\mathbf{x})$, $\sigma_2(\mathbf{x})$ αποτελούν ένα νέο σύνολο προτύπων **Φ**={(+1,-1),(-1,+1),(+1,+1)}. Στο Σχ.3.4.2.3 φαίνεται ο χώρος του **Φ**. Ένας γραμμικός ταξινομητής Τ μπορεί να προσδιορίσει την **ε**. Το όλο σύστημα φαίνεται στο Σχ.3.2.2.3. Αφήνεται σαν άσκηση στον αναγνώστη ο υπολογισμός των συναπτικών βαρών του ΝΔ του σχήματος 3.2.2.3 με γραμμικές διακριτικές συναρτήσεις όπως των ευθειών των σχημάτων 3.2.2.1, 3.2.2.2. Προσέξτε παρατηρώντας τον πίνακα 3.4.2.2, ότι ο ένας νευρώνας του κρυφού επιπέδου υλοποιεί την λογική πράξη OR (α+β) και ο άλλος την λογική πράξη AND (α·β).



Σχ.3.2.2.2.



Σχ.3.2.2.3

Με τη χρήση περισσότερων νευρώνων παρατεταγμένων σε επίπεδα όπως στο Σχ.3.2.2.3 δημιουργούμε πολυεπίπεδα δίκτυα perceptron (*MLP: Multi Layer Perceptron*) που μπορούν να διαχωρίσουν κλάσεις μη γραμμικά διαχωρίσιμες. Τα επίπεδα ανάμεσα στο επίπεδο εισόδου των τιμών του ανύσματος **x** και του επιπέδου των νευρώνων εξόδου, ονομάζονται κρυφά επίπεδα. Το νευρωνικό δίκτυο του Σχ.3.2.2.3 έχει ένα κρυφό επίπεδο δύο νευρώνων. Ο διαχωρισμός μη γραμμικά διαχωρίσιμων κλάσεων με την χρήση MLP όπως παρουσιάσθηκε μέχρι τώρα, απαιτεί την παρέμβαση του ανθρώπου σχεδιαστή και δεν είναι μια γενική και αυτοματοποιημένη διαδικασία. Οι περισσότεροι αλγόριθμοι εκπαίδευσης βασίζονται στην εύρεση του ελαχίστου μιας κατάλληλης συνάρτησης κόστους των συναπτικών βαρών. Μια τέτοια προσέγγιση βασίζεται στην κατάβαση αντίθετα από την κλίση της συνάρτησης αυτής (*gradient descent*). Οι βηματικές συναρτήσεις που χρησιμοποιήθηκαν μέχρι τώρα δεν είναι συνεχείς στο μηδέν και έχουν μηδενικές παραγώγους για τις άλλες τιμές. Άλλες συναρτήσεις παραγωγίσιμες σε όλο το πεδίο ορισμού τους μπορούν να χρησιμοποιηθούν και να δώσουν ικανοποιητική λύση όπως θα δούμε ακολούθως.

3.3. Πολυεπίπεδοι ταξινομητές – Διόρθωση σφάλματος με οπισθοδιάδοση (Back Error Propagation)

Ο perceptron αν και εντυπωσίασε με την ομοιότητα που εμφάνιζε με τα βιολογικά νευρωνικά δίκτυα εν τούτοις δεν μπορούσε να επιλύσει αυτόματα ακόμη και απλά προβλήματα. Για περισσότερο από μια δεκαετία η εφαρμογή του και το έντονο ερευνητικό ενδιαφέρον ατόνησαν αλλά δεν εγκαταλείφθηκαν. Οι Roumelhart et al[3] πρότειναν μια νέα εκδοχή MLP ΤΝΔ που έδωσε μεγάλη ώθηση στον αντίστοιχο επιστημονικό πεδίο.

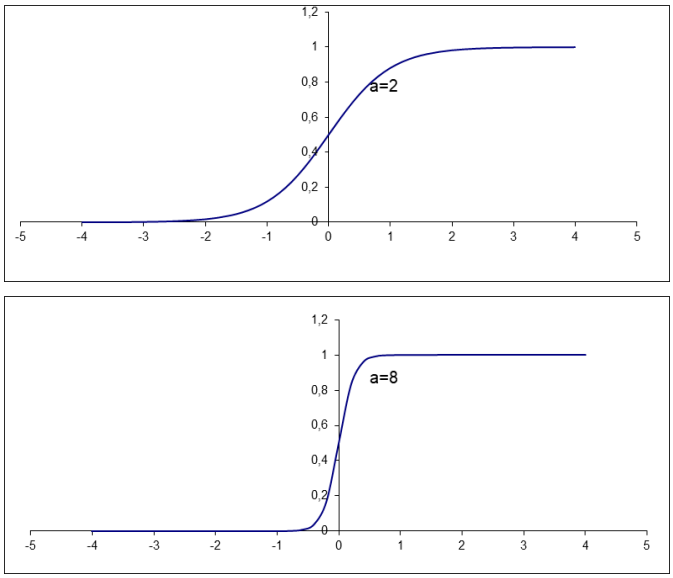
[3] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (1986a). "*Learning representations by back-propagating errors*". *Nature*. **323** (6088): 533–536.

Στην νέα αυτή εκδοχή η βηματική συνάρτηση εξόδου αντικαθίσταται από συνεχείς και παραγωγίσιμες συναρτήσεις που την προσεγγίζουν. Τέτοιες συναρτήσεις είναι οι ακόλουθες:

- *Σιγμοειδής-λογιστική* (*sigmoid logistic*):

$$f(\sigma) = \frac{1}{1+e^{-a\sigma}} \qquad (3.3.1)$$

Η παράμετρος α λέγεται παράμετρος *κλίσης* (*slope*) ή *λοξότητος*, (Σχ.3.3.1)



Σχήμα.3.3.1

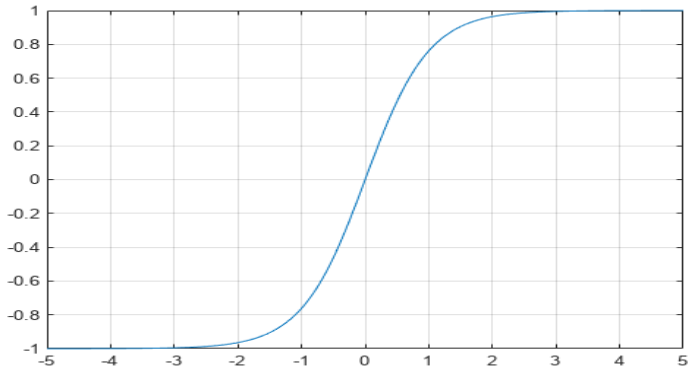
Αξίζει να αναφερθεί ότι

$$f'(\sigma) = a \cdot f(\sigma) \cdot \big(1 - f(\sigma)\big)$$

Άρα για y=f(σ), $y'(\sigma) = a \cdot y \cdot (1 - y)$

- Σιγμοειδής-Υπερβολική *εφαπτομένη (hyperbolic tangent)*:

$$f(\sigma) = tanh(\sigma) = \frac{1 - e^{-a\sigma}}{1 + e^{-a\sigma}}$$

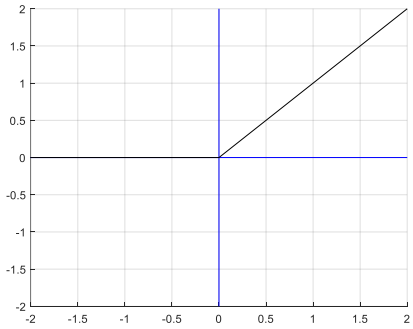


Σχήμα. Η συνάρτηση της υπερβολικής εφαπτομένης για a=2

- *Συνάρτηση ράμπας (ramp function)*: $f(\sigma) = \begin{cases} 0 & \alpha\nu \quad \sigma < 0 \\ \sigma & \alpha\nu \quad \sigma \geq 0 \end{cases} = \max(0, \sigma)$.

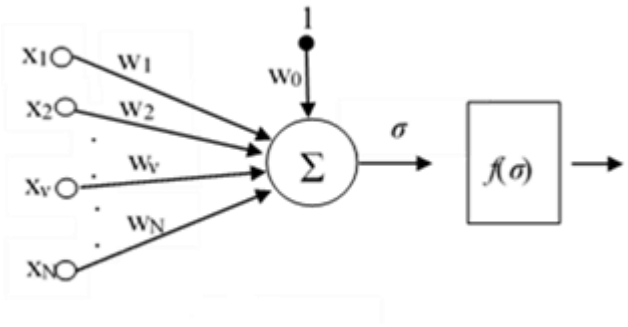
  Ονομάζεται και (*ReLU: Rectified Linear Unit*). Η ReLU χρησιμοποιείται

κατά κόρον στα *συνελικτικά νευρωνικά δίκτυα* που θα παρουσιασθούν σε επόμενο κεφάλαιο.



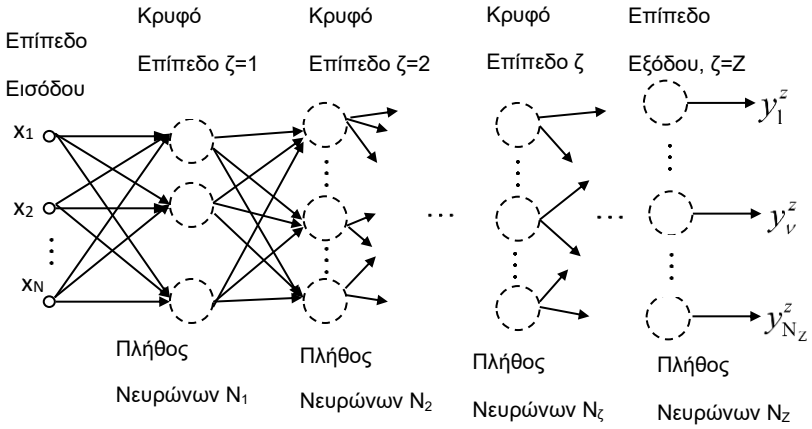
Σχήμα: Η συνάρτηση ReLU.

Η χρήση της f(x) διευκολύνει την εφαρμογή μιας μεθόδου ελαχιστοποίησης της κατάλληλης συνάρτησης κόστους. Ο τεχνητός νευρώνας έχει τώρα την μορφή του Σχ. 3.3.2



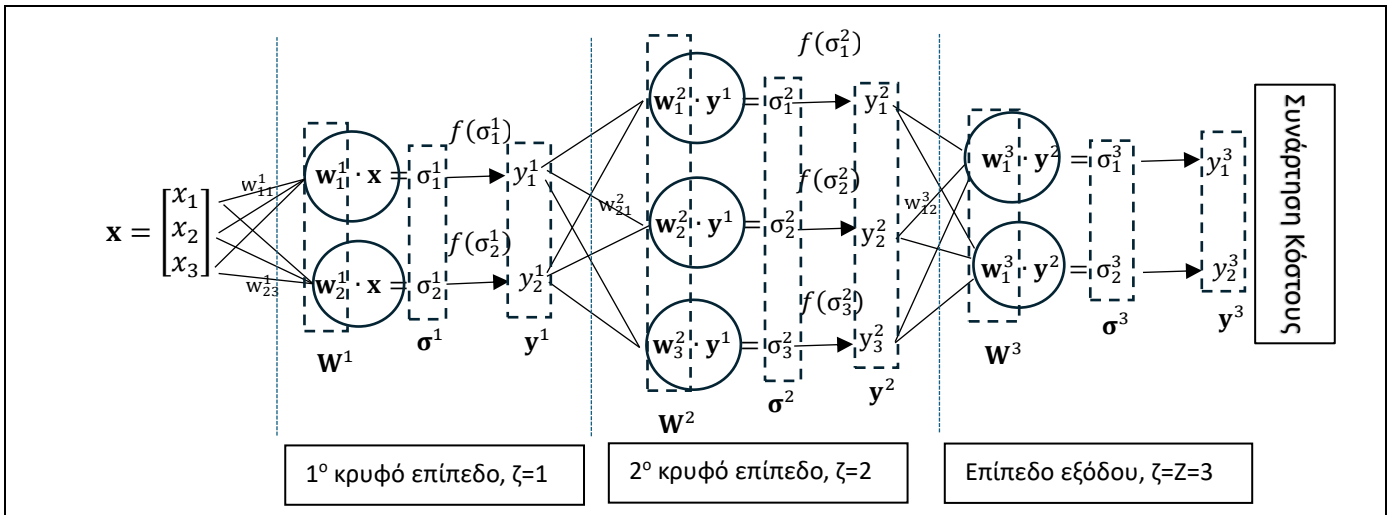
Σχ. 3.3.2

Το όλο ΝΔ όπως θα παρουσιασθεί και θα αναλυθεί ακολούθως βασίζεται στα παραπάνω. Είναι ένας πολυεπίπεδος perceptron και ονομάζεται λόγω της μεθόδου εκπαίδευσής του, ΝΔ Back error propagation ή για συντομία Back propagation. Στο νευρωνικό δίκτυο υπάρχουν επίπεδα δηλαδή ομάδες νευρώνων, πλήρως διασυνδεδεμένα μεταξύ τους Σχ.3.3.3. Η έξοδος κάθε νευρώνα ενός επιπέδου εισέρχεται σε κάθε νευρώνα του επόμενου επιπέδου, με αντίστοιχη σύναψη (βάρος). Οι τιμές του πίνακα εισόδου αποτελούν το επίπεδο εισόδου. Εκτός από το επίπεδο εισόδου υπάρχουν Z το πλήθος επόμενα διαδοχικά επίπεδα, το τελευταίο των οποίων ονομάζεται επίπεδο εξόδου. Τα επίπεδα που ενδεχομένως υπάρχουν μεταξύ των επιπέδων εισόδου και εξόδου, λέγονται κρυφά επίπεδα. Για την αρίθμηση των επιπέδων θα χρησιμοποιούμε τον δείκτη ζ με τιμές ζ=0,…,Z, με ζ=0 για το επίπεδο εισόδου και ζ=Z για το επίπεδο εξόδου.



Σχ.3.3.3.

Στο Σχ. 3.3.4. φαίνεται αναλυτικά ένα νευρωνικό όπου το πρότυπο εισόδου έχει δύο χαρακτηριστικά $[x_1,x_2]^T$ που εισέρχονται σε ένα κρυφό επίπεδο δύο νευρώνων οι έξοδοι των οποίων είναι είσοδοι ενός επόμενου επιπέδου με τρείς νευρώνες, και επίπεδο εξόδου με δύο νευρώνες, δύο τελικές τιμές εξόδου.



Σχ. 3.3.4. Αναλυτικό διάγραμμα ενός πολυεπίπεδου Perceptron (Αντιλήπτωρ)

Ονομάζουμε N_ζ το πλήθος των νευρώνων ενός επιπέδου ζ. Ως εκ τούτου το πλήθος των στοιχείων του πίνακα εισόδου x είναι N_0 και του πίνακα εξόδου y είναι N_Z. Αν ν είναι δείκτης για αρίθμηση των νευρώνων του επιπέδου ζ, τότε οι συνάψεις του συνθέτουν έναν πίνακα γραμμής $\boldsymbol{w}_\nu^\zeta$. Αν μ δείκτης για την αρίθμηση των νευρώνων του προηγουμένου επιπέδου ζ-1 που αποτελείται από $N_{\zeta-1}$ νευρώνες, τότε

$$\boldsymbol{w}_\nu^\zeta = [w_{\nu 1}^\zeta,\ldots,w_{\nu\mu}^\zeta,\ldots,w_{\nu,N_{\zeta-1}}^\zeta,w_{\nu 0}^\zeta]$$

Η έξοδος του αθροιστή του νευρώνα θα είναι σ_ν^ζ

$$\sigma_\nu^\zeta = \mathbf{w}_\nu^\zeta \cdot \mathbf{y}^{\zeta-1} \quad \text{με } \mathbf{y}^{\zeta-1} = \left[y_1^{\zeta-1},\ldots,y_{N_{\zeta-1}}^{\zeta-1}\right]$$

Η τελική έξοδος του νευρώνα, y_ν^ζ προκύπτει από την σχέση

$$y_\nu^\zeta = f\left(\sigma_\nu^\zeta\right)$$

όπου f(.) κατάλληλη μη γραμμική συνάρτηση π.χ. η σιγμοειδής λογιστική, η υπερβολική εφαπτομένη κ.α. Η f(.) ονομάζεται και ως συνάρτηση ενεργοποίησης και πρέπει να είναι συνεχής και παραγωγίσιμη.

Γράφοντας υπό μορφή πινάκων τα παραπάνω ορίζουμε:

$$\boldsymbol{\sigma}^{\zeta}=\begin{bmatrix}\sigma_1^{\zeta}\\ \vdots\\ \sigma_{\nu}^{\zeta}\\ \vdots\\ \sigma_{N_\zeta}^{\zeta}\end{bmatrix},\ \mathbf{W}^{\zeta}=\begin{bmatrix}\mathbf{w}_1^{\zeta}\\ \vdots\\ \mathbf{w}_{\nu}^{\zeta}\\ \vdots\\ \mathbf{w}_{N_\zeta}^{\zeta}\end{bmatrix}=\begin{bmatrix}w_{11}^{\zeta}\cdots w_{1\nu}^{\zeta}\cdots w_{1N_{\zeta-1}}^{\zeta}\\ \vdots\quad\vdots\quad\vdots\\ w_{\nu 1}^{\zeta}\cdots w_{\nu\mu}^{\zeta}\cdots w_{\nu N_{\zeta-1}}^{\zeta}\\ \vdots\quad\vdots\quad\vdots\\ w_{N_\zeta 1}^{\zeta}\cdots w_{N_\zeta\mu}^{\zeta}\cdots w_{N_\zeta N_{\zeta-1}}^{\zeta}\end{bmatrix},\quad \mathbf{y}^{\zeta}=\begin{bmatrix}y_1^{\zeta}\\ \vdots\\ y_{\nu}^{\zeta}\\ \vdots\\ y_{N_\zeta}^{\zeta}\end{bmatrix}=\begin{bmatrix}f(\sigma_1^{\zeta})\\ \vdots\\ f(\sigma_{\nu}^{\zeta})\\ \vdots\\ f(\sigma_{N_\zeta}^{\zeta})\end{bmatrix}$$

μ=1,....,N_ζ, ν=1,....,$N_{\zeta-1}$ και $f(.)$ μία μη γραμμική συνάρτηση. Τα βάρη του ν-οστού νευρώνα παρατάσσονται σε διάνυσμα γραμμής.

ΣΥΝΑΡΤΗΣΗ ΚΟΣΤΟΥΣ

Έστω i=1,…,I με I το πλήθος των ζευγών $(\mathbf{x}_i,\mathbf{y}_i)$ που είναι τα στοιχεία του συνόλου εκπαίδευσης S={$(\mathbf{x}_i,\mathbf{y}_i)$} όπου $\mathbf{x}_i$ πίνακας στήλης της εισόδου και $\mathbf{y}_i$ ο αντίστοιχος επιθυμητός πίνακας στήλης της εξόδου. Η συνάρτηση $\mathcal{L}(\mathbf{W}^1,\cdots,\mathbf{W}^{\zeta},\cdots,\mathbf{W}^{Z},S)=\sum_i \Delta_i$ ονομάζεται συνάρτηση συνολικού κόστους. Το ζητούμενο είναι η εύρεση των κατάλληλων $\mathbf{W}^{\zeta}$ ώστε αυτή να ελαχιστοποιηθεί (πρόβλημα βελτιστοποίησης). Υπάρχουν διάφορες μέθοδοι προσέγγισης/βελτιστοποίησης που λέγονται *optimizers* (GD, GDM, SGDM, ADAM κ.α.). Οι μέθοδοι αυτές είναι επαναληπτικές. Επαναληπτικά γίνεται διόρθωση των $\mathbf{W}^{\zeta}$ έως ότου η $\mathcal{L}(.)$ να ελαχιστοποιηθεί Η διόρθωση μπορεί να γίνεται αφού προστεθούν όλα τα Δ_i, η κατά ομάδες (bathes:δέσμες) ή για κάθε πρότυπο. Σε κάθε περίπτωση πρέπει να υπολογισθεί η ποσότητα

$$\Delta_i(\mathbf{W}^1,\cdots,\mathbf{W}^{\zeta},\cdots,\mathbf{W}^{Z},\mathbf{x}_i)$$

Υπάρχουν συναρτήσεις κόστους όπως αυτή του Μέσου Τετραγωνικού Σφάλματος (Mean Square Error, MSE) και της Διασταυρούμενης Εντροπίας (Cross-Entropy). Η συνάρτηση κόστους του Μέσου Τετραγωνικού Σφάλματος για $\mathbf{y}_i$ επιθυμητές τιμές και εκτιμήσεις/εξόδους $\mathbf{y}_i^Z$, i=1…I, δίνεται από την σχέση:

$$\mathrm{MSE}=\frac{1}{I}\sum_{i=1}^{N}(\mathbf{y}_i^Z-\mathbf{y}_i)^T(\mathbf{y}_i^Z-\mathbf{y}_i)$$

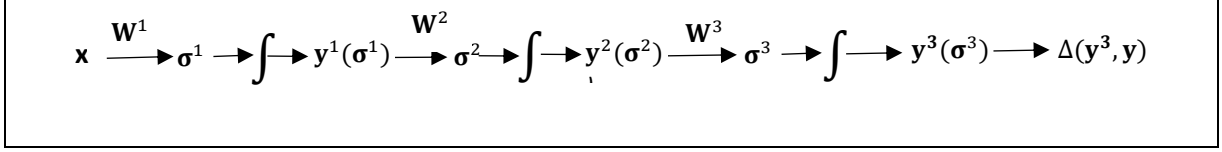
Η ποσότητα I παραμένει σταθερή. Για την ελαχιστοποίηση της ενδιαφέρει βασικά η ποσότητα $\Delta_i=(\mathbf{y}_i^Z-\mathbf{y}_i)^T(\mathbf{y}_i^Z-\mathbf{y}_i)$. Γενικά για $\Delta=\Delta_i/2$, $\mathbf{y}=\mathbf{y}_i$ και $\mathbf{x}=\mathbf{x}_i$, ισχύει:

$$\Delta(\mathbf{y}^z,\mathbf{y})=\frac{1}{2}\cdot \mathrm{D}_{\mathrm{E}}^2(\mathbf{y}^z,\mathbf{y})=\frac{1}{2}(\mathbf{y}^z-\mathbf{y})^T\cdot(\mathbf{y}^z-\mathbf{y})\quad\Rightarrow\frac{\partial\Delta}{\partial\mathbf{y}^z}=(\mathbf{y}^z-\mathbf{y})^T$$

Η συνάρτηση κόστους που βασίζεται στο MSE μπορεί να χρησιμοποιηθεί για προβλήματα ταξινόμησης και προβλήματα παλινδρόμησης και πρόβλεψης. Στις πλείστες των περιπτώσεων στο επίπεδο εξόδου δεν χρησιμοποιούμε κάποια συνάρτηση ενεργοποίησης δηλαδή μη γραμμικότητας. Για τον λόγο αυτό $\mathbf{y}(\boldsymbol{\sigma})=\boldsymbol{\sigma}$, και προφανώς $\frac{\partial\mathbf{y}}{\partial\boldsymbol{\sigma}}=\mathbf{I}$, $\mathbf{I}$ ο μοναδιαίος πίνακας.

Υπολογισμός των βαρών με οπισθοδιάδοση του σφάλματος (back-error propagation).

Το δίκτυο του Σχ.3.3.4 μπορεί να αναπαρασταθεί διαγραμματικά όπως στο ακόλουθο Σχήμα 3.3.5



	1ο επίπεδο		2ο επίπεδο		3ο επίπεδο	Κόστος Δ
$\mathbf{x}$ →	$\boldsymbol{\sigma}^1=\mathbf{W}^1\mathbf{x}$ → $\mathbf{y}^1=\mathbf{y}^1(\boldsymbol{\sigma}^1)$	→	$\boldsymbol{\sigma}^2=\mathbf{W}^2\mathbf{y}^1$ → $\mathbf{y}^2=\mathbf{y}^2(\boldsymbol{\sigma}^2)$	→	$\boldsymbol{\sigma}^3=\mathbf{W}^3\mathbf{y}^2$ → $\mathbf{y}^3=\mathbf{y}^3(\boldsymbol{\sigma}^3)$	$\Delta(\mathbf{y}^3,\mathbf{y})$

$$y_\nu^\zeta = f\left(\sigma_\nu^\zeta\right) = \frac{1}{1+e^{-a\cdot\sigma_\nu^\zeta}}$$

Σχήμα [3.3.5]

Τα διανύσματα είναι επαυξημένα και οι πίνακες περιέχουν τους σταθερούς όρους. Σύμφωνα με τον κανόνα παραγώγισης της αλυσίδας:

$\frac{\partial\Delta}{\partial\mathbf{W}^3}=\underbrace{\frac{\partial\Delta}{\partial\mathbf{y}^3}\cdot\frac{\partial\mathbf{y}^3}{\partial\boldsymbol{\sigma}^3}}_{\boldsymbol{\delta}^3}\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\mathbf{W}^3}$	$\frac{\partial\Delta}{\partial\mathbf{W}^3}=\boldsymbol{\delta}^3\frac{\partial\boldsymbol{\sigma}^3}{\partial\mathbf{W}^3}$	$\boldsymbol{\delta}^3=\frac{\partial\Delta}{\partial\mathbf{y}^3}\cdot\frac{\partial\mathbf{y}^3}{\partial\boldsymbol{\sigma}^3}$
$\frac{\partial\Delta}{\partial\mathbf{W}^2}=\underbrace{\underbrace{\frac{\partial\Delta}{\partial\mathbf{y}^3}\cdot\frac{\partial\mathbf{y}^3}{\partial\boldsymbol{\sigma}^3}}_{\boldsymbol{\delta}^3}\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\boldsymbol{\sigma}^2}}_{\boldsymbol{\delta}2}\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\mathbf{W}^2}=\underbrace{\boldsymbol{\delta}^3\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\boldsymbol{\sigma}^2}}_{\boldsymbol{\delta}^2}\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\mathbf{W}^2}$	$\frac{\partial\Delta}{\partial\mathbf{W}^2}=\boldsymbol{\delta}^2\frac{\partial\boldsymbol{\sigma}^2}{\partial\mathbf{W}^2}$	$\boldsymbol{\delta}^2=\boldsymbol{\delta}^3\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\boldsymbol{\sigma}^2}$ $=\boldsymbol{\delta}^3\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\mathbf{y}^2}\cdot\frac{\partial\mathbf{y}^2}{\partial\boldsymbol{\sigma}^2}$
$\frac{\partial\Delta}{\partial\mathbf{W}^1}=\underbrace{\underbrace{\frac{\partial\Delta}{\partial\mathbf{y}^3}\cdot\frac{\partial\mathbf{y}^3}{\partial\boldsymbol{\sigma}^3}}_{\boldsymbol{\delta}^3}\cdot\frac{\partial\boldsymbol{\sigma}^3}{\partial\boldsymbol{\sigma}^2}}_{\boldsymbol{\delta}2}\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\boldsymbol{\sigma}^1}\cdot\frac{\partial\boldsymbol{\sigma}^1}{\partial\mathbf{W}^1}=\underbrace{\boldsymbol{\delta}^2\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\boldsymbol{\sigma}^1}}_{\boldsymbol{\delta}1}\cdot\frac{\partial\boldsymbol{\sigma}^1}{\partial\mathbf{W}^1}$	$\frac{\partial\Delta}{\partial\mathbf{W}^1}=\boldsymbol{\delta}^1\frac{\partial\boldsymbol{\sigma}^1}{\partial\mathbf{W}^1}$	$\boldsymbol{\delta}^1=\boldsymbol{\delta}^2\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\boldsymbol{\sigma}^1}$ $=\boldsymbol{\delta}^2\cdot\frac{\partial\boldsymbol{\sigma}^2}{\partial\mathbf{y}^1}\cdot\frac{\partial\mathbf{y}^1}{\partial\boldsymbol{\sigma}^1}$

Από τα παραπάνω μπορούμε να γράψουμε γενικά ότι

$$\frac{\partial\Delta}{\partial\mathbf{W}^Z}=\boldsymbol{\delta}^Z\cdot\frac{\partial\boldsymbol{\sigma}^Z}{\partial\mathbf{W}^Z}\quad \text{με } \boldsymbol{\delta}^Z=\frac{\partial\Delta}{\partial\mathbf{y}^Z}\cdot\frac{\partial\mathbf{y}^Z}{\partial\boldsymbol{\sigma}^Z}$$
$$\frac{\partial\Delta}{\partial\mathbf{W}^\zeta}=\boldsymbol{\delta}^\zeta\cdot\frac{\partial\boldsymbol{\sigma}^\zeta}{\partial\mathbf{W}^\zeta}\quad \text{με } \boldsymbol{\delta}^\zeta=\boldsymbol{\delta}^{\zeta+1}\cdot\frac{\partial\boldsymbol{\sigma}^{\zeta+1}}{\partial\boldsymbol{\sigma}^\zeta}=\boldsymbol{\delta}^{\zeta+1}\cdot\frac{\partial\boldsymbol{\sigma}^{\zeta+1}}{\partial\mathbf{y}^\zeta}\cdot\frac{\partial\mathbf{y}^\zeta}{\partial\boldsymbol{\sigma}^\zeta}$$

Ο τρόπος αυτός υπολογισμού του $\frac{\partial\Delta}{\partial\mathbf{W}^\zeta}$ ονομάζεται οπισθοδιάδοση του σφάλματος (Back error propagation). Σημειωταίον ότι για το προτελευταίο επίπεδο ζ=Ζ-1, Ζ=ζ+1 και ότι ο πολλαπλασιασμός διανύσματος γραμμής επί ένα δισδιάστατο πίνακα δίνει διάνυσμα γραμμής, άρα τα $\boldsymbol{\delta}^\zeta$ είναι διανύσματα γραμμής. Θα αναφέρουμε τώρα τις απαραίτητες μαθηματικές έννοιες και σχέσεις που θα μας βοηθήσουν στους υπολογισμούς μας.

Παράγωγοι συναρτήσεων πολλών μεταβλητών και υπολογισμοί

Έστω $\mathbf{y} \in \mathcal{R}^{\mathrm{N}\times 1}$ και $\Delta(\mathbf{y}) \in \mathcal{R}$ τότε $\frac{\partial \Delta}{\partial \mathbf{y}} = \left[\frac{\partial \Delta}{\partial \mathrm{y}_1} \quad \cdots \quad \frac{\partial \Delta}{\partial \mathrm{y}_\nu} \quad \cdots \quad \frac{\partial \Delta}{\partial \mathrm{y}_\mathrm{N}}\right]$

Το διάνυσμα γραμμής $\frac{\partial \Delta}{\partial \mathbf{y}}$ συμβολίζεται και ως $\nabla_{\mathbf{y}}\Delta$ και διαβάζεται ανάδελτα (namda) του Δ ως προς $\mathbf{y}$.

Αν $\Delta(\mathbf{y}^{\mathrm{z}}, \mathbf{y}) = \frac{1}{2}\mathrm{D}_{\mathrm{E}}^2(\mathbf{y}^{\mathrm{z}}, \mathbf{y}) \Rightarrow \delta^{\mathrm{Z}} = \frac{1}{2}(\mathbf{y}^{\mathrm{z}} - \mathbf{y})^{\mathrm{T}} \cdot (\mathbf{y}^{\mathrm{z}} - \mathbf{y}) \quad = \frac{\partial \Delta}{\partial \mathbf{y}^{\mathrm{Z}}} = (\mathbf{y}^{\mathrm{z}} - \mathbf{y})^{\mathrm{T}}$

Έστω $\mathbf{y}=\mathbf{y}(\boldsymbol{\sigma})$ με $\mathbf{y} \in \mathcal{R}^{\mathrm{N}\,\mathrm{x}\,1}$ και $\boldsymbol{\sigma} \in \mathcal{R}^{\mathrm{M}\,\mathrm{x}\,1}$ τότε

$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\sigma}} = \underbrace{\left[\frac{\vartheta y_\nu}{\vartheta \sigma_\mu}\right]}_{N\times \mathrm{M}} = \begin{bmatrix} \frac{\vartheta y_1}{\vartheta \sigma_1} & \cdots & \frac{\vartheta y_1}{\vartheta \sigma_\mathrm{M}} \\ \vdots & \ddots & \vdots \\ \frac{\vartheta y_N}{\vartheta \sigma_1} & \cdots & \frac{\vartheta y_N}{\vartheta \sigma_\mathrm{M}} \end{bmatrix}$ που λέγεται Ιακωβιανός πίνακας και συμβολίζεται ως

$\boldsymbol{J_y(\sigma)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\sigma}}$

Παράδειγμα 1: Έστω $\boldsymbol{y} \in \mathcal{R}^{3\,x\,1}, \boldsymbol{\sigma} \in \mathcal{R}^{3x\,1}$ και f(.) βαθμωτή συνάρτηση.

$$\boldsymbol{y} = [f(\sigma_\nu)]_{\nu=1}^{3} = \begin{bmatrix} f(\sigma_1) \\ f(\sigma_2) \\ f(\sigma_3) \end{bmatrix} \Rightarrow \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\sigma}} = \begin{bmatrix} \frac{\partial y_1}{\partial \sigma_1} & \frac{\partial y_1}{\partial \sigma_2} & \frac{\partial y_1}{\partial \sigma_3} \\ \frac{\partial y_2}{\partial \sigma_1} & \frac{\partial y_2}{\partial \sigma_2} & \frac{\partial y_2}{\partial \sigma_3} \\ \frac{\partial y_3}{\partial \sigma_1} & \frac{\partial y_3}{\partial \sigma_2} & \frac{\partial y_{32}}{\partial \sigma_3} \end{bmatrix} = \begin{bmatrix} f'(\sigma_1) & 0 & 0 \\ 0 & f'(\sigma_2) & 0 \\ 0 & 0 & f'(\sigma_3) \end{bmatrix}$$

Έστω τώρα $\mathbf{D} = [\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3]$ τότε

$$[\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3] \cdot \begin{bmatrix} f'(\sigma_1) & 0 & 0 \\ 0 & f'(\sigma_2) & 0 \\ 0 & 0 & f'(\sigma_3) \end{bmatrix} = [\mathrm{d}_1 \cdot f'(\sigma_1),\ \mathrm{d}_2 \cdot f'(\sigma_2),\ \mathrm{d}_3 \cdot f'(\sigma_3)] =$$

$$[\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3] \odot [\mathrm{f}'(\sigma_1), \mathrm{f}'(\sigma_2), \mathrm{f}'(\sigma_3)] = \mathbf{D} \odot \mathrm{diag}_\mathrm{r}(\frac{\partial \mathbf{y}}{\partial \boldsymbol{\sigma}})$$

Το σύμβολο $\odot$ συμβολίζει τον πολλαπλασιασμό στοιχείο προς στοιχείο δύο πινάκων ίσων διαστάσεων (γινόμενο Hadamar) και $\mathrm{diag}_\mathrm{r}(.)$ η διαγώνιος τετραγωνικού πίνακα ως διάνυσμα γραμμής.

Για τον λόγο αυτό στο επίπεδο εξόδου Ζ ισχύει

$\boldsymbol{\delta}^{Z} = \frac{\partial \Delta}{\partial \mathbf{y}^{\mathrm{Z}}} \cdot \frac{\partial \mathbf{y}^{\mathrm{Z}}}{\partial \boldsymbol{\sigma}^{Z}} = \left[\frac{\partial \Delta}{\partial \mathrm{y}_1} \cdot \mathrm{f}'(\sigma_1^{\mathrm{Z}}) \quad \cdots \quad \frac{\partial \Delta}{\partial \mathrm{y}_\nu} \cdot \mathrm{f}'(\sigma_\nu^{\mathrm{Z}}) \quad \cdots \quad \frac{\partial \Delta}{\partial \mathrm{y}_\mathrm{N}} \cdot \mathrm{f}'(\sigma_\nu^{\mathrm{Z}})\right] \Rightarrow$

$$\boldsymbol{\delta}^{Z} = \left[\frac{\partial \Delta}{\partial y_1^{\mathrm{Z}}} \quad \cdots \quad \frac{\partial \Delta}{\partial \mathrm{y}_{\mathrm{N_Z}}^{\mathrm{Z}}}\right] \odot \left[f'(\sigma_1^{Z}), \cdots, f'(\sigma_{\mathrm{N_Z}}^{Z})\right] = \frac{\partial \Delta}{\partial \mathbf{y}^{\mathrm{Z}}} \odot \mathrm{diag}_\mathrm{r}(\frac{\partial \mathbf{y}^{\mathrm{Z}}}{\partial \boldsymbol{\sigma}^{Z}}) = \frac{\partial \Delta}{\partial \mathbf{y}^{\mathrm{Z}}} \odot [f'(\sigma_\nu^{\mathrm{Z}})]_{\boldsymbol{\nu=1}}^{\mathbf{N_Z}}$$

με $[f'(\sigma_\nu^{\mathrm{Z}})]_{\boldsymbol{\nu=1}}^{\mathbf{N_Z}}$διάνυσμα γραμμής, (3.3.2)

Παράδειγμα 2: Έστω σ=W•y με $\boldsymbol{\sigma} \in \mathcal{R}^{2x1}$, $\boldsymbol{y} \in \mathcal{R}^{3x1}$, $\boldsymbol{W} \in \mathcal{R}^{2x3}$ τότε

$$\frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\vartheta\sigma_1}{\vartheta y_1} & \frac{\vartheta\sigma_1}{\vartheta y_2} & \frac{\vartheta\sigma_1}{\vartheta y_3} \\ \frac{\vartheta\sigma_2}{\vartheta y_1} & \frac{\vartheta\sigma_2}{\vartheta y_2} & \frac{\vartheta\sigma_2}{\vartheta y_3} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{12} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} = \mathbf{W}$$

Για τον λόγο αυτό στα κρυφά επίπεδα όπου $\boldsymbol{\sigma}^{\zeta+1}=\mathbf{W}^{\zeta+1}\bullet\mathbf{y}^{\zeta}$ ισχύσει:

$\frac{\partial \boldsymbol{\sigma}^{\zeta+1}}{\partial \mathbf{y}^{\zeta}} = \mathbf{W}^{\zeta+1}$, επειδή $\boldsymbol{\delta}^{\zeta+1} \cdot \frac{\partial \boldsymbol{\sigma}^{\zeta+1}}{\partial \mathbf{y}^{\zeta}}$ διάνυσμα γραμμής τότε

$$\boldsymbol{\delta}^{\zeta} = \boldsymbol{\delta}^{\zeta+1} \cdot \frac{\partial \boldsymbol{\sigma}^{\zeta+1}}{\partial \mathbf{y}^{\zeta}} \cdot \frac{\partial \mathbf{y}^{\zeta}}{\partial \boldsymbol{\sigma}^{\zeta}} = \boldsymbol{\delta}^{\zeta+1} \cdot \mathbf{W}^{\zeta+1} \odot \left[f'\left(\sigma_v^{\zeta}\right)\right]_{\mathbf{v=1}}^{\mathbf{N_Z}} \qquad (3.3.3)$$

με $\left[f'(\sigma_v^{Z})\right]_{\mathbf{v=1}}^{\mathbf{N_Z}}$ διάνυσμα γραμμής.

Για $\boldsymbol{\sigma}=\boldsymbol{\sigma}(\mathbf{W})$, $\boldsymbol{\sigma} \in \mathcal{R}^{N_s \times 1}$, $\mathbf{W} \in \mathcal{R}^{N_r \times N_c}$,

$$\frac{\partial \boldsymbol{\sigma}}{\partial \mathbf{W}} = \underbrace{\left[\frac{\vartheta\sigma_s}{\vartheta w_{rc}}\right]}_{N_s \times N_r \times N_c} = \mathbf{T}$$

Ο $\mathbf{T}$ είναι τρισδιάστατος πίνακα $N_s \times N_r \times N_c$ και λέγεται τανυστής δεδομένων(data tensor). Οι δείκτες r και c είναι, ως γνωστόν, τα πρώτα γράμματα των λέξεων rows και columns. Ο δείκτης s είναι το πρώτο γράμμα της λέξεως slice (φέτα) και αριθμεί τους πίνακες του τανυστή. Μπορούν βέβαια να χρησιμοποιηθούν και άλλα γράμμα ως δείκτες όμως σημασία έχει η σειρά τους (slice, row, column). Για παράδειγμα το στοιχείο $T_{213} = \frac{\partial \sigma_2}{\partial w_{13}}$. Από τούδε θα χρησιμοποιείται ο όρος tensor όταν το πλήθος των διαστάσεων είναι μεγαλύτερο του δύο.

Παράδειγμα 3

$\boldsymbol{\sigma} \in \mathcal{R}^{N_s \times 1}, \mathbf{W} \in \mathcal{R}^{N_r \times N_c}$

s=1…N_s, r=1…N_r, c=1…N_c

N_s=2, N_r=2, N_c=3

$\boldsymbol{\sigma}(\mathbf{W})$=W•y. ήτοι

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\frac{\partial \boldsymbol{\sigma}}{\partial \mathbf{W}} = \underbrace{\left[\frac{\vartheta\sigma_s}{\vartheta w_{rc}}\right]}_{N_s \times N_r \times N_c} = \mathbf{T}$$

$$\underbrace{\mathbf{T}}_{2\times2\times3} = \begin{bmatrix} \frac{\partial\sigma_2}{\partial w_{11}} & \frac{\partial\sigma_2}{\partial w_{12}} & \frac{\partial\sigma_2}{\partial w_{13}} \\ \frac{\partial\sigma_2}{\partial w_{21}} & \frac{\partial\sigma_2}{\partial w_{22}} & \frac{\partial\sigma_2}{\partial w_{23}} \end{bmatrix} \begin{bmatrix} \frac{\partial\sigma_1}{\partial w_{11}} & \frac{\partial\sigma_1}{\partial w_{12}} & \frac{\partial\sigma_1}{\partial w_{13}} \\ \frac{\partial\sigma_1}{\partial w_{21}} & \frac{\partial\sigma_1}{\partial w_{22}} & \frac{\partial\sigma_1}{\partial w_{23}} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \\ 0 & 0 & 0 \end{bmatrix}$$

Το γινόμενο δύο tensors A,B, με τον τρόπο που ορίσθηκε το γινόμενο των πινάκων απαιτεί ο τελευταίος δείκτης του A να παίρνει το ίδιο πλήθος τιμών με τον πρώτο δείκτη του B ή όπως θα λέγαμε να είναι κοινός. Η μαθηματική έκφραση είναι:

$$\mathbf{A}\cdot\mathbf{B}=\mathbf{C}\ \mu\varepsilon\ c_{i_1\dots i_N,k_1\dots k_M}=\sum_j a_{i_1\dots i_N,j}\cdot b_{j,k_1\dots k_M}$$

Ο tensor A έχει N διαστάσεις, ο B έχει M διαστάσεις, ο C έχει N+M-2, διαστάσεις. Το πλήθος των διαστάσεων ενός τανυστή λέγεται τάξη αυτού και γράφεται συνήθως επάνω και δεξιά του, π.χ. $\mathbf{T}^{(3)}$. Τα βαθμωτά μεγέθη έχουν τάξη 0, τα διανύσματα 1 και οι πίνακες 2.

Παράδειγμα 4

$$\boldsymbol{\delta}=[\delta_1,\delta_2],\qquad \mathbf{T}=\begin{matrix}\begin{bmatrix}0&0&0\\y_1&y_2&y_3\end{bmatrix}\\ \begin{bmatrix}y_1&y_2&y_3\\0&0&0\end{bmatrix}\end{matrix}$$

$$\boldsymbol{\delta}\cdot\mathbf{T}=\sum_j \delta_j\cdot t_{jrc}=\sum_{j=1}^{2}\delta_j\cdot t_{jrc}=\delta_1\cdot t_{1rc}+\delta_2\cdot t_{2rc}==\delta_1\cdot\begin{bmatrix}y_1&y_2&y_3\\0&0&0\end{bmatrix}+\delta_2\cdot\begin{bmatrix}0&0&0\\y_1&y_2&y_3\end{bmatrix}$$

$$=\begin{bmatrix}\delta_1\cdot y_1&\delta_1\cdot y_2&\delta_1\cdot y_3\\\delta_2\cdot y_1&\delta_2\cdot y_2&\delta_2\cdot y_3\end{bmatrix}=\begin{bmatrix}\delta_1\\\delta_1\end{bmatrix}\cdot[y_1\quad y_2\quad y_3]=\boldsymbol{\delta}^{\mathbf{T}}\cdot\mathbf{y}^{\mathbf{T}}$$

Για το λόγο αυτό

$$\frac{\partial\boldsymbol{\Delta}}{\partial\mathbf{W}^{\zeta}}=\left(\boldsymbol{\delta}^{\zeta}\right)^{\mathrm{T}}\cdot\left(\mathbf{y}^{\zeta-1}\right)^{\mathrm{T}}\quad(3.3.4)$$

Από τις σχέσεις (3.3.2),(3,3,3),(3.3.4) σύμφωνα με την μέθοδο κατάβασης κατά την κλίση (GD) η διόρθωση των βαρών των νευρώνων ενός επιπέδου δίνεται από τις σχέσεις:

$$\mathbf{W}^{\zeta}(t+1)=\mathbf{W}^{\zeta}(t)-\rho\sum_{\forall(\mathbf{x},\mathbf{y})}\left(\boldsymbol{\delta}^{\zeta}\right)^{\mathrm{T}}\cdot\left(\mathbf{y}^{\zeta-1}\right)^{\mathrm{T}}$$

$\boldsymbol{\delta}^{Z}=(\mathbf{y}^{z}-\mathbf{y})^{\mathrm{T}}\odot[f'(\sigma_v^Z)]_{\mathbf{v=1}}^{\mathbf{N_Z}}$ για το επίπεδο εξόδου

$\boldsymbol{\delta}^{\zeta}=\boldsymbol{\delta}^{\zeta+1}\cdot\mathbf{W}^{\zeta+1}\odot\left[f'\left(\sigma_v^{\zeta}\right)\right]_{\mathbf{v=1}}^{\mathbf{N_Z}}$ για τα κρυφά επίπεδα

Η μέθοδος μπορεί να περιγραφεί μαθηματικά υπολογίζοντας τις μερικές παραγώγους ανά γραμμή των βαρών ενός νευρώνα ή ακόμη για κάθε βάρος, π.χ. για το ν-οστό νευρώνα του ζ-επιπέδου από τα παραπάνω δίνεται από τις σχέσεις.

$$\mathbf{w}_v^{\zeta}(t+1)=\mathbf{w}_v^{\zeta}(t)-\rho\sum_{\forall(\mathbf{x},\mathbf{y})\in V}\delta_v^{\zeta}\cdot(\mathbf{y}^{\zeta-1})^{\mathrm{T}}$$

$$\delta_v^{Z}=(y_v^{Z}-y_v)\cdot f'\left(\sigma_v^{Z}\right)\ \text{για το επίπεδο εξόδου}$$

$$\delta_\nu^\zeta = \boldsymbol{\delta}^{\zeta+1} \cdot \mathbf{W}_{:\nu}^{\zeta+1} \cdot f'\left(\sigma_\nu^\zeta\right) = \left(\sum_{\mu=1}^{N_{\zeta+1}} \delta_\mu^{\zeta+1} \cdot w_{\mu\nu}^{\zeta+1}\right) \cdot f'\left(\sigma_\nu^\zeta\right)$$ για τα κρυφά επίπεδα

Υπενθυμίζουμε ότι αν y=f(σ) είναι η σιγμοειδής λογιστική συνάρτηση, $f'(\sigma) = a \cdot y \cdot (1-y)$

Υπενθυμίζουμε ακόμη ότι η διόρθωση μπορεί να γίνεται αφού προστεθούν όλα τα Δ_i, ή κατά ομάδες (bathes:δέσμες) ή για κάθε πρότυπο.

Η χρήση πινάκων και tensors απλοποιεί την μαθηματική γραφή και συμβάλλει στην προγραμματιστική υλοποίηση. Για παράδειγμα η βιβλιοθήκη TensorFlow διαθέτει τις σχετικές συναρτήσεις και δομές για μια τέτοια υλοποίηση και παράλληλη εκτέλεση τους με την αρχιτεκτονική λογισμικού παράλληλης επεξεργασίας CUDA (Compute Unified Device Architecture) σε GPU (Graphics Processing Unit) επεξεργαστή. Ο GPU είναι επεξεργαστής (Hardware) που φέρουν οι κάρτες NVIDIA. Η διαδοχή είναι:

TensorFlow → CUDA → NVIDIA Driver → NVIDIA GPU (Hardware)

Η ταχύτητα υπολογισμών API με την χρήση των παραπάνω τεχνολογιών είναι μέχρι της τάξεως 50:1 σε σύγκριση με αυτή της χρήσης μόνο του κεντρικού επεξεργαστή. Αυτό είναι απαραίτητο για μεγάλα δίκτυα π.χ. CNN που θα παρουσιάσουμε σε επόμενο κεφάλαιο και εκατομμύρια βάρη. Η TensorFlow είναι γραμμένη κυρίως σε γλώσσα C και C++. Εν τούτοις για ευκολία όσων προγραμματίζουν σε υψηλού επιπέδου γλώσσες (φιλικές στον προγραμματιστή) χρησιμοποιούνται εντολές της γλώσσας Python ως διεπαφές που καλούν τον εκάστοτε αντικειμενικό κώδικα της βιβλιοθήκης της TensorFlow.

Σύγκριση τεχνικών χαρακτηριστικών CPU με GPU

CPU (Κεντρική Μονάδα Επεξεργασίας)	GPU (Μονάδα Επεξεργασίας Γραφικών)
Ισχυροί πυρήνες (συνήθως 4–64)	Εκατοντάδες έως χιλιάδες απλούστεροι πυρήνες
Βελτιστοποιημένη για σειριακές εργασίες	Βελτιστοποιημένη για μαζικά παράλληλες εργασίες
Χαμηλή απόδοση διεκπεραίωσης (throughput), μικρή καθυστέρηση (latency)	Υψηλή απόδοση διεκπεραίωσης (throughput), μεγαλύτερη καθυστέρηση (latency)
Επεξεργασία γενικού σκοπού	Παράλληλοι αριθμητικοί υπολογισμοί

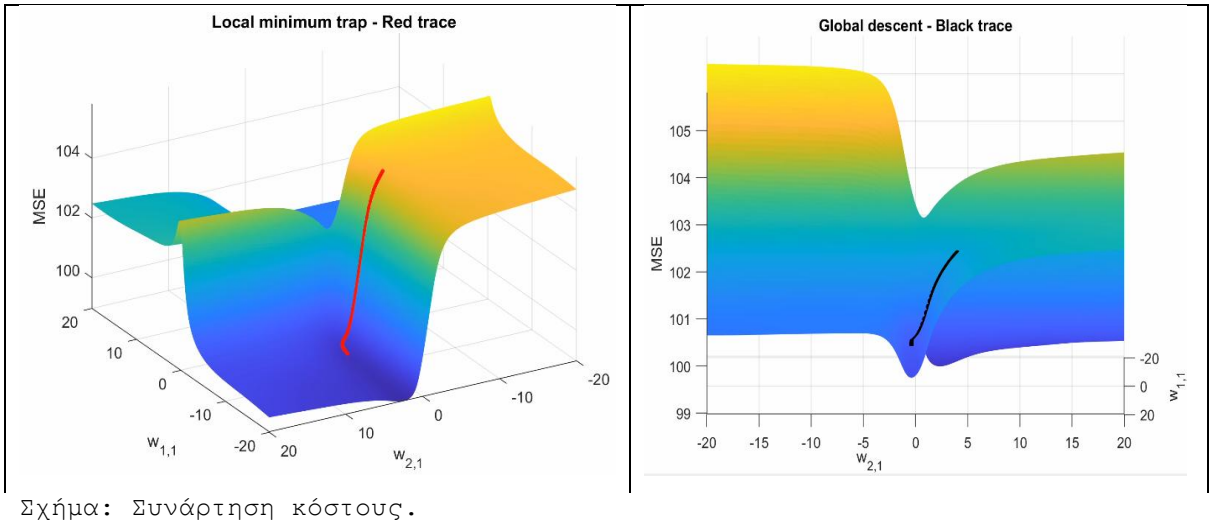
Η σχέση $\boldsymbol{\sigma}^\zeta = \mathbf{W}^\zeta \mathbf{y}^{\zeta-1}$ αποτελεί έναν αφινικό (affine) μετασχηματισμό που μετασχηματίζει τον χώρο των **y**, σύμφωνα με τις τιμές του πίνακα **W**. Γενικά οι ευθείες γραμμές διατηρούνται καθώς και οι αναλογίες και η παραλληλία μεταξύ ευθυγράμμων τμημάτων. Δεν διατηρούνται γωνίες ή αποστάσεις (εκτός αν ο **W** είναι ορθογώνιος πίνακας και det(A)=1). Ακολούθως η μη γραμμική συνάρτηση ενεργοποίησης (activation function) **y**=**y**(**σ**) στρεβλώνει τον χώρο των **σ** έτσι που τα μη γραμμικά προβλήματα να μετατρέπονται σε γραμμικά. Για να επιτευχθεί αυτό βέβαια πρέπει πρώτα να είναι κατάλληλες οι τιμές του πίνακα **W**. Αυτό θα επιτευχθεί με την εκπαίδευση του δικτύου όπως θα δούμε παρακάτω. Θα δούμε ακόμη την γεωμετρική αναπαράσταση των παραπάνω σε σχετικό παράδειγμα για την πύλη XOR.

Το ΤΝΔ ενός MLP είναι ένας συνδυασμούς από σύνθετες συναρτήσεις και πολλαπλασιασμούς πινάκων. Αν ζ=1,2...Z αριθμεί τα επίπεδα, η έξοδος θα είναι

$\mathbf{y}^{z}(\mathbf{W}^{1},\cdots,\mathbf{W}^{\zeta},\cdots,\mathbf{W}^{Z},\mathbf{x}) = \mathbf{y}^{z}(\mathbf{W}^{z}\cdot\mathbf{y}^{z-1}(\mathbf{W}^{z-1}\cdot\mathbf{y}^{z-2}(\mathbf{W}^{z-2}\cdots\mathbf{y}^{\zeta}(\mathbf{W}^{\zeta}\mathbf{y}^{\zeta-1}(\cdots\mathbf{y}^{3}(\mathbf{W}^{2}\mathbf{y}^{1}(\mathbf{W}^{1}\cdot\mathbf{x}))\cdots)$ όπου $\mathbf{W}^{\zeta}$ πίνακας με γραμμές τα βάρη των νευρώνων του ζ επιπέδου.

Παραδείγματα εφαρμογής του δικτύου υπάρχουν στο παράρτημα των εργαστηριακών ασκήσεων.

Το ΝΔ MLP με Back error propagation είναι από τα πλέον χρησιμοποιηθέντα ΝΔ σε πληθώρα εφαρμογών από διαφορετικές επιστημονικές περιοχές. Στα κρυφά επίπεδά του προσδιορίζονται ουσιαστικά αυτόματα τα αποτελεσματικά χαρακτηριστικά για την ταξινόμηση. Το πλήθος των κλάσεων μπορεί να είναι το ίδιο με αυτό των ανυσμάτων εισόδου του συνόλου εκπαίδευσης και έτσι το ΝΔ να προσεγγίζει συναρτήσεις των ανυσμάτων εισόδου στα ανύσματα εξόδου. Το σημαντικότερο μειονέκτημα του back propagation είναι ο χρόνος ολοκλήρωσης της εκπαίδευσης του, (χρόνος σύγκλισης). Είναι δυνατόν να χρειασθούν εκατοντάδες χιλιάδες επαναλήψεις εωσότου συγκλίνει ακόμη και για σχετικά απλές εφαρμογές. Σε κάποιες εφαρμογές χρειάσθηκαν μερικές μέρες για την σύγκλιση του συστήματος. Ο εγκλωβισμός της διαδικασίας σύγκλισης σε τοπικά ελάχιστα της συνάρτησης κόστους είναι ένα επιπρόσθετο πρόβλημα. Πρόβλημα επίσης εμφανίζεται όταν το μέτρο της κλίσης έχει πολύ μικρές τιμές σε κάποια επίπεδα του δικτύου επειδή η συνάρτηση κόστους είναι σχεδόν σταθερή (επίπεδη) και η σύγκλιση αργεί (vanishing gradient problem), ή όταν για κάποια επίπεδα μεταβάλλεται ραγδαία (exploding gradient problem).



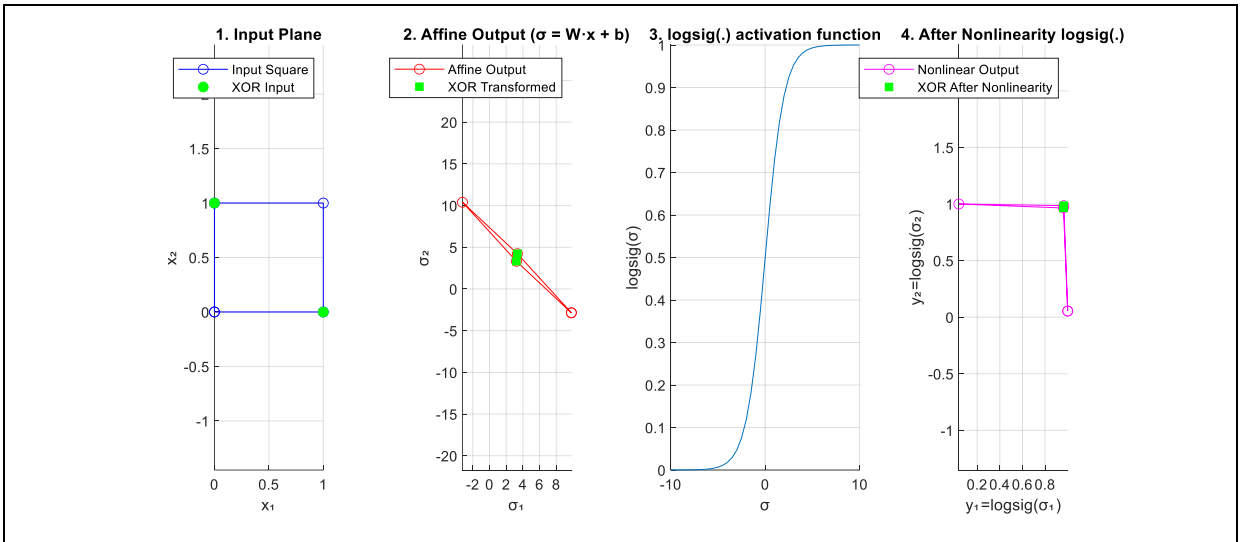
Σχήμα: Συνάρτηση κόστους.

(α) Εύρεση ολικού ακρότατου (β)Εγκλωβισμός σε τοπικό ακρότατο

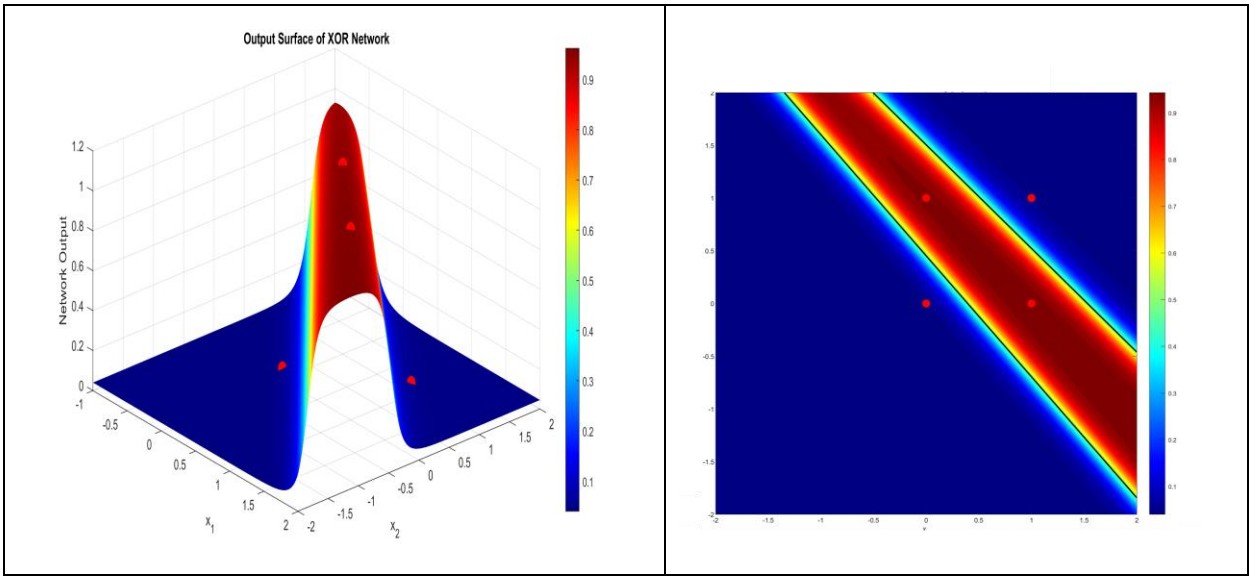
Για παράδειγμα για την εκμάθηση της πύλης XOR δομούμε έναν MLP με δύο νευρώνες εισόδου συνάρτηση ενεργοποίησης την σιγμοειδή λογιστική και έναν νευρώνα εξόδου χωρίς συνάρτηση ενεργοποίησης. Μετά την εκπαίδευση ο πίνακας των νευρώνων εισόδου $\mathbf{W}^{1}$ έχει τιμές βαρών

$\mathbf{W}^{1} = \begin{bmatrix} -6.47 & -6.58 & +9.82 \\ +7.08 & +6.17 & -2.85 \end{bmatrix}$ και ο εξόδου $\mathbf{W}^{2} = [6.44 \quad 6.56 \quad -9.62]$

Η σχέση $\boldsymbol{\sigma = W^1 \cdot x}$ είναι ένα αφινικός μετασχηματισμός που όμως αν και μετατοπίζει τις θέσεις των **x** στις θέσεις **σ** δεν δίδει γραμμικά διαχωρίσιμες τιμές. Αυτό το κάνει η εφαρμογή της σιγμοειδούς λογιστικής συνάρτησης που



στρεβλώνει τον χώρο παράγοντας τις τιμές **y** που είναι γραμμικά διαχωρίσιμές. Στο παρακάτω σχήμα φαίνονται οι μετασχηματισμοί των χώρων. Στο παρακάτω σχήμα φαίνεται η συνάρτηση της τελικής εξόδου y^2 με πεδίο ορισμού το αρχικό επίπεδο (x_1, x_2) και η κάτοψη της όπου φαίνονται τα όρια διαχωρισμού των κλάσεων με μαύρες γραμμές.



Η διόρθωση των βαρών για κάθε πρότυπο $\mathbf{x}_i$ του συνόλου εκπαίδευσης με τυχαία επιλογή ονομάζεται ***στοχαστική κάθοδο κατά την κλίση*** (***stochastic gradient descent, SGD***). Οι μέθοδοι αυτοί είναι αργές και απαιτούν πολλές επαναλήψεις. Μία βελτίωση τους εύκολα υλοποιήσιμη είναι η πρόσθεση ενός ακόμη όρου στην σχέση διόρθωσης που ονομάζουμε όρο ***ορμής*** ή ***ροπής*** (***momentum term***). Συγκεκριμένα

η κατάβαση κατά την κλίση για μια συνάρτηση κόστους $\mathcal{L}(\boldsymbol{w}_\nu^\zeta)$ δίνεται σύμφωνα με όσα προαναφέρθηκαν από τις σχέσεις:

$$\boldsymbol{w}_\nu^\zeta(t+1) = \boldsymbol{w}_\nu^\zeta(t) - \rho \left.\frac{\partial K}{\partial \boldsymbol{w}_\nu^\zeta}\right|_{\boldsymbol{w}_\nu^\zeta(t)}$$

Στην διόρθωση με ορμή

$$\boldsymbol{w}_\nu^\zeta(t+1) = \boldsymbol{w}_\nu^\zeta(t) - \rho \left.\frac{\partial K}{\partial \boldsymbol{w}_\nu^\zeta}\right|_{\boldsymbol{w}_\nu^\zeta(t)} + \alpha(\boldsymbol{\Delta w}_\nu^\zeta(t))$$

$$\boldsymbol{\Delta w}_\nu^\zeta(t) = \boldsymbol{w}_\nu^\zeta(t) - \boldsymbol{w}_\nu^\zeta(t-1)$$

$\alpha \in (0\,1)$, με συνήθεις τιμές από 0.2 έως 0.8.

Προστίθεται δηλαδή και ένα μέρος της προηγούμενης μεταβολής των βαρών του νευρώνα. Η ποσότητα $\alpha\boldsymbol{\Delta w}_\nu^\zeta(t)$ είναι ο όρος της ορμής (***momentum term***). Η χρήση του μειώνει το πλήθος των επαναλήψεων αυξάνοντας το βήμα διόρθωσης των βαρών κατά την κάθοδο όταν η συνάρτηση κόστους μεταβάλλεται αργά, δηλαδή το πλάτος της βάθμωσης ($|\nabla K_{\boldsymbol{w}_\nu^\zeta(t)}|$) είναι πολύ μικρό. Επίσης εξομαλύνει τις απότομες αλλαγές της διεύθυνσης της διόρθωσης. Αυτές οι ιδιότητες της ορμής φαίνονται αλγεβρικά αν υπολογίσουμε της διόρθωση μετά από Τ επαναλήψεις από μία αρχική τιμή $\mathbf{w}(0)$ των βαρών κάποιου νευρώνα (τα σύμβολα ν, ζ έχουν παραληφθεί για απλότητα). Θέτοντας $\boldsymbol{G}(t) = -\rho\nabla K_{\boldsymbol{w}_\nu^\zeta(t)}$ θα ισχύει:

$$\boldsymbol{\Delta w}(T) = -\rho\boldsymbol{G}(T) + \alpha\cdot\boldsymbol{\Delta w}(T-1)$$

$$\boldsymbol{\Delta w}(T-1) = -\rho\boldsymbol{G}(T-1) + \alpha\cdot\boldsymbol{\Delta w}(T-2) \Rightarrow \alpha\boldsymbol{\Delta w}(T-1) = -\rho\alpha\boldsymbol{G}(T-1) + \alpha^2\cdot\boldsymbol{\Delta w}(T-2)$$

$$\boldsymbol{\Delta w}(T-2) = -\rho\boldsymbol{G}(T-2) + \alpha\cdot\boldsymbol{\Delta w}(T-3) \Rightarrow \alpha^2\boldsymbol{\Delta w}(T-2) = -\rho\alpha^2\boldsymbol{G}(T-2) + \alpha^3\cdot\boldsymbol{\Delta w}(T-3)$$

$$\vdots$$

$$\boldsymbol{\Delta w}(2) = -\rho\boldsymbol{G}(2) + \alpha\cdot\boldsymbol{\Delta w}(1) \Rightarrow \alpha^{T-2}\boldsymbol{\Delta w}(2) = -\rho\alpha^{T-2}\boldsymbol{G}(2) + \alpha^{T-1}\cdot\boldsymbol{\Delta w}(1)$$

$$\boldsymbol{\Delta w}(1) = -\rho\boldsymbol{G}(t) + \alpha\cdot\boldsymbol{\Delta w}(0) \Rightarrow \alpha^{T-1}\boldsymbol{\Delta w}(1) = -\rho\alpha^{T-1}\boldsymbol{G}(1) + \alpha^T\cdot\boldsymbol{\Delta w}(0)$$

Προσθέτοντας τις εξισώσεις μετά της συνεπαγωγές προκύπτει:

$$\boldsymbol{\Delta w}(T) = -\rho\sum_{t=0}^{T-1}\alpha^t\,\boldsymbol{G}(T-t) + \alpha^T\cdot\boldsymbol{\Delta w}(0)$$

Μετά από μερικές επαναλήψεις, επειδή 0<α<1, η μεταβολή $\boldsymbol{\Delta w}(T)$ επηρεάζεται από τις προηγούμενες τιμές της $\boldsymbol{G}$ που φθίνουν κατά α^t και ο όρος $\alpha^T\cdot\boldsymbol{\Delta w}(0)$ γίνεται αμελητέος. Αν η $\mathbf{G}(.)$ στο διάστημα αυτό των επαναλήψεων έχει σταθερή τιμή $\mathbf{G}$ τότε

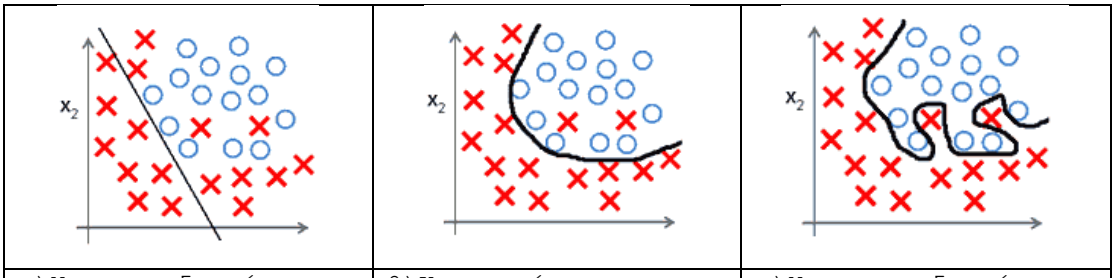
$$\boldsymbol{\Delta w}(T) = -\rho(1+\alpha+\alpha^2+\cdots+\alpha^{T-1})\boldsymbol{G} = -\boldsymbol{\rho}\frac{\mathbf{1}-\boldsymbol{\alpha}^{\boldsymbol{T}}}{\mathbf{1}-\boldsymbol{\alpha}}\boldsymbol{G} \approx -\frac{\boldsymbol{\rho}}{\mathbf{1}-\boldsymbol{\alpha}}\boldsymbol{G}$$

Για ρ=0.2 και α=0.5 η αλλαγή των τιμών των βαρών θα είναι $0.4\boldsymbol{G}$ ενώ χωρίς τον όρο ορμής θα ήταν $0.2\boldsymbol{G}$. Όταν η μέθοδος κατάβασης της κλίσης με όρο ορμής (Gradient Descent with Momentum, ***GDM***), εφαρμόζεται με διόρθωση ανά πρότυπο με τυχαία επιλογή και τιτλοφορείται ως Stochastic Gradient Descent with Momentum, ***SGDM***).

Γενικά στα νευρωνικά δίκτυα το πλήθος των επιπέδων και των νευρώνων των είναι ένα ζήτημα που μας απασχολεί. Δεν υπάρχει ένας γενικός τύπος-αλγόριθμος που να μας οδηγεί σε ένα βέλτιστο πλήθος παραμέτρων-βαρών. Ένας μεγάλος αριθμός

βαρών αυξάνει το υπολογιστικό κόστος και μπορεί να οδηγήσει σε λεπτομερή εκμάθηση του συνόλου εκπαίδευσης ακόμη και του θορύβου των μετρήσεων, εις βάρος της γενικότητας του, δηλαδή της επιτυχούς ταξινόμησης νέων αγνώστων προτύπων. Αυτό το φαινόμενο το ονομάζουμε *υπερεκπαίδευση*. Μικρός αριθμός παραμέτρων μπορεί να οδηγήσει σε αδυναμία εκπαίδευσης ή χαμηλή απόδοση-ακρίβεια ταξινόμησης. Την περίπτωση αυτή αναφέρουμε ως *υποεκπαίδευση*. Στο Σχήμα 3.3.5 φαίνεται η διαχωριστική καμπύλη τριών συστημάτων ενός προβλήματος δύο διαστάσεων για τις περιπτώσεις που αναφέραμε.

Μία απλή μέθοδος είναι να ξεκινάμε με ένα μικρό πλήθος παραμέτρων και να το αυξάνουμε προοδευτικά μέχρι να έχουμε ικανοποιητικό αποτέλεσμα. Εκείνο που μπορεί να συμβεί επίσης είναι το φαινόμενο της υπερ-εκπαίδευσης. Τότε έχουμε υψηλή απόδοση εκπαίδευσης για τα πρότυπα που εκπαιδεύτηκε το δίκτυο και χαμηλή για τα πρότυπα που δεν εκπαιδεύτηκε (σαν να «παπαγαλίζει»). Για να εντοπίσουμε νωρίς το πρόβλημα χωρίζουμε το αρχικό σύνολο S των διαθέσιμων δεδομένων σε δύο μέρη. Το ένα χρησιμοποιείται ως σύνολο εκπαίδευσης (Training set, T) και το άλλο ως σύνολο επικύρωσης του αποτελέσματος (Validation set, V). Θα πρέπει το ποσοστό επιτυχίας επιτυχούς ταξινόμησης (accuracy) να είναι υψηλό με παραπλήσιες τιμές και για τα δύο σύνολα. Σε κάποια δίκτυα από το σύνολο επικύρωσης συνάγονται τιμές παραμέτρων που καθοδηγούν τον αλγόριθμο βελτιστοποίησης. Τέλος ένα τρίτο σύνολο δοκιμής (test set) που αποτελείται από πρότυπα που δεν χρησιμοποιήθηκαν αξιολογεί το τελικό αποτέλεσμα.



α) Υποεκπαιδευμένο σύστημα	β) Κανονικά εκπαιδευμένο σύστημα	γ) Υπερεκπαιδευμένο σύστημα

Σχήμα 3.3.5

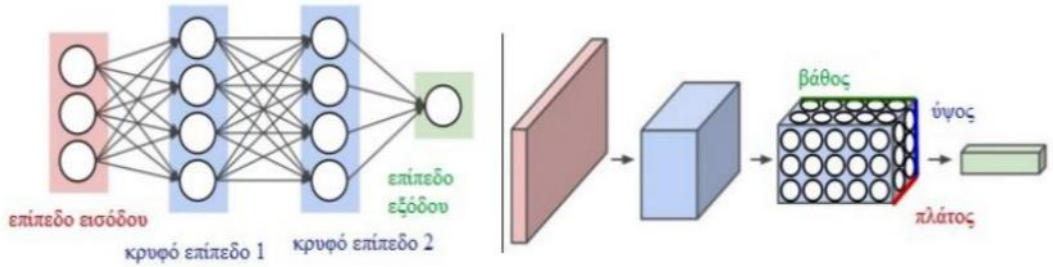
3.4. Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) (CNN:Convolutional Neural Networks) έχουν χρησιμοποιηθεί ευρέως στην αναγνώριση εικόνων και μοιάζουν με τα Τεχνητά Νευρωνικά Δίκτυα πρόσθιας τροφοδότησης (feed-forward networks, FFN). Υπάρχουν διαφορές ανάμεσα σε αυτά τα δύο είδη δικτύων, που καθιστούν τα ΣΝΔ πιο ελκυστικά για χρήση. Πιο συγκεκριμένα στα κλασσικά FFN η κλιμάκωση μεγάλων εικόνων οδηγεί σε μεγάλο μη διαχειρίσιμο πλήθος βαρών. Για παράδειγμα στο σύνολο δεδομένων CIFAR-10 οι εικόνες είναι μεγέθους, μόνο 32 x 32 x 3 (πλάτος × ύψος × χρωματικά κανάλια), και επομένως ένας πλήρως συνδεδεμένος νευρώνας στο πρώτο κρυφό επίπεδο

ενός ΤΝΔ θα είχε 32*32*3 = 3072 βάρη. Παρά το γεγονός ότι αυτός ο αριθμός δείχνει διαχειρίσιμος, για είσοδο εικόνας μεγαλύτερων διαστάσεων, π.χ 200 x 200 x 3 θα είχαμε νευρώνες με 200*200*3 = 120.000 βάρη ο καθένας.

Για περισσότερους νευρώνες ο αριθμός των παραμέτρων θα μεγάλωνε ραγδαία. Συνεπώς η πλήρης συνδεσιμότητα είναι σπάταλη, και μπορεί λόγω των πολλών παραμέτρων να οδηγήσει εύκολα σε *υπερπροσαρμογή* (*overfitting*) του δικτύου. Από την άλλη πλευρά, η χρήση των ΣΝΔ εκμεταλλεύεται το γεγονός ότι η είσοδος αποτελείται από εικόνες, και περιορίζει την αρχιτεκτονική του με πιο έξυπνο τρόπο. Γενικότερα τα ΣΝΔ, διαθέτουν νευρώνες οι οποίοι έχουν 3 διαστάσεις (πλάτος-ύψος-βάθος), και έχουν την ιδιαιτερότητα να συνδέονται διαδοχικά με μία μικρή περιοχή του προηγούμενου επιπέδου αντίθετα με ότι γινόταν σε μία πλήρη σύνδεση. Κατ' ουσίαν ο νευρώνας είναι ένα φίλτρο που *συνελίσεται* στα δεδομένα εισόδου του. Θα πρέπει να διευκρινίσουμε εδώ ότι λέγοντας «συνελίσεται» εννοούμε την εφαρμογή της δισδιάστατης συνέλιξης με κάποια διευρυμένη χρήση. Οι όροι νευρώνας, κόμβος (node), φίλτρο και πυρήνας (kernel)χρησιμοποιούνται εναλλακτικά στην βιβλιογραφία των συνελικτικών δικτύων για το ίδιο πράγμα.

Στο σχήμα 3.4.1 απεικονίζεται αριστερά μία αρχιτεκτονική ενός κλασικού νευρωνικού δικτύου και δεξιά μία αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου. Κάθε επίπεδο ενός ΣΝΔ μετασχηματίζει τον τρισδιάστατο όγκο εισόδου (input volume), σε έναν τρισδιάστατο όγκο εξόδου (output volume. Στο συγκεκριμένο παράδειγμα η έγχρωμη εικόνα του παραπάνω σχήματος αποτελεί το επίπεδο εισόδου επομένως θα έχει βάθος 3, όσα και τα κανάλια μιας RGB εικόνας.



Σχήμα 3.4.1. (Αρχιτεκτονική CNN)

Στα ΣΝΔ εκτελούνται τέσσερις βασικές λειτουργίες:

1. Συνέλιξη (Convolution)
2. Εφαρμογή μη γραμμικότητας (Non Linearity) – ReLU
3. Συγκέντρωση ή Υπο-Δειγματοληψία (Pooling / Sub sampling)
4. Κατηγοριοποίηση με πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer for Classification)

Επίπεδο Συνέλιξης (Convolution)

Θεμελιώδες δομικό επίπεδο ενός Συνελικτικού Νευρωνικού Δικτύου, όπως έχει προαναφερθεί, είναι το επίπεδο στο οποίο συνελίσεται η είσοδος με Ν το πλήθος φίλτρα (νευρώνες βαρών). Το αποτέλεσμα είναι πίνακες που αποτελούν χάρτες χαρακτηριστικών. Το επίπεδο αυτό ονομάζεται συνελικτικό επίπεδο. Τα φίλτρα αυτά στην περίπτωση των έγχρωμων εικόνων είναι τρισδιάστατα, καθώς το μέγεθός τους καθορίζεται από το ύψος *R* (πλήθος γραμμών), το πλάτος (πλήθος στηλών) *C* και το βάθος *D* (πλήθος πινάκων R×C). Στην περίπτωση που η είσοδος είναι μία έγχρωμη εικόνα το βάθος είναι τρία (3) για τα τρία βασικά χρώματα (Red Green Blue: RGB). Για την παραγωγή αυτών των χαρακτηριστικών, σαρώνεται ολόκληρη η εικόνα, πραγματοποιούνται πράξεις εσωτερικού γινομένου μεταξύ των τιμών του φίλτρου και της υποκείμενης περιοχής του πίνακα και στο τέλος εξάγεται το αποτέλεσμα το οποίο τοποθετείται στον χάρτη χαρακτηριστικών.

Η πράξη της *δισδιάστατης συνέλιξης* για δύο διακριτά σήματα x(r,c), w(r,c),(δισδιάστατες ακολουθίες)ορίζεται τυπικά από την σχέση:

$$x * w = \{x * w\}(r,c) = \sum_k \sum_l x(k,l) \cdot w(r-k, c-l) \text{ με } r,c,k,l \in \mathbb{Z}$$

Η πράξη της *δισδιάστατης συσχέτισης* για δύο διακριτά σήματα x(m,n), y(m,n),(δισδιάστατες ακολουθίες)ορίζεται τυπικά από την σχέση:

$$x \circ w = \{x \circ w\}(r,c) = \sum_k \sum_l x(k,l) \cdot w(k-r, l-c) \text{ με } r,c,k,l \in \mathbb{Z}$$

Αν w(k,l)->w(-k,-l) οι δύο πράξεις δίνουν το ίδιο αποτέλεσμα. Για ακολουθίες πεπερασμένου μήκους τα όρια των δεικτών *k,l* καθορίζονται από τις θέσεις πέραν των οποίων οι ακολουθίες έχουν όλες τις τιμές τους μηδενικές.

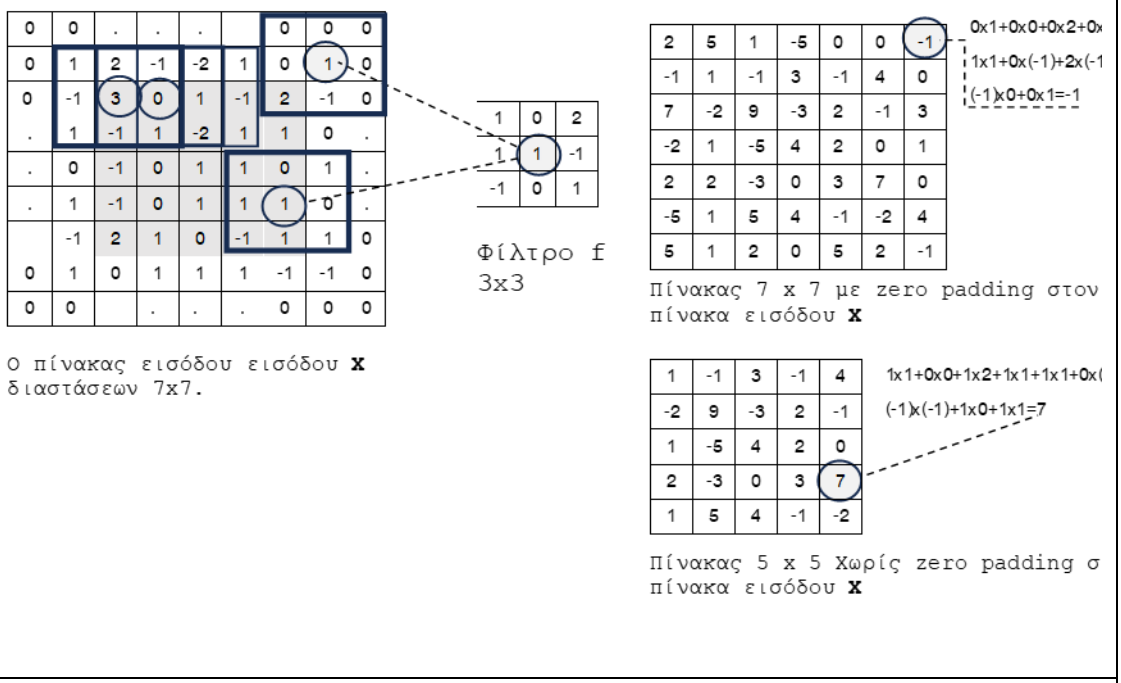
Στα συνελικτικά δίκτυα εφαρμόζεται η πράξη της συσχέτισης. Εν τούτοις ιστορικά επικράτησε να χρησιμοποιείται ο όρος συνέλιξη από την ευρεία χρήση του στην επεξεργασία σημάτων. Θα διατηρήσουμε αυτήν την ορολογία στην συνέχεια.

Ανεξαρτήτως των ορίων μέσα στα οποία ορίζεται η πράξη της συνέλιξης στο $\mathbb{Z} \times \mathbb{Z}$, εκείνο που μας ενδιαφέρει είναι οι τιμές των ακολουθιών και τις θεωρούμε ως στοιχεία πινάκων α) της εισόδου **X** με διαστάσεις (R_x x C_x) και β) του φίλτρου **W** με διαστάσεις (R_w x C_w).

Στο ακόλουθο παράδειγμα (Σχ.3.4.2) φαίνεται η εφαρμογή της πράξης της συνέλιξης (συσχέτισης). Οι τιμές του του φίλτρου *w*(.,.) ολισθαίνουν επάνω στις τιμές των γραμμών και των στηλών της *x*(.,.) και υπολογίζεται το άθροισμα των γινομένων των ομοιόθετων τιμών. Το αποτέλεσμα θα είναι μια ακολουθία πεπερασμένου μήκους με διαστάσεις (R_x+ R_w-1)x(C_x+ C_w-1). Αν θέλουμε οι διαστάσεις της προκύπτουσας ακολουθίας να είναι ίσες με αυτές τις εισόδου, αφαιρούμε R_w-1

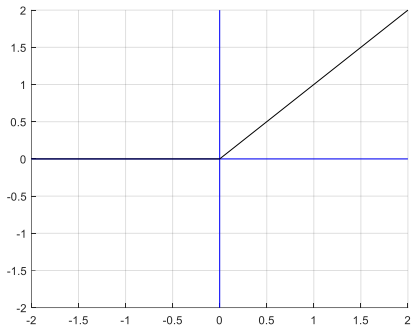
γραμμές και C_w-1 στήλες από τα όριά της. Απλούστερα μπορούμε να ολισθαίνουμε το φίλτρο έτσι ώστε η κεντρική τιμή του φίλτρου *w* να φτάνει στα όρια της *x* και αποδίδοντας μηδενικές τιμές στη *x* εκτός των ορίων της (zero padding).

Είναι βολικό το φίλτρο να έχει περιττό αριθμό γραμμών και στηλών π.χ. 3×3, 5×5 ώστε να υπάρχει κεντρική τιμή. Στο τέλος της πράξης της συνέλιξης σε κάθε τιμή που θα προκύψει προσθέτουμε και ένα σταθερό όρο w_0 (bias) η σημασία του οποίου είναι ίδια με αυτήν που αναφέρθηκε στον γραμμικό ταξινομητή.



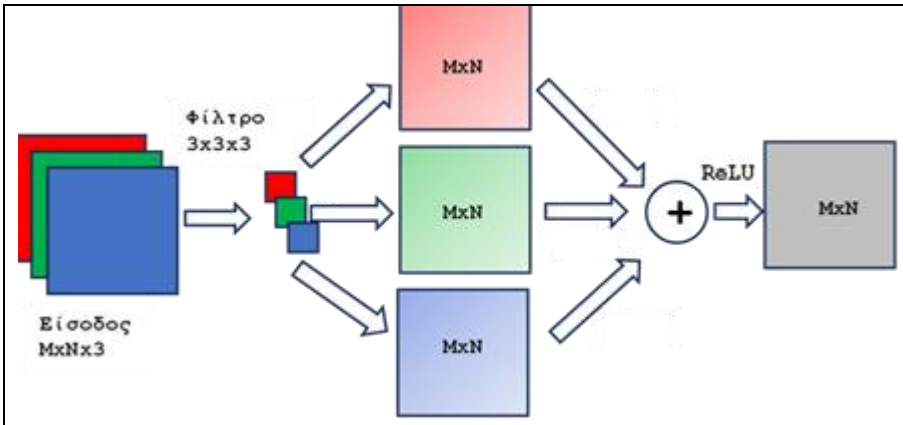
Σχήμα 3.4.2: Υπολογισμοί στο συνελικτικό επίπεδο

Η τιμή κάθε στοιχείου του παραγόμενου γίνεται όρισμα μιας συνάρτησης ενεργοποίησης κυρίως της ReLU.

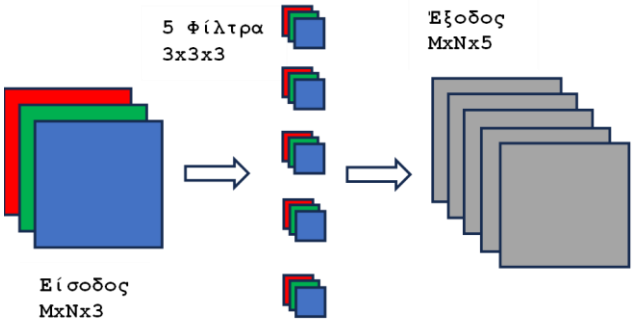


Η συνάρτηση ReLU (Rectified Linear Unit)

$f(x)=\max(0,x)$



Σχ.3.4.3. Η εφαρμογή ενός φίλτρου σε είσοδο τριών καναλιών στο συνελικτικό επίπεδο



Σχ.3.4.4. Συνελικτικό επίπεδο με πέντε (5) φίλτρα

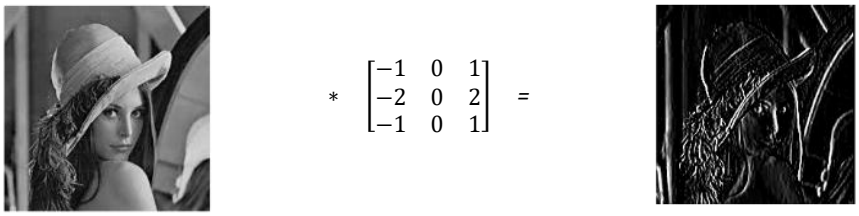
Αν η είσοδος έχει βάθος D δηλαδή αποτελείται από D το πλήθος πίνακες, ονομάζονται και κανάλια (channels), (π.χ. D=3, για είσοδο μία RGB έγχρωμη εικόνα), τότε και κάθε φίλτρο πρέπει να έχει βάθος D, δηλαδή να αποτελείται από D πίνακες. Για παράδειγμα αν η είσοδος είναι R_xxC_x×3 το φίλτρο πρέπει να είναι R_wxC_w×3. Στην περίπτωση αυτή μετά την εφαρμογή του φίλτρου θα προκύψουν και D πίνακες ώς αποτελέσματα. Οι πίνακες αυτοί είναι ίσων διαστάσεων και προστίθενται.

Τελικά μετά την άθροιση προκύπτει ένας πίνακας για κάθε φίλτρο. Αν εφαρμοσθούν N φίλτρα θα προκύψουν N πίνακες, όσοι και τα φίλτρα. Στην ορολογία των ΣΝΔ συχνά ένας τρισδιάστατος πίνακας καλείται και *όγκος* (*volume*). Άρα η είσοδος μας είναι ένας όγκος R_xxC_x×3 και με την εφαρμογή N φίλτρων παράγεται ως αποτέλεσμα ένας όγκος R_xxC_x×N. Τα παραπάνω απεικονίζονται στα ακόλουθα σχήματα (Σχ.3.4.3, Σχ.3.4.4)

Το φίλτρο κατά την ολίσθηση του είναι αποδεκτό να κινείται και με βήμα (δρασκελιά, stride) μεγαλύτερο του ένα. Αυτό είναι μια παραλλαγή συνέλιξης που οδηγεί σε αποτελέσματα μικρότερων διαστάσεων και ακρίβειας. Έχει χρήση στην

περίπτωση μεγάλων εικόνων και φίλτρων (π.χ για φίλτρο 15×15, stride=2). Τα παραπάνω, δηλαδή η συνέλιξη της εισόδου με τα φίλτρα και η εφαρμογή της μη γραμμικότητας αποτελούν ένα *Επίπεδο Συνέλιξης* (*Convolutional Layer*).

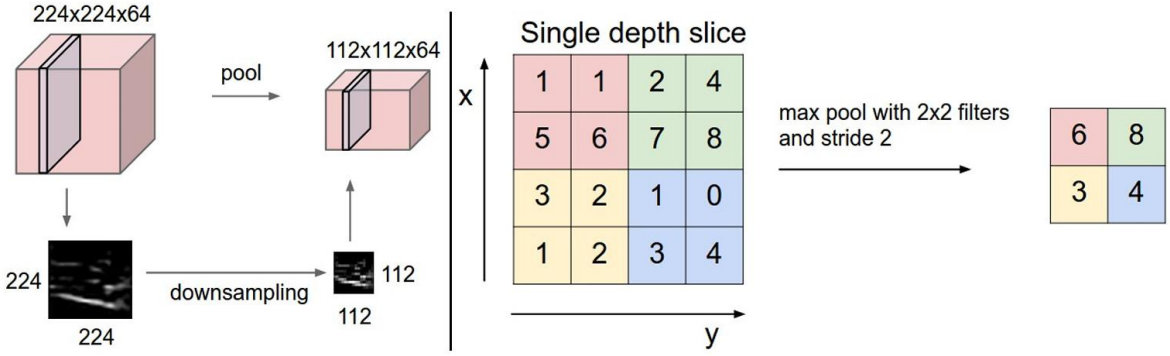
Οι πίνακες που εξάγονται από το επίπεδο συνέλιξης καλούνται ***χάρτες χαρακτηριστικών***. Στο ακόλουθο Σχήμα 3.4.5 φαίνεται η εφαρμογή ενός φίλτρου *Sobel* στην εικόνα «lena».



Σχ. 3.4.5.

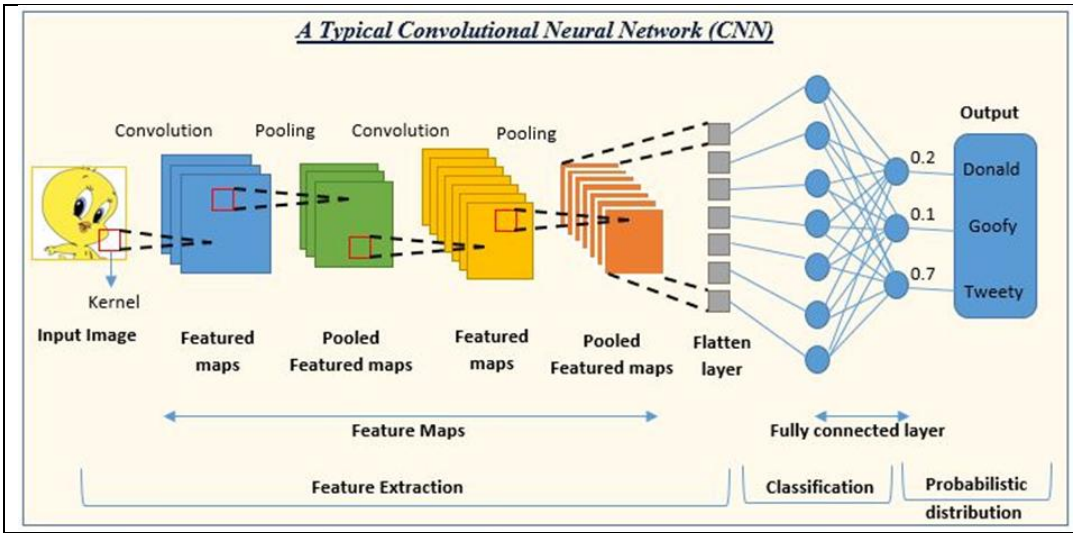
Επίπεδο δειγματολειψίας (Pooling Layer)

Τα επίπεδα δειγματοληψίας συναντώνται ανάμεσα στα επίπεδα συνέλιξης των ΣΝΔ. Αυτό που κάνει αυτό το επίπεδο είναι να μειώνει το μέγεθος των πινάκων εισόδων και κατά συνέπεια και τους υπολογισμούς ενός δικτύου και θέτοντας υπό έλεγχο προβλήματα υπερπροσαρμογής (over-fitting). Λαμβάνοντας τις τιμές μιας μικρής περιοχής π.χ. (M×M) ενός πίνακα εισόδου και με βήμα M παίρνουμε την μεγαλύτερη από αυτές (Max Pooling) ή τον μέσο όρο τους (Average Pooling) και δημιουργούμε έναν νέο πίνακα με μικρότερες διαστάσεις. Αν για παράδειγμα ο πίνακας είναι 224×224 για M=2 θα έχουμε ως αποτέλεσμα ένα πίνακα 112×112. Ένα παράδειγμα φαίνεται στο παρακάτω σχήμα 3.4.6.



Σχήμα 3.4.6. (Επίπεδο Υποδειγματοληψίας)

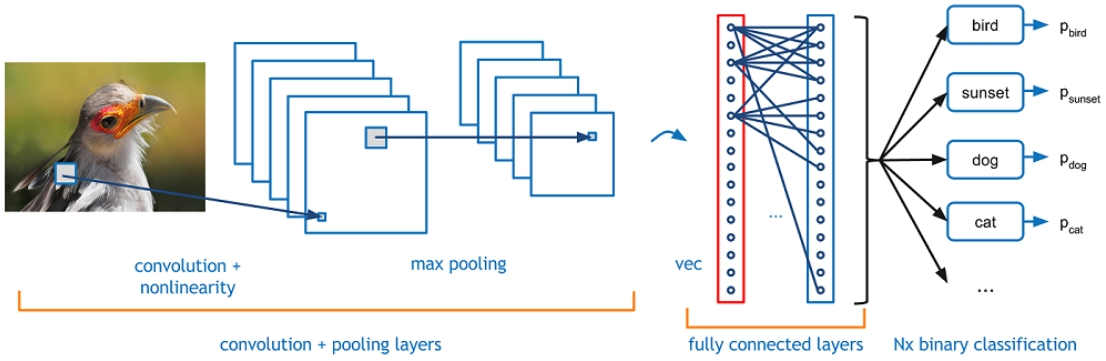
Επίπεδα συνέλιξης και υποδειγματοληψίας μπορούν να τοποθετούνται διαδοχικά καταλήγοντας σε πίνακες μικρότερων διαστάσεων κάθε φορά. Μετά από αυτήν την διαδοχή οι τιμές των τελικών πινάκων τοποθετούνται σε ένα άνυσμα στήλης (*flattening*). Το άνυσμα αυτό θα είναι η είσοδος ενός τελικού επιπέδου ταξινόμησης με πλήρη διασύνδεση, Σχήμα 3.4.7.



Σχ. 3.4.7.

Επίπεδο Ταξινόμησης

Σε αυτό το επίπεδο οι νευρώνες έχουν πλήρεις συνδέσεις με όλες τις εξόδους των νευρώνων του προηγούμενου επιπέδου. Το πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer) είναι ένας Perceptron πολλών επιπέδων (MLP), που χρησιμοποιεί ως συνάρτηση κόστους την πολυωνυμική λογιστική (Multinomial Logistic) στο επίπεδο εξόδου. Για την είσοδο των δεδομένων στο επίπεδο αυτό θα πρέπει αυτά να δομήσουν ένα μονοδιάστατο πίνακα (άνυσμα στήλης). Στο ακόλουθο Σχήμα 3.4.8 παρουσιάζεται μια εκδοχή ενός CNN τεσσάρων κλάσεων.



Σχήμα 3.4.8.

Υπολογισμός του πλήθους βαρών και εκπαίδευση στα ΣΝΔ

Επίπεδο συνέλιξης : Το συγκεκριμένο επίπεδο είναι αυτό που δημιουργεί τον πίνακα χαρακτηριστικών, οπότε έχουμε πίνακες βαρών. Οι αρχικές τιμές των βαρών είναι τυχαίες. Ο αριθμός των βαρών (παραμέτρων) σε ένα επίπεδο συνέλιξης με N κανάλια και M φίλτρα διαστάσεων r × c θα είναι: r×c×N×M

Επίπεδο δειγματοληψίας : Στο επίπεδο δειγματοληψίας μειώνουμε το μέγεθος των διαστάσεων. Για παράδειγμα αν σε έναν χάρτη χαρακτηριστικών 224×224 στοιχείων εφαρμοστεί δειγματοληψία σε 2×2 υπό-περιοχές τότε αριθμός των στοιχείων του χάρτη χαρακτηριστικών είναι 112×112. Στο επίπεδο δειγματοληψία δεν υπάρχουν βάρη.

Πλήρως συνδεδεμένο επίπεδο (FC) : Το συγκεκριμένο επίπεδο έχει τον υψηλότερο αριθμό βαρών από κάθε άλλο επίπεδο. Για τον υπολογισμό του πλήθους των βαρών υπολογίζεται το γινόμενο του πλήθους των νευρώνων κάθε επιπέδου επί το πλήθος των εξόδων του προηγούμενου επιπέδου. Έτσι το πλήθος των βαρών είναι: πλήθος νευρώνων στο τρέχον επίπεδο × (πλήθος εξόδων στο προηγούμενο επίπεδο+1), το +1 αφορά τον σταθερό όρο. Για παράδειγμα στο ***CNN VGGNET*** το πλήθος των βαρών φαίνεται στον ακόλουθο Πίνακα.

```
INPUT: [224x224x3]        memory:  224*224*3=150K   weights: 0
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*3)*64 = 1,728
CONV3-64: [224x224x64]  memory:  224*224*64=3.2M   weights: (3*3*64)*64 = 36,864
POOL2: [112x112x64]  memory:  112*112*64=800K   weights: 0
CONV3-128: [112x112x128]  memory:  112*112*128=1.6M   weights: (3*3*64)*128 = 73,728
CONV3-128: [112x112x128]  memory:  112*112*128=1.6M   weights: (3*3*128)*128 = 147,456
POOL2: [56x56x128]  memory:  56*56*128=400K   weights: 0
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*128)*256 = 294,912
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*256)*256 = 589,824
CONV3-256: [56x56x256]  memory:  56*56*256=800K   weights: (3*3*256)*256 = 589,824
POOL2: [28x28x256]  memory:  28*28*256=200K   weights: 0
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*256)*512 = 1,179,648
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [28x28x512]  memory:  28*28*512=400K   weights: (3*3*512)*512 = 2,359,296
POOL2: [14x14x512]  memory:  14*14*512=100K   weights: 0
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
CONV3-512: [14x14x512]  memory:  14*14*512=100K   weights: (3*3*512)*512 = 2,359,296
POOL2: [7x7x512]  memory:  7*7*512=25K  weights: 0
FC: [1x1x4096]  memory:  4096  weights: 7*7*512*4096 = 102,760,448
FC: [1x1x4096]  memory:  4096  weights: 4096*4096 = 16,777,216
FC: [1x1x1000]  memory:  1000 weights: 4096*1000 = 4,096,000

TOTAL memory: 24M * 4 bytes ~= 93MB / image (only forward! ~*2 for bwd)
TOTAL params: 138M parameters
```

Πίνακας 3.4.1: Πλήθος των βαρών στο CNN VGGNet 16

Εκπαίδευση των Συνελικτικών Νευρωνικών Δικτύων

Στα CNN η συνάρτηση κόστους βασίζεται στην διασταυρούμενη εντροπία (cross entropy) και όχι στο μέσο τετραγωνικό σφάλμα. Η προσέγγιση ακρότατου σημείου (ελαχίστου) επιδιώκεται κυρίως με μεθόδους όπως η SGDM (Stochastic Gradient Descent with Momentum) και η ADAM (Adaptive Moment Estimation). Στο τελικό επίπεδο εξόδου του ταξινομητή, συνάρτηση ενεργοποίησης είναι η *softmax* που μετατρέπει τις τιμές των αθροιστών σε πιθανότητες. Ακολούθως θα παρουσιάσουμε την διαδικασίας εκπαίδευσης και διόρθωσης σφάλματος στα επίπεδα ενός CNN.

Εκπαίδευση στο επίπεδο πλήρους διασύνδεσης

Στο τελικό επίπεδο εξόδου το πλήθος των νευρώνων ορίζεται να είναι ίσο με το πλήθος των κλάσεων, έστω Μ. Κάθε κλάση αντιστοιχίζεται σε έναν νευρώνα (τον ίδιο πάντοτε) και ονοματίζεται με την θέση του m=1...Μ στο επίπεδο εξόδου. Έστω το $\boldsymbol{x} = [x_1 \dots x_n \dots x_{N,} 1]$ επαυξημένο διάνυσμα που εισέρχεται στο επίπεδο εξόδου και $\boldsymbol{w_m} = [w_{1m}, \dots w_{nm}, \dots w_{Nm}, w_{0m}]^T$ τα επαυξημένα διανύσματα των βαρών των νευρώνων σ' αυτό.

Αρχικά υπολογίζονται οι ποσότητες $\sigma_m = \boldsymbol{w}_m^T \cdot \boldsymbol{x}$. Οι τιμές αυτές μετασχηματίζονται σύμφωνα την συνάρτηση softmax σε

$$y_m = \frac{e^{\boldsymbol{w}_m^T \cdot \boldsymbol{x}}}{\sum_k e^{\boldsymbol{w}_k^T \cdot \boldsymbol{x}}}$$

Οι τιμές y_m είναι θετικές και το άθροισμα τους $\sum_{m=1}^{M} y_m = 1$. Επιθυμούμε οι τιμές εξόδου y_m να είναι τέτοιες ώστε αν το άνυσμα εισόδου ανήκει στην κλάση c, $\mathbf{x}^c$, $c \in \{1, \dots, M\}$ τότε η έξοδος y_c του c νευρώνα να είναι μεγαλύτερη των άλλων εξόδων και ει δυνατόν ίση με την μονάδα εκφράζοντας έτσι υψηλή πιθανότητα να ανήκει η είσοδος στην επιθυμητή κλάση. Με όρους πιθανοτήτων αν $1, \dots, M$ είναι οι τιμές μιας τυχαίας διακριτής μεταβλητής ω, τότε $y_m = p(\omega = m|\boldsymbol{x^c})$ είναι οι πιθανότητες που αντιστοιχούν στις τιμές τις ω δηλαδή $y1=p(1|\mathbf{x}^c), \dots, y_m=p(m|\mathbf{x}^c), \dots, y_M=p(M|\mathbf{x}^c)$. Το ιδανικό θα ήταν οι πιθανότητες $p(\omega = m|\boldsymbol{x^c})$ να ήταν ή να συνέκλιναν με πιθανότητες

$$q(\omega = m|\boldsymbol{x^c}) = \begin{cases} 1 \; \alpha\nu \; m = c \\ 0 \; \alpha\nu \; m \neq c \end{cases}$$

Ένα μέτρο σύγκρισης των p(ω) και q(ω) δίνεται από την σχέση της διασταυρούμενης εντροπίας (cross entropy):

$$H(q,p) = -\sum_{\forall \omega} q(\omega) \cdot \log\big(p(\omega)\big) =$$

$$-\sum_{m=1}^{M} q(\omega = m|\boldsymbol{x}^c) \cdot \log\big(p(\omega = m|\boldsymbol{x}^c)\big) = -\sum_{m=1}^{M} q(\omega = m|\boldsymbol{x}^c) \cdot \log(y_m) = -\log(y_c)$$

Για παράδειγμα σε ένα σύστημα ταξινόμησης τεσσάρων κλάσεων εισέρχεται το άνυσμα $\mathbf{x}^2$ και οι έξοδοι των τεσσάρων νευρώνων είναι

$y_1=0.2$, $y_2=0.5$ $y_3=0.1$ $y_4=0.2$

και προφανώς

$q(1|x_2)=0$, $q(2|x_2)=1$, $q(3|x_2)=0$, $q(4|x_2)=0$

H(q,p)=-(0·log(0.2)+1·log(0.5)+0·log(0.1)+ 0·log(0.2))=2

Όσο η y_2 τείνει στην μονάδα η H(q,p) θα τείνει στο μηδέν.

Κατόπιν αυτού μπορούμε να εισάγουμε ως συνάρτηση κόστους (cost function) ή απώλειας (loss) την σχέση

$$L(\boldsymbol{W}) = -\sum_i \sum_m q(\omega = m|\boldsymbol{x}_i^c) \cdot \log(y_{mi}) = -\sum_i \log(y_{c,i})$$

που αθροίζει τις τιμές της cross entropy που προκύπτουν από την εισαγωγή στο σύστημα κάθε προτύπου $\mathbf{x}_i$, i=1...I, I το πλήθος των προτύπων.

$$y_{c,i} = \frac{\mathrm{e}^{\boldsymbol{w}_c^T \cdot \boldsymbol{x}_i}}{\sum_m \mathrm{e}^{\boldsymbol{w}_m^T \cdot \boldsymbol{x}_i}}$$

Για τον νευρώνα j στο επίπεδο εξόδου $j \in \{1 \dots M\}$ και wj το άνυσμα στήλης των βαρών του ισχύει

$$\frac{\partial L}{\partial \boldsymbol{w}_j} = \frac{\partial}{\vartheta \boldsymbol{w}_j}\left(-\sum_i \log(y_{ci})\right) = -\sum_i \frac{\partial}{\vartheta \boldsymbol{w}_j} \log(y_{ci}) = -\sum_i \frac{\partial}{\vartheta \boldsymbol{w}_j} \log\left(\frac{\mathrm{e}^{\boldsymbol{w}_c^T \cdot \boldsymbol{x}_i}}{\sum_m \mathrm{e}^{\boldsymbol{w}_m^T \cdot \boldsymbol{x}_i}}\right)$$

Για τον υπολογισμό της παραγώγου γράφουμε τα παραπάνω υπό μορφή σύνθετης συνάρτησης. Θα παραλείψουμε για λόγους διευκόλυνσης της γραφής τον δείκτη i (θέτω x=xi)

$\sigma_j = \boldsymbol{w}_j^T \cdot \boldsymbol{x}$, $y_c = \frac{e^{\sigma_c}}{\sum_m e^{\sigma m}}$, $l = \log(y_c)$

$$\frac{\partial l}{\partial \boldsymbol{w}_j} = \frac{\partial l}{\partial y_c}\frac{\partial y_c}{\partial \sigma_j}\frac{\partial \sigma_j}{\partial \boldsymbol{w}_j}$$

$$\frac{\partial \sigma_j}{\partial \boldsymbol{w}_j} = \boldsymbol{x}_i^T, \quad \frac{\partial l_i}{\partial y_c} = \frac{1}{y_c}, \frac{\partial y_c}{\partial \sigma_j} = \frac{\partial}{\partial z_j}\left(\frac{e^{\sigma_c}}{\sum_m e^{\sigma m}}\right) = \begin{cases} \frac{e^{\sigma} \cdot \sum_m e^{\sigma m} - e^{\sigma_c} \cdot e^{\sigma_c}}{(\sum_m e^{\sigma m})^2} = y_c(1 - y_c) & \alpha\nu\, j = c \\ \frac{-e^{\sigma_c} \cdot e^{\sigma_j}}{(\sum_m e^{\sigma m})^2} = -y_c \cdot y_j & \alpha\nu\, j \neq c \end{cases}$$

άρα $\frac{\partial l_i}{\partial w_j} = \begin{cases} (1 - y_c) \cdot \boldsymbol{x}_i^T & \alpha\nu\, j = c \\ -y_j \cdot \boldsymbol{x}_i^T & \alpha\nu\, j \neq c \end{cases}$, για κάθε συναπτικό βάρος του j νευρώνα

$\frac{\partial l_i}{\partial w_{nj}} = \begin{cases} (1 - y_c) \cdot x_{ni} & \alpha\nu\, j = c \\ -y_j \cdot x_{ni} & \alpha\nu\, j \neq c \end{cases}$ και τελικά

$\frac{\partial L}{\partial w_{nj}} = -\sum_i \frac{\partial l_i}{\partial w_{nj}}$

Διόρθωση σφάλματος με οπισθοδιάδοση στο επίπεδο δειγματοληψίας.

Τα επίπεδα δειγματοληψίας (pooling) δεν διαθέτουν παραμέτρους που θα αλλάξουν κατά την εκπαίδευση του ΣΝΔ. Αυτό που έχουμε μόνο να κάνουμε είναι να υπολογίσουμε τις κλίσεις σε σχέση με το προηγούμενο επίπεδο. Αν R είναι μια περιοχή του επιπέδου που υποδειγματοληπτείται με την διαδικασία του maxpooling η έξοδος θα είναι $y = \max\{x_i \in R\}$ και $\frac{\partial y}{\partial x_\iota} = \begin{cases} 1 \; \alpha\nu \; y = x_i \\ 0 \; \alpha\nu \; y \neq x_i \end{cases}$. Για παράδειγμα αν η περιοχή είναι 2×2 θα έχουμε:

x_1=5	x_2=8
x_3=3	x_4=1

->Max Pooling-> y=8

$\frac{\partial y}{\partial x_1} = 0$	$\frac{\partial y}{\partial x_2} = 1$
$\frac{\partial y}{\partial x_3} = 0$	$\frac{\partial y}{\partial x_4} = 0$

Σύμφωνα με τα παραπάνω η διόρθωση των βαρών του φίλτρου θα γίνει μόνο για την θέση του επί του πίνακα εισόδου (κατά την συνελικτική διαδικασία) που παρήχθη η μέγιστη τιμή (το 8). Στην περίπτωση του average pooling, η κλίση θα είναι παντού $1/n$ όπου n είναι το πλήθος των στοιχείων του παραθύρου (για το παράδειγμα μας 1/4, n=4).

Διόρθωση σφάλματος με οπισθοδιάδοση στο επίπεδο συνέλιξης

Στο συνελικτικό επίπεδο ένα φίλτρο είναι ένας νευρώνας με βάρη που το πλήθος τους καθορίζεται από τις διαστάσεις του φίλτρου. Οι τιμές εισόδου προέρχονται από τις τιμές του καναλιού καθώς αυτές συνελίσονται με τα βάρη του φίλτρου. Αυτό μπορεί να θεωρηθεί σε μια παράλληλη υλοποίηση όπως φαίνεται στο Σχήμα 3.4.9, για δεδομένα μιας διάστασης και φίλτρο εύρους τρία (3×1).

Έστω $\delta_0 = \frac{\partial_L}{\partial_{\sigma_0}}$, $\delta_1 = \frac{\partial_L}{\partial_{\sigma_1}}$, $\delta_2 = \frac{\partial_L}{\partial_{\sigma_2}}$, $\delta_3 = \frac{\partial_L}{\partial_{\sigma_3}}$ οι μεταβολές της συνάρτησης κόστους ως προς τα αθροίσματα (έξοδοι αθροιστών), $\boldsymbol{w} = [w_0, w_1, w_2]$, $\boldsymbol{\sigma} = [\sigma_0, \sigma_1, \sigma_2]^T$. Θα ισχύουν οι σχέσεις

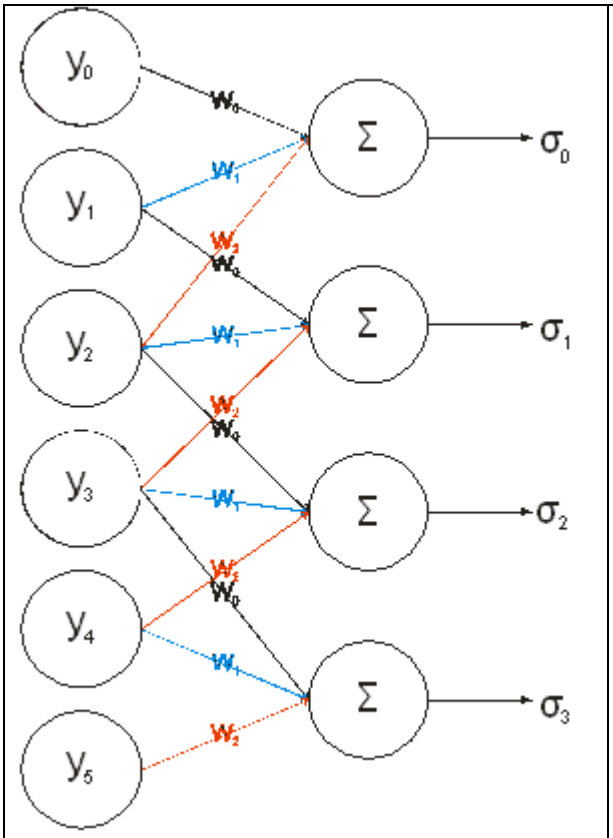
$\sigma_0 = w_0 y_0 + w_1 y_1 + w_2 y_2$, $\sigma_1 = w_0 y_1 + w_1 y_2 + w_2 y_3$

$\sigma_2 = w_0 y_2 + w_1 y_3 + w_2 y_4$, $\sigma_3 = w_0 y_3 + w_1 y_4 + w_2 y_5$

$$\frac{\partial_L}{\partial_{w_0}} = \frac{\partial_L}{\partial_{\boldsymbol{\sigma}}} \cdot \frac{\partial_{\boldsymbol{\sigma}}}{\partial_{w_0}} = [\delta_0, \delta_1, \delta_2, \delta_3] \cdot \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \delta_0 y_0 + \delta_1 y_1 + \delta_2 y_2 + \delta_3 y_3$$

$$\frac{\partial_L}{\partial_{w_1}} = \frac{\partial_L}{\partial_{\boldsymbol{\sigma}}} \cdot \frac{\partial_{\boldsymbol{\sigma}}}{\partial_{w_1}} = [\delta_0, \delta_1, \delta_2, \delta_3] \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \delta_0 y_1 + \delta_1 y_2 + \delta_2 y_3 + \delta_3 y_4$$

$$\frac{\partial_L}{\partial_{w_2}} = \frac{\partial_L}{\partial_{\boldsymbol{\sigma}}} \cdot \frac{\partial_{\boldsymbol{\sigma}}}{\partial_{w_2}} = [\delta_0, \delta_1, \delta_2, \delta_3] \cdot \begin{bmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \delta_0 y_2 + \delta_1 y_3 + \delta_2 y_4 + \delta_3 y_5$$



$$\sigma_0 = w_0 y_0 + w_1 y_1 + w_2 y_2$$

$$\sigma_1 = w_0 y_1 + w_1 y_2 + w_2 y_3$$

$$\sigma_2 = w_0 y_2 + w_1 y_3 + w_2 y_4$$

$$\sigma_3 = w_0 y_3 + w_1 y_4 + w_2 y_5$$

Σχήμα 3.4.8.

Για την οπισθοδιάδοση της διόρθωσης Θα πρέπει να υπολογίσουμε τις ποσότητες:

$$\frac{\partial_L}{\partial_{y_0}}, \frac{\partial_L}{\partial_{y_1}}, \dots, \frac{\partial_L}{\partial_{y_5}}$$

$$\frac{\partial_L}{\partial_{y_0}} = \frac{\partial_L}{\partial_{\boldsymbol{\sigma}}} \cdot \frac{\partial_{\boldsymbol{\sigma}}}{\partial_{y_0}} = [\delta_0, \delta_1, \delta_2, \delta_3] \cdot \begin{bmatrix} w_0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \delta_0 w_0$$

$$\frac{\partial_L}{\partial_{y_1}} = \frac{\partial_L}{\partial_{\boldsymbol{\sigma}}} \cdot \frac{\partial_{\boldsymbol{\sigma}}}{\partial_{y_1}} = [\delta_0, \delta_1, \delta_2, \delta_3] \cdot \begin{bmatrix} w_1 \\ w_0 \\ 0 \\ 0 \end{bmatrix} = \delta_0 w_1 + \delta_1 w_0$$

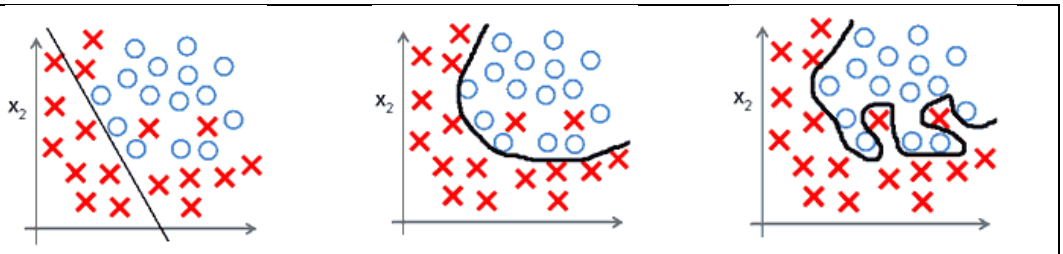
$$\frac{\partial_L}{\partial_{y_2}} = \delta_0 w_2 + \delta_1 w_1 + \delta_2 w_0, \ \frac{\partial_L}{\partial_{y_3}} = \delta_1 w_2 + \delta_2 w_1 + \delta_3 w_0$$

$$\frac{\partial_L}{\partial_{y_4}} = \delta_2 w_2 + \delta_3 w_1, \ \frac{\partial_L}{\partial_{y_5}} = \delta_3 w_2$$

Γενικότερα, σε ένα συνελικτικό επίπεδο, η κλίση του τρέχοντος επιπέδου προκύπτει με την συνέλιξη των δ_i με τα συνελικτικά φίλτρα αντεστραμμένα.

Το πλήθος των επιπέδων και των νευρώνων των, είναι ένα ζήτημα που μας απασχολεί. Δεν υπάρχει ένας γενικός τύπος-αλγόριθμος που να μας οδηγεί σε ένα βέλτιστο πλήθος παραμέτρων-βαρών. Ένας μεγάλος αριθμός βαρών αυξάνει το υπολογιστικό κόστος και μπορεί να οδηγήσει σε λεπτομερή εκμάθηση του συνόλου εκπαίδευσης ακόμη και του θορύβου των μετρήσεων, εις βάρος της γενικότητας του, δηλαδή της επιτυχούς ταξινόμησης νέων αγνώστων προτύπων. Αυτό το φαινόμενο το ονομάζουμε *υπερεκπαίδευση*. Μικρός αριθμός παραμέτρων μπορεί να οδηγήσει σε αδυναμία εκπαίδευσης ή χαμηλή απόδοση-ακρίβεια ταξινόμησης. Την περίπτωση αυτή αναφέρουμε ως *υποεκπαίδευση*. Στο Σχήμα 3.4.10 φαίνεται η διαχωριστική καμπύλη τριών συστημάτων ενός προβλήματος δύο διαστάσεων για τις περιπτώσεις που αναφέραμε.

Μία απλή μέθοδος είναι να ξεκινάμε με ένα μικρό πλήθος παραμέτρων και να το αυξάνουμε προοδευτικά μέχρι να έχουμε ικανοποιητικό αποτέλεσμα. Εκείνο που συχνά εμφανίζεται όταν έχουμε υψηλή απόδοση εκπαίδευσης είναι η υπερεκπαίδευση που διαπιστώνεται όταν το σύστημά μας ταξινομεί πρότυπα για τα οποία δεν εκπαιδεύτηκε και παρουσιάζει χαμηλή απόδοση. Για να εντοπίσουμε νωρίς το



(α) Υποεκπαιδευμένο σύστημα

(β) Κανονικά εκπαιδευμένο σύστημα

(γ) Υπερεκπαιδευμένο σύστημα

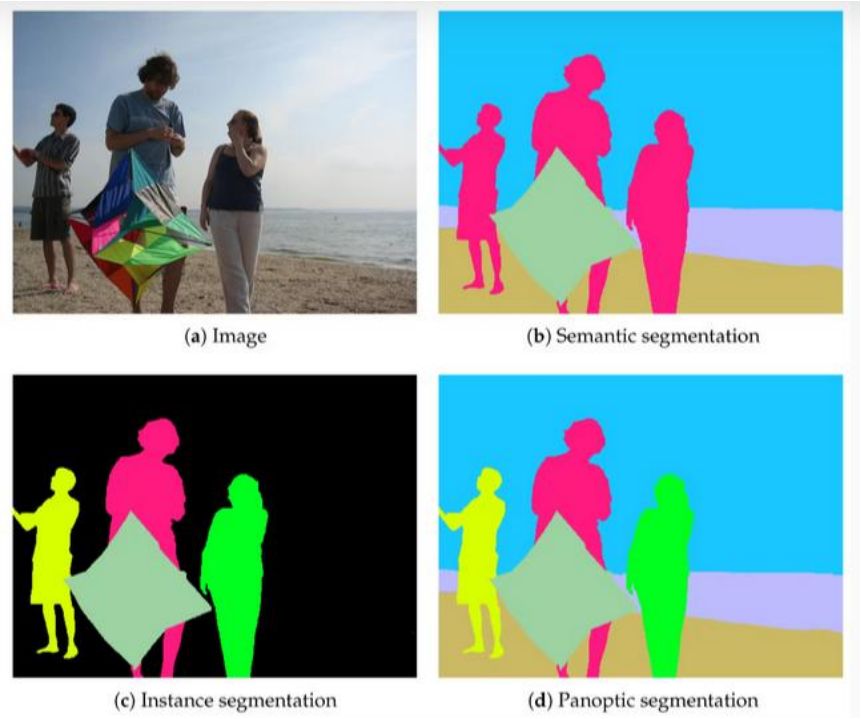
Σχήμα 3.4.10.

πρόβλημα χωρίζουμε το αρχικό σύνολο S των διαθέσιμων δεδομένων σε δύο μέρη. Το ένα χρησιμοποιείται ως σύνολο εκπαίδευσης (Training set, T) και το άλλο ως σύνολο επικύρωσης του αποτελέσματος (Validation set, V). Θα πρέπει το ποσοστό επιτυχίας επιτυχούς ταξινόμησης (accuracy) να είναι υψηλό με παραπλήσιες τιμές.

3.4.1 Συνελικτικά νευρωνικά δίκτυα για σημασιολογική κατάτμηση εικόνων

Η κατάτμηση εικόνας είναι μια διαδικασία επεξεργασίας της εικόνας για τον εντοπισμό περιοχών με συγκεκριμένες ιδιότητες. Η σημασιολογική κατάτμηση ταξινομεί όλα τα συσχετιζόμενα εικονοστοιχεία σε μια ομάδα. Η περιπτωσιολογική κατάτμηση διακρίνει ως ξεχωριστές οντότητες αντικείμενα της ίδιας κλάσης. Η πανοπτική κατάτμηση συνδυάζει τα δυο παραπάνω είδη κατάτμησης. (Σχήμα 3.4.1.1).

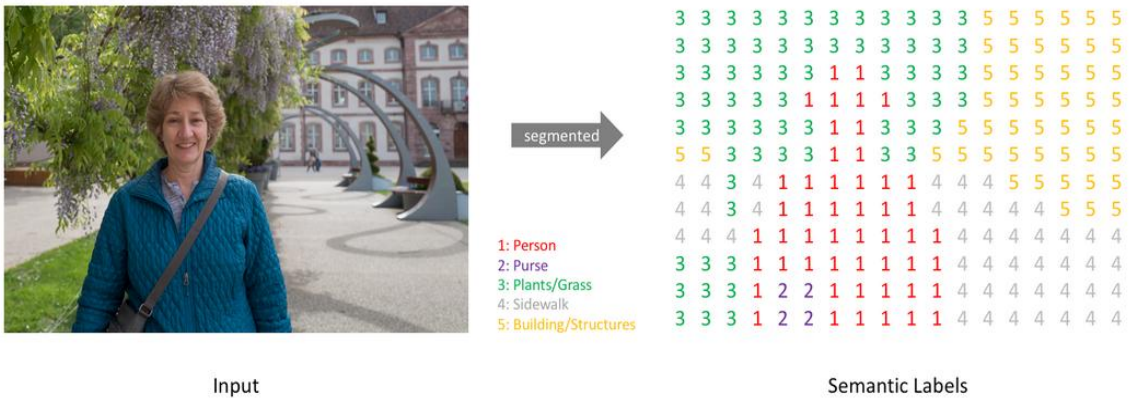
Η Τμηματοποίηση Περιοχών διαιρεί την εικόνα σε περιοχές που μοιράζονται παρόμοια χαρακτηριστικά, όπως χρώμα και υφή.



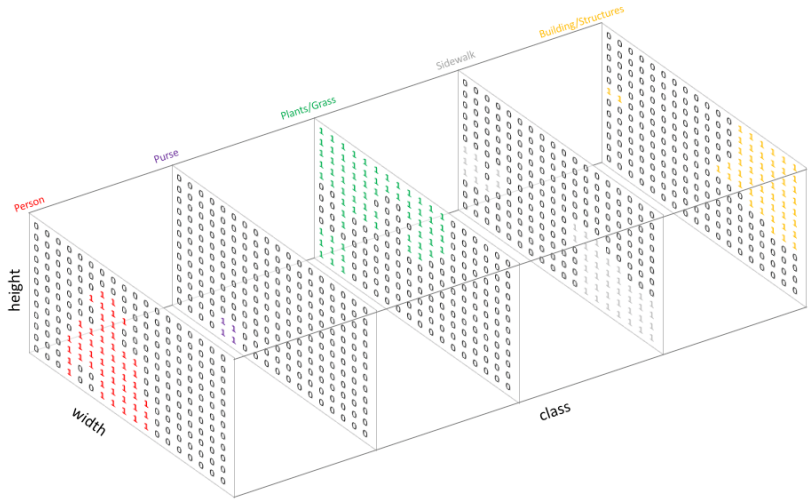
Σχήμα 3.4.1.1: Παραδείγματα των τριών ειδών κατάτμησης εικόνας.

Στη σημασιολογική κατάτμηση το ΣΝΔ ενεργεί όπως στη διαδικασία ταξινόμησης με τη μόνη διαφορά πως αντί να ταξινομεί ολόκληρη την εικόνα, ταξινομεί το κάθε εικονοστοιχείο της σε μια προκαθορισμένη κλάση (π.χ άνθρωπος). Τα ΣΝΔ σημασιολογικής κατάτμησης χρησιμοποιούνται σε πολλούς τομείς της Βιομηχανίας, της Ιατρικής, της Τηλεπισκόπησης κ.α.

Συνήθως η είσοδος είναι μια ψηφιακή H × W × D με D = 3 για εικόνες RGB, ή D = 1 για εικόνες αποχρώσεων του γκρι. Η έξοδος είναι ένας πίνακας τμηματοποίησης H × W που κάθε εικονοστοιχείο αντιπροσωπεύεται από έναν ακέραιο αριθμό (ετικέτα, label) αντιπροσωπευτικό της κλάσης που ανήκει (Σχήμα 3.4.1.2). Στη πραγματικότητα ο χάρτης, πριν πάρει τη τελική του μορφή, έχει βάθος ίσο με τον αριθμό των κλάσεων (Σχήμα 3.4.1.3).

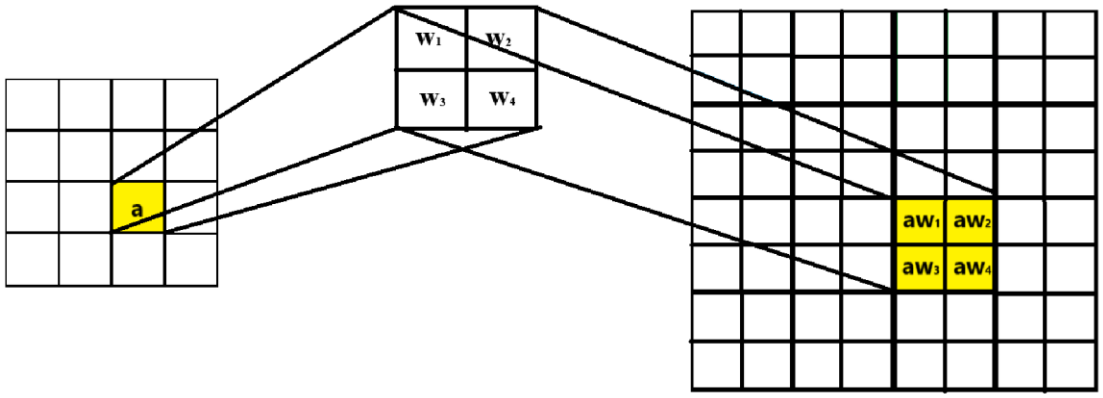


Σχήμα 3.4.1.2: Χάρτης κατάτμησης για μια εικόνα εισόδου.



Σχήμα 3.4.1.3: Πίνακας κατάτμησης για μια εικόνα εισόδου πριν εφαρμοστεί η argmax.

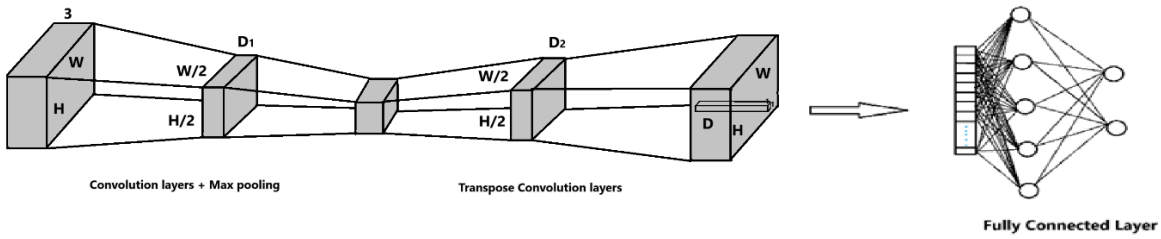
Μια συνηθισμένη προσέγγιση για εκπαίδευση ΣΝΔ για κατάτμηση εικόνων είναι η δομή κωδικοποιητή/αποκωδικοποιητή. Όπου ο πρώτος θα μειώνει τη χωρική ανάλυση της εισόδου, και ο ρόλος του θα είναι να οδηγεί σε χάρτες χαρακτηριστικών χαμηλής ανάλυσης και ο δεύτερος θα αυξάνει τις χωρικές διαστάσεις του χάρτη χαρακτηριστικών. Η μείωση των διαστάσεων γίνεται με την διαδικασία του *pooling* όπως έχει αναλυθεί στο Κεφαλαίο 2. Η πιο δημοφιλής μέθοδος για την αύξηση της ανάλυσης είναι η ανάστροφη συνέλιξη (ΑΣ), (*transposed convolution*:*TC*). Στην ανάστροφη συνέλιξη κάθε τιμή από τον χαμηλής ανάλυσης χάρτη χαρακτηριστικών (που έρχεται από τα προηγούμενα επίπεδα που έδρασε η max pooling) πολλαπλασιάζεται με τις τιμές κάποιου φίλτρου. Έπειτα, τοποθετούνται σε διαδοχικές θέσεις του χάρτη εξόδου χαρακτηριστικών και έτσι αυξάνεται η ανάλυση του. Συνεπώς, ενώ η κλασική συνέλιξη συρρικνώνει τη πληροφορία, η ανάστροφη συνέλιξη διαστέλλει τη πληροφορία και έτσι αυξάνονται οι διαστάσεις της εικόνας (ή του πίνακα). Για παράδειγμα αν ενας από τους υπο-πίνακες χαρακτηριστικών που προέκυψε από ένα συνελικτικό επίπεδο (Κεφ.2) είναι 4×4 τότε με την ανάστροφο συνέλιξη (ΑΣ) με ένα φίλτρο 2×2 και βήμα 2 (stride = 2) (άρα δεν έχουμε επικάλυψη) θα προκύψει ένας υπο-πίνακας 8×8 κάθε στοιχείο του οποίου έχει την τιμή όπως δείχνεται στο ακόλουθο Σχήμα 3.4.1.4.



Σχήμα 3.4.1.4: Ολίσθηση του φίλτρου σε έναν πίνακα Α με βήμα 2.

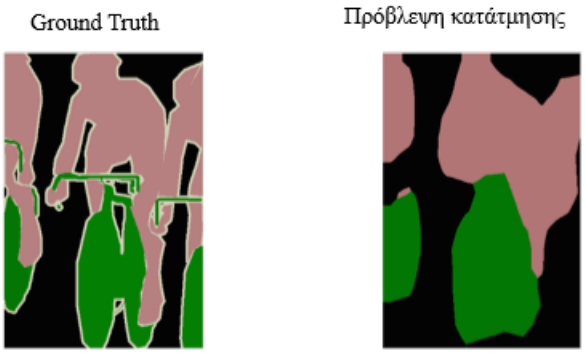
Αν το βάθος του πίνακα στο οποίο εφαρμόσθηκε η ΑΣ είναι D τότε κάθε 2×2 φίλτρο θα πρέπει να έχει βάθος D δηλαδή θα είναι ένας όγκος 2×2×D. Οι τιμές των D υποπινάκων που θα προκύψουν θα προστεθούν. Αν το πλήθος των 2×2×D φίλτρων της ΑΣ είναι N ο όγκος των αποτελεσμάτων που θα προκύψει για το παράδειγμά μας θα είναι 8×8×N. Στην όλη διαδικασία συμμετέχει η συνάρτηση *ReLU* όπως και στο συνελικτικό επίπεδο.

Η ακολουθία των ΑΣ ακολουθεί αφού ολοκληρωθεί η διαδικασία των ΣΕ. (Σχήμα 3.4.1.5).



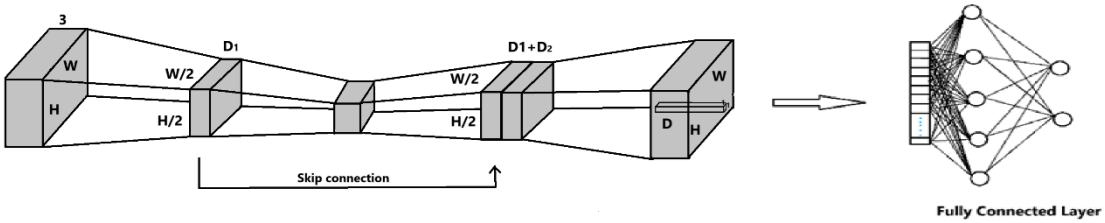
Σχήμα 3.4.1.5: Εφαρμογή επιπέδων ΑΣ έπειτα από χρήση ΣΕ

Η ιδέα χρήσης σε μια αρχιτεκτονική κωδικοποιητή αποκωδικοποιητή όπου σε ένα ΣΝΔ προστίθενται επίπεδα ανάστροφης συνέλιξης ο κωδικοποιητής μειώνει τις διαστάσεις της εικόνας. Αυτό σημαίνει πως οι τελικοί χάρτες χαρακτηριστικών δεν περιγράφουν τα λεπτομερή χαρακτηριστικά και το τελικό αποτέλεσμα υστερεί σε ακρίβεια (Σχήμα 3.4.1.6).



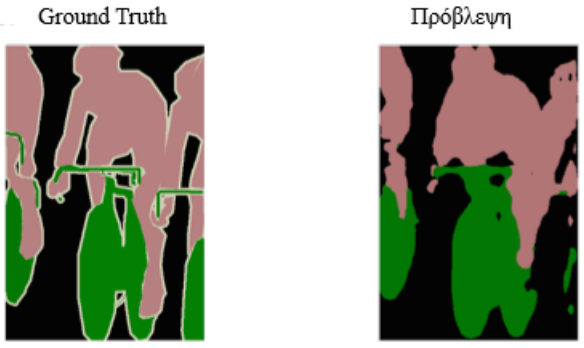
Σχήμα 3.4.1.6: Πρόβλεψη σημασιολογικής κατάτμησης ενός πλήρους συνελικτικού δικτύου.

Για την επίλυση του προβλήματος προτάθηκε μια βελτίωση της παραπάνω αρχιτεκτονικής. Σε αυτή παρατίθενται οι πίνακες χαρακτηριστικών που προέκυψαν στα αρχικά κυρίως συνελικτικά επίπεδα σε πίνακες που προέκυψαν σε στάδια ΑΣ αρκεί βέβαια να έχουν το ίδιο πλήθος γραμμών και στηλών. Για παράδειγμα αν $H \times W \times D_1$ είναι όγκος που προέκυψε από κάποιο συνελικτικό επίπεδο και $H \times W \times D_2$ όγκος που προέκυψε από κάποιο επίπεδο ΑΣ, προκύπτει ένας όγκος $H \times W \times (D_1+D_2)$. Η σύνδεση αυτή ονομάζεται *Συνδέσεις Παράκαμψης* (*skip connections*).



Σχήμα 3.4.1.7: Διάγραμμα των συνδέσεων (skip connections).

Αυτό έχει ως αποτέλεσμα, στη διάρκεια αποκωδικοποίησης οι χάρτες χαρακτηριστικών να είναι πιο ακριβείς καθώς μεταφέρονται λεπτομερείς χωρικές πληροφορίες που χάνονταν κατά τη διάρκεια της υποδειγματοληψίας. Έτσι το ΣΝΔ παράγει πιο ακριβή περιγράμματα στην τελική κατάτμηση (Σχήμα 3.4.1.8).



Σχήμα 3.4.1.8: Πρόβλεψη με χρήση συνδέσεων παράκαμψης.

Ο τελευταίος όγκος που παράγεται από την εφαρμογή των επιπέδων ανάστροφης συνέλιξης για μια αρχική έγχρωμη εικόνα $H \times W \times 3$ θα είναι $H \times W \times D$. Για κάθε εικονοστοιχείο θα έχει προκύψει ένα άνυσμα με D στοιχεία και θα ανήκει σε μια

από τις επιθυμητές κλάσεις. Από το σημείο αυτό τα ανύσματα αποτελούν τις εισόδους ενός νευρωνικού δικτύου εκπαίδευσης με πλήρης συνδέσεις με τελική συνάρτηση ενεργοποίησης την *softmax* και συνάρτηση κόστους την *cross entropy* όπως περιγράφηκαν στην προηγούμενη παράγραφο 3.3.

3.5. ΑΝΑΔΡΟΜΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Ένα επαναληπτικό νευρωνικό δίκτυο (Recurrent neural network, RNN) είναι σχεδιασμένο για την επεξεργασία ακολουθιακών δεδομένων, όπως κείμενο, ήχος ή χρονοσειρές. Σε αντίθεση με τα δίκτυα εμπρόσθιας τροφοδότησης (feed forward nets), τα οποία επεξεργάζονται κάθε είσοδο ξεχωριστά, τα RNNs διατηρούν μια «κρυφή» εσωτερική μονάδα που συν-τροφοδοτείται από προηγούμενες τιμές εισόδου, επιτρέποντας στο δίκτυο να "θυμάται" προηγούμενες πληροφορίες.

Απλά ανδρομικά νευρωνικά δίκτυα (VANILLA Recurrent Neural Network)

Ένα απλό Επαναληπτικό Νευρωνικό Δίκτυο αποτελεί την βασική δομή ενός επαναληπτικού νευρωνικού δικτύου και είναι σχεδιασμένο για την επεξεργασία ακολουθιακών δεδομένων. Στο επίπεδο εισόδου μαζί την τρέχουσα είσοδο εισέρχονται και τιμές που προέκυψαν από την προηγούμενη είσοδο. Ένα επίπεδο βαρών εφαρμόζεται και σ’αυτές. Ο συνδυασμός των δύο επιπέδων έχει ως αποτέλεσμα που ονομάζεται κρυφή κατάσταση (hidden state). Σε κάθε χρονικό βήμα, η νέα εσωτερική κατάσταση υπολογίζεται με βάση την τρέχουσα είσοδο και την προηγούμενη εσωτερική κατάσταση, συνήθως μέσω μιας μη γραμμικής συνάρτησης όπως η tanh ή η ReLU.

Η εκπαίδευση γίνεται μέσω της μεθόδου οπισθοδιάδοσης του σφάλματος στο χρόνο (Backpropagation Through Time (BPTT)), η οποία επεκτείνει την παραδοσιακή οπισθοδιάδοση για να χειριστεί τις χρονικές εξαρτήσεις.

Τα Vanilla RNNs αντιμετωπίζουν δυσκολίες στην εκμάθηση μακροπρόθεσμων εξαρτήσεων των τιμών της ακολουθίας λόγω του προβλήματος της εξαφανιζόμενης ή εκρηκτικής κλίσης (vanishing/exploding gradient problem). Αυτό σημαίνει ότι καθώς η ακολουθία μεγαλώνει, το δίκτυο δυσκολεύεται να μάθει συσχετισμούς μεταξύ απομακρυσμένων χρονικών βημάτων.

Παρά τους περιορισμούς τους, τα Vanilla RNNs χρησιμοποιούνται σε εφαρμογές όπως:

Ανάλυση συναισθήματος σε κείμενα

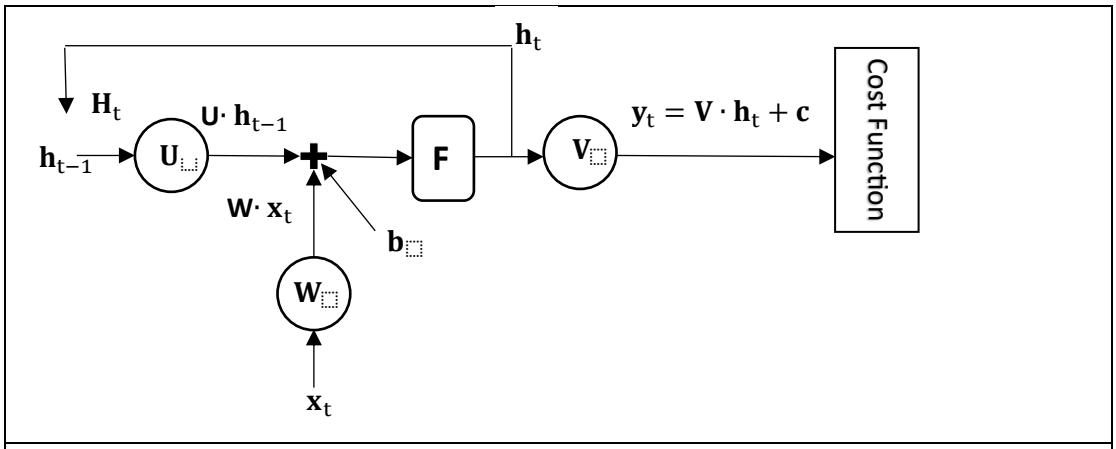
Αναγνώριση φωνής

Πρόβλεψη χρονοσειρών

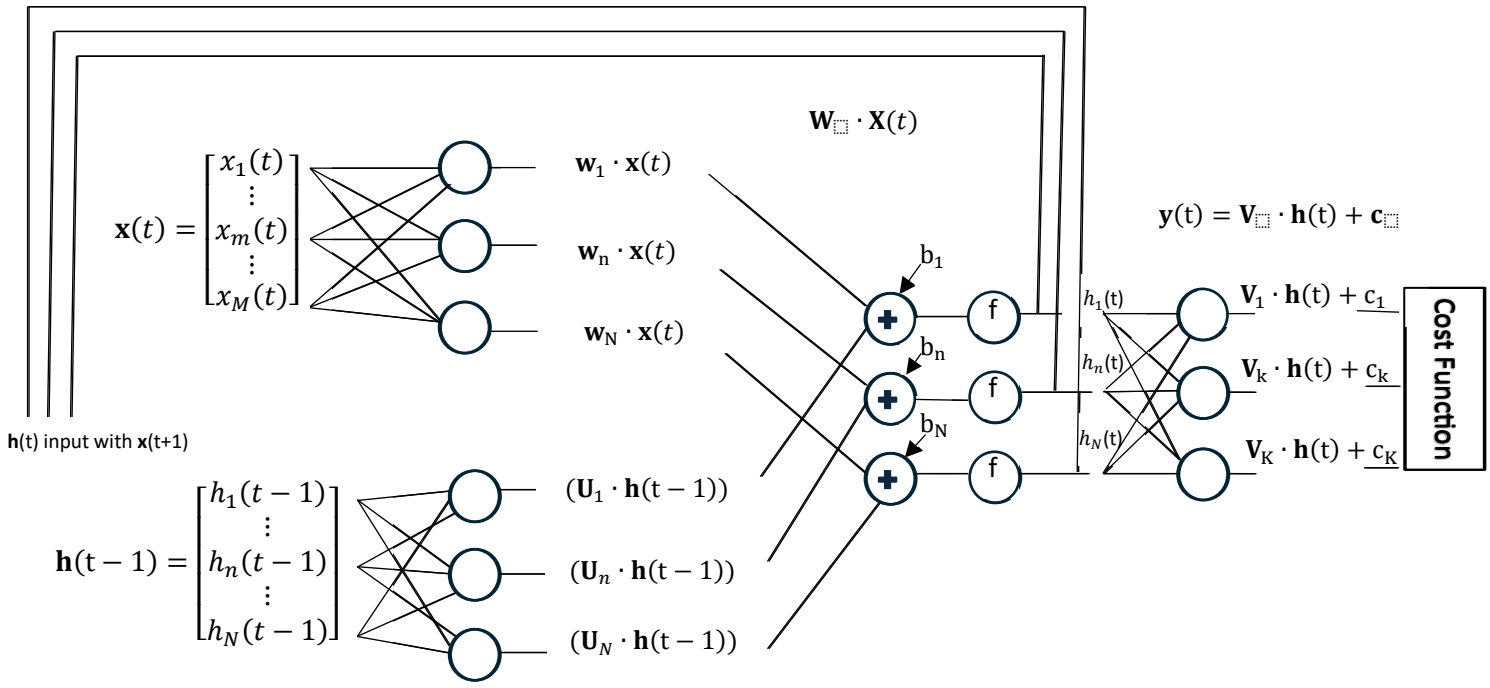
Ανάλυση σειρών γεγονότων

Για πιο σύνθετες εργασίες που απαιτούν την κατανόηση μακροπρόθεσμων εξαρτήσεων, προτιμώνται αρχιτεκτονικές όπως τα Long Short-Term Memory (LSTM) και τα Gated Recurrent Units (GRU), τα οποία έχουν σχεδιαστεί για να αντιμετωπίζουν τα προβλήματα των Vanilla RNNs

Τοπολογία του δικτύου



Διαγραμμα ενός RNN



Αναλυτικό διάγραμμα ενός RNN

Ακολουθούν οι μαθηματικές εκφράσεις του συστήματος καθώς και οι σχέσεις για την διόρθωση των βαρών με την μέθοδο οπισθοδιάδοσης του σφάλματος στο χρόνο.

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \\ \vdots \\ x_M(t) \end{bmatrix},\ \mathbf{x}(t) \in \mathbb{R}^M,\ \mathbf{h}(t) = \begin{bmatrix} h_1(t) \\ \vdots \\ h_n(t) \\ \vdots \\ h_N(t) \end{bmatrix},\ \mathbf{h}(t) \in \mathbb{R}^N$$

$$\mathbf{w}_n = [w_{n1}, w_{n2}, \cdots, w_{nm}, \cdots, w_{nM}],\quad \mathbf{w}_n^{\mathrm{T}} \in \mathbb{R}^M$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \\ \vdots \\ \mathbf{w}_N \end{bmatrix} = \begin{bmatrix} w_{11}, w_{12}, \cdots, w_{1m}, \cdots, w_{1M} \\ \vdots \\ w_{n1}, w_{n2}, \cdots, w_{nm}, \cdots, w_{nM} \\ \vdots \\ w_{N1}, w_{N2}, \cdots, w_{Nm}, \cdots, w_{NM} \end{bmatrix},\ \mathbf{W} \in \mathbb{R}^{N \times M}$$

$$\mathbf{u}_n = [u_{n1}, u_{n2}, \cdots, u_{ni}, \cdots, u_{nN}],\quad \mathbf{u}_n^{\mathrm{T}} \in \mathbb{R}^N$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \\ \vdots \\ \mathbf{u}_N \end{bmatrix} = \begin{bmatrix} u_{11}, u_{12}, \cdots, u_{1i}, \cdots, u_{1N} \\ \vdots \\ u_{n1}, u_{n2}, \cdots, u_{ni}, \cdots, u_{nN} \\ \vdots \\ u_{N1}, u_{N2}, \cdots, u_{Ni}, \cdots, u_{NN} \end{bmatrix},\ \mathbf{U} \in \mathbb{R}^{N \times N}\ ,\ \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \\ \vdots \\ b_N \end{bmatrix},\ \mathbf{b} \in \mathbb{R}^N,$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_k \\ \vdots \\ \mathbf{v}_K \end{bmatrix} = \begin{bmatrix} v_{11}, v_{12}, \cdots, v_{1j}, \cdots, v_{1N} \\ \vdots \\ v_{k1}, v_{k2}, \cdots, v_{kj}, \cdots, v_{kN} \\ \vdots \\ u_{K1}, u_{K2}, \cdots, u_{Kj}, \cdots, u_{KN} \end{bmatrix},\ \mathbf{V} \in \mathbb{R}^{K \times N}\ ,\ \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \\ \vdots \\ c_K \end{bmatrix},\ \mathbf{c} \in \mathbb{R}^K$$

$$\mathbf{h}(\mathrm{t}) = \mathbf{f}(\mathbf{W} \cdot \mathbf{x}(\mathrm{t}) + \mathbf{U} \cdot \mathbf{h}(\mathrm{t}-1) + \mathbf{b}) = \begin{bmatrix} \mathrm{f}(\mathbf{w}_1 \cdot \mathbf{x}(\mathrm{t}) + \mathbf{u}_1 \cdot \mathbf{h}(\mathrm{t}-1) + \mathrm{b}_1) \\ \vdots \\ \mathrm{f}(\mathbf{w}_\mathrm{n} \cdot \mathbf{x}(\mathrm{t}) + \mathbf{u}_\mathrm{n} \cdot \mathbf{h}(\mathrm{t}-1) + \mathrm{b}_\mathrm{n}) \\ \vdots \\ \mathrm{f}(\mathbf{w}_\mathrm{N} \cdot \mathbf{x}(\mathrm{t}) + \mathbf{u}_\mathrm{N} \cdot \mathbf{h}(\mathrm{t}-1) + \mathrm{b}_\mathrm{N}) \end{bmatrix}$$

Όπου f(.) μη γραμμική συνάρτηση ενεργοποίησης π.χ. tanh(.)

$$\mathbf{y}(\mathrm{t}) = \mathbf{V} \cdot \mathbf{h}(\mathrm{t}) + \mathbf{c}$$

Εκπαίδευση με οπισθοδιάδοση του σφάλματος στο χρόνο (BPTT)

Θεωρούμε το κόστος Δ(t) που εξαρτάται από την έξοδο $\mathbf{y}(t)$ και την επιθυμητή τιμή y΄(t) κατά την είσοδο του $\mathbf{x}(t)$. Η συνάρτηση κόστους μπορεί είναι το μέσο τετραγωνικό σφάλμα (RMSE) ή η ετερο-εντροπία (cross entropy) και είναι άθροισμα των όρων Δ(t). Εν τούτοις την κύρια σημασία έχει η ποσότητα Δ(t). Υπολογίζουμε αρχικά την ποσότητα. Γνωρίζουμε από την ανάλυση του BP

$$\mathbf{W}^{\zeta}(\mathrm{t}+1) = \mathbf{W}^{\zeta}(\mathrm{t}) - \rho \sum_{\forall(\mathbf{x},\mathbf{y})} \left(\boldsymbol{\delta}^{\zeta}\right)^{\mathrm{T}} \cdot \left(\mathbf{y}^{\zeta-1}\right)^{\mathrm{T}}$$

$$\boldsymbol{\delta}^{\mathrm{Z}} = (\mathbf{y}^{\mathrm{z}} - \mathbf{y})^{\mathrm{T}} \odot [f'(\sigma_{\mathrm{v}}^{\mathrm{Z}})]_{:\mathbf{v}=\mathbf{1}}^{\mathbf{N_Z}}$$ για το επίπεδο εξόδου

$$\boldsymbol{\delta}^{\zeta} = \boldsymbol{\delta}^{\zeta+1} \cdot \mathbf{W}^{\zeta+1} \odot \left[f'\left(\sigma_{\mathrm{v}}^{\zeta}\right)\right]_{:\mathbf{v}=\mathbf{1}}^{\mathbf{N_Z}}$$ για τα κρυφά επίπεδα

Στο επίπεδο εξόδου Z=2

$\boldsymbol{\delta}^2 = \left[\delta_1^2, \cdots, \delta_k^2, \cdots, \delta_K^2\right]$, $\delta_\kappa^2 = \frac{\partial \Delta}{\partial y_k(t)}$ (Παραγώγιση βαθμωτής συνάρτησης ως προς διάνυσμα).

Στα κρυφά επίπεδα ζ=1, Z=ζ+1=2

Για f(x)=tanh(x) τότε $\tanh'(x) = 1 - \tanh^2(x) = 1 - f(x)^2$ άρα

$$\tanh'(\mathbf{w}_n \cdot \mathbf{x}(t) + \mathbf{u}_n \cdot \mathbf{h}(t-1) + b_n) = (1 - h_n(t)^2)$$

$$\boldsymbol{\delta}^1 = \boldsymbol{\delta}^2 \cdot \mathbf{V} \odot [1 - h_n(t)^2]_{:\boldsymbol{\nu}=\mathbf{1}}^{\mathbf{Nz}}$$

Μας ενδιαφέρει ο υπολογισμός των $\frac{\partial \Delta}{\partial \mathbf{V}}$, $\frac{\partial \Delta}{\partial \mathbf{c}}$, $\frac{\partial \Delta}{\partial \mathbf{W}}$, $\frac{\partial \Delta}{\partial \mathbf{U}}$, $\frac{\partial \Delta}{\partial \mathbf{b}}$.

Με την εφαρμογή των παραπάνω

$\frac{\partial \Delta}{\partial \mathbf{V}} = \boldsymbol{\delta}^{2^T} \cdot \mathbf{h}^T(t) = (\mathbf{h}(t) \cdot \boldsymbol{\delta}^2)^T$ και για τον σταθερό όρο

$\frac{\partial \Delta}{\partial \mathbf{c}} = \boldsymbol{\delta}^{2^T}$

επειδή δεν χρησιμοποιήσαμε επαυξημένα διανύσματα

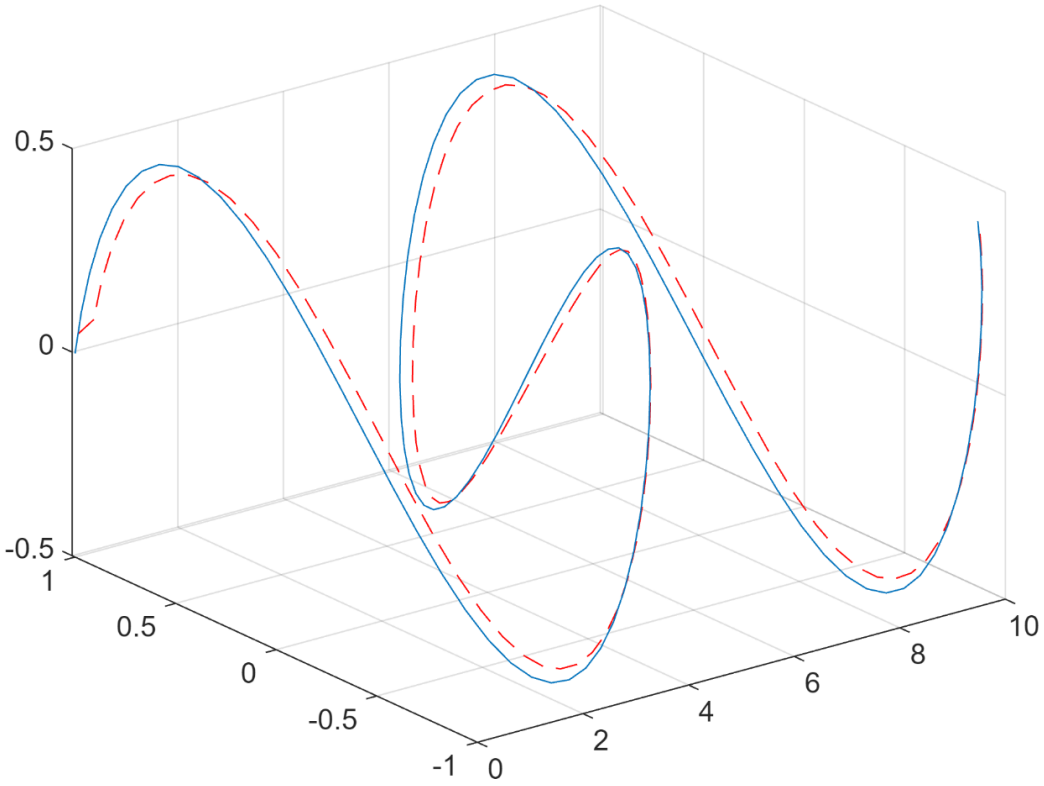
Άρα

$\frac{\partial \Delta}{\partial \mathbf{W}} = \boldsymbol{\delta}^{1^T} \cdot \mathbf{x}^T(t)$ όμοια

$\frac{\partial \Delta}{\partial \mathbf{U}} = \boldsymbol{\delta}^{1^T} \cdot \mathbf{h}^T(t-1)$

$\frac{\partial \Delta}{\partial \mathbf{b}} = \boldsymbol{\delta}^{1^T}$

ΥΛΟΠΟΙΗΣΗ ΣΤΟ MATLAB


0.5
0
-0.5
1
0.5
0
-0.5
-1
0
2
4
6
8
10

ΚΕΦΑΛΑΙΟ 4

ΕΚΠΑΙΔΕΥΣΗ ΧΩΡΙΣ ΕΠΟΠΤΗ

Η εύρεση των συγκεντρώσεων των προτύπων όταν δεν είναι γνωστό το πλήθος τους και η μορφολογία τους είναι ένα πρόβλημα που απασχολεί την Υπολογιστική Νοημοσύνη, την ανάλυση και την εξόρυξη δεδομένων. Για την λύση του προτείνονται ενδιαφέρουσες τεχνικές με δομή ή όχι νευρωνικού δικτύου. Υπάρχουν ποικίλες ενδιαφέρουσες παραλλαγές του όπως η ιεραρχική συσταδοποίηση. Κρίσιμα ερωτήματα είναι το πλήθος των κλάσεων, οι αποστάσεις μεταξύ των προτύπων ή εσωτερική διασπορά και η μορφολογία των συγκεντρώσεων. Η μέτρηση πολλών χαρακτηριστικών και η ποικιλία των προτύπων είναι βασικοί παράγοντες που επιτείνουν την δυσκολία του προβλήματος.

Ακολούθως θα παρουσιάσουμε τρεις μεθόδους εκπαίδευσης χωρίς επόπτη. Οι δύο πρώτες είναι αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την επίλυση σχετικών προβλημάτων ή να χρησιμοποιηθούν ως εργαλεία σε περισσότερο σύνθετες εργασίες. Η τρίτη είναι μία μέθοδος που βασίζεται στην λειτουργία ενός νευρωνικού δικτύου και μπορεί να αποτελέσει βάση για την σχεδίαση ταξινομητών. Η μέθοδος αυτή δίνει την δυνατότητα εποπτείας σε χώρων με περισσότερες των τριών διαστάσεων που η αναπαράστασή τους σε ένα σύστημα αξόνων είναι γεωμετρικά αδύνατη.

4.1 Ο Αλγόριθμος ISODATA ή Κ-Μέσων (k-means ή c-means)

Ο Αλγόριθμος των Κ-μέσων τιτλοφορείται και ως ISODATA ή αλγόριθμος Lloyd's και είναι από τους δημοφιλέστερους αλγορίθμους συσταδοποίησης. Οι ομοιότητα των προτύπων περιγράφεται με την Ευκλείδεια Απόσταση. Κάθε συγκέντρωση καθορίζεται από ένα σημείο-αντιπρόσωπο στο χώρο των προτύπων και σ' αυτήν ανήκουν τα πρότυπα που απέχουν μικρότερη απόσταση από τον αντιπρόσωπο αυτής συγκριτικά με τους αντιπροσώπους των άλλων συγκεντρώσεων. Ο αντιπρόσωπος είναι ο μέσος όρος των

προτύπων που αποτελούν την συγκέντρωση και ονομάζεται κέντρο της κλάσης. Ακολούθως ο αλγόριθμος περιγράφεται για προκαθορισμένο πλήθος συγκεντρώσεων.

Έστω,

- Ι το πλήθος των προτύπων, Κ το πλήθος το συγκεντρώσεων,

- Ν το πλήθος των χαρακτηριστικών,

- $\mathbf{x}_i \in E^N$ πίνακας στήλης που περιγράφει το i πρότυπο, i=1…I

- $\mathbf{c}_k \in E^N$ πίνακας στήλης που περιγράφει τον αντιπρόσωπο (κέντρο) της k κλάσης, k=1…K

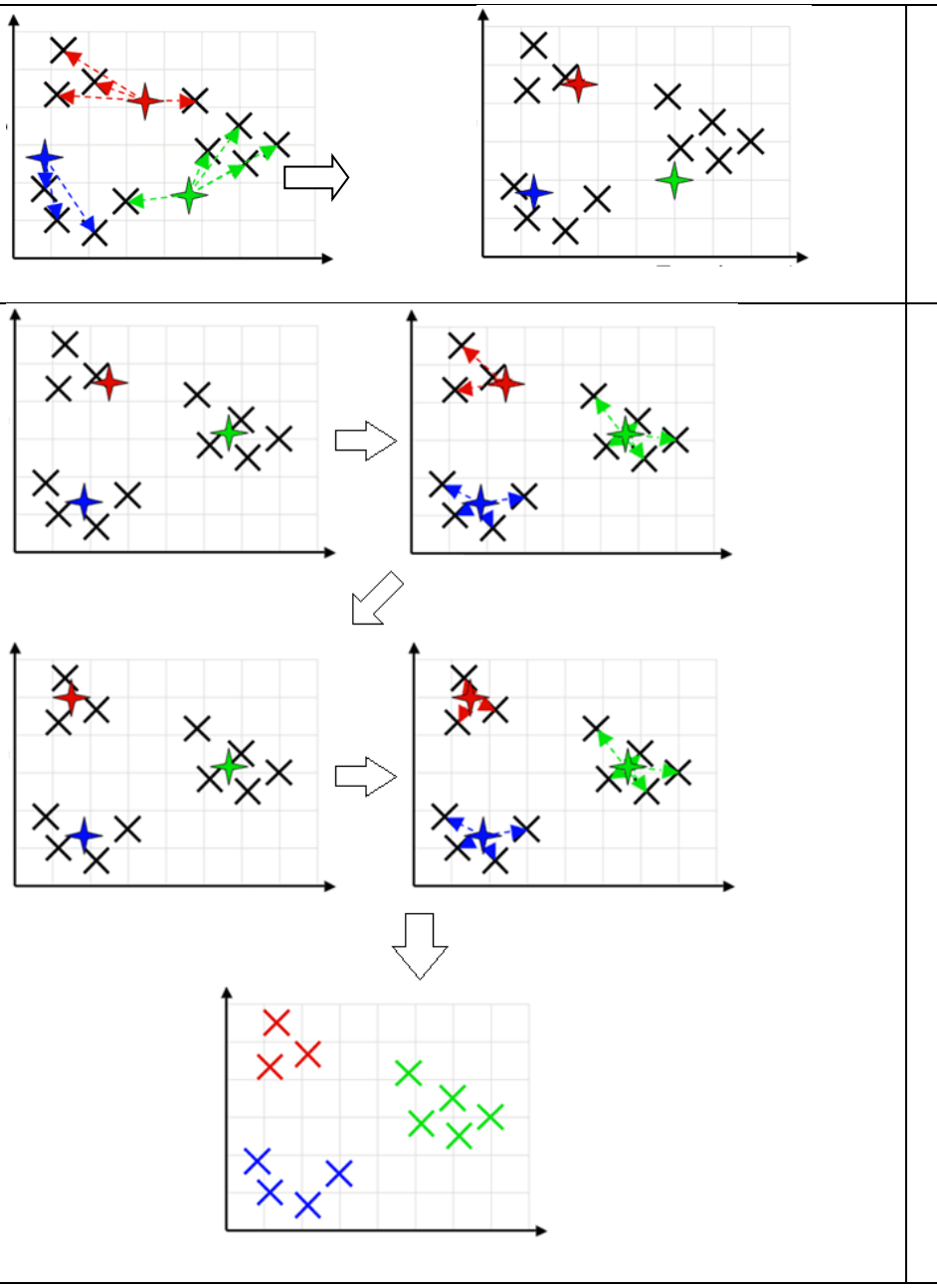
- Β πίνακας Ι × 1 σε κάθε στοιχείο i του οποίου καταχωρείται η συγκέντρωση που ανήκει το i πρότυπο, (πχ b_i=k),

- t μετρητής επανάληψης.

Για t=0 αποδίδονται τυχαίες τιμές στα $\mathbf{c}_k$, για κάθε k.

Για κάθε πρότυπο υπολογίζονται οι αποστάσεις του από τα κέντρα όλων των κλάσεων και καταχωρείται σ' αυτή που έχει κοντινότερο κέντρο στο πρότυπο. Για κάθε κλάση υπολογίζεται το πλήθος P των προτύπων που ανήκουν στην κλάση Υπολογίζεται το κέντρο της κλάσης Αν τα κέντρα παραμείνουν ίδια ο αλγόριθμος τερματίζεται	Επανάληψη Για i από 1 έως I $b_i = \min_k \left\{ \sqrt{\sum_{v=1}^{N} (x_{vi} - c_{vk})^2} \right\}$ Τέλος για. t = t+1 Για k από 1 έως K $P = \sum_{\forall b_i = k} 1$ $c_k = \frac{\sum_{\forall b_i = k} x_i}{P}$ Τέλος για. Αν $c_k(t) = c_k(t+1)$ για κάθε k, τέλος επανάληψης

Στο Σχήμα 4.1.1 φαίνεται η μετακίνηση των κέντρων των συγκεντρώσεων σε διαδοχικές επαναλήψεις του αλγορίθμου.

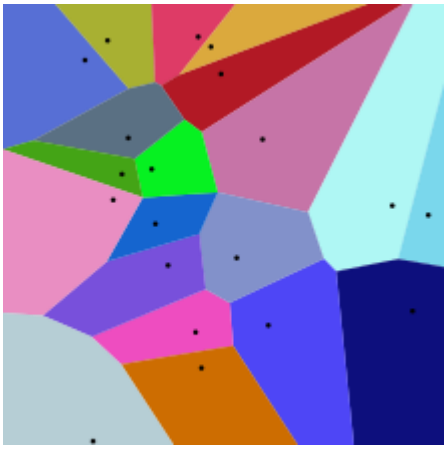


Σχ. 4.1.1.

Είναι δυνατόν επίσης κάποιο ή κάποια κέντρα να μην έχουν κανένα κοντινότερο πρότυπο και ως εκ τούτου να έχουμε λιγότερες από Κ συστάδες. Στην περίπτωση αυτή το κέντρο τοποθετείται κοντά σε πρότυπο που απέχει περισσότερο όλων από το κέντρο της συστάδας του.

Ο k-means θεωρείται ένας γρήγορος, αρκετά αποτελεσματικός αλγόριθμος με ευκολία στην υλοποίηση. Δίνει καλύτερα αποτελέσματα όταν το σύνολο των δεδομένων είναι δομημένο σε διαχώρισιμες συγκεντρώσεις. Μειονέκτημα είναι ότι απαιτεί τον προκαθορισμό του αριθμού των κέντρων των συστάδων. Η τυχαία επιλογή της τιμής αυτών των κέντρων μπορεί να μη μας οδηγήσει σε καλά αποτέλεσμα. Επίσης, ο

αλγόριθμος παρουσιάζει αδυναμία στον χειρισμό δεδομένων με θόρυβο ή ακραίες τιμές και αποτυγχάνει για μη συμπαγές σύνολο δεδομένων (σφαιρικές συγκεντρώσεις). Οι μεσοκάθετοι των κέντρων αποτελούν τις γραμμικές διακριτικές συναρτήσεις των συγκεντρώσεων και συνθέτουν το λεγόμενο διάγραμμα Voronoi (Voronoi tessellation, Σχήμα 4.1.2.)



Σχ.4.1.2.

4.2 Απεικόνιση αλυσίδας κοντινών γειτόνων.

Η *απεικόνιση αλυσίδας* (*chain mapping*) είναι μία μέθοδος απλή αλλά υπολογιστικά δαπανηρή που στοχεύει κυρίως στην εύρεση διαδοχικών γειτονικών σημείων που ονομάζουμε αλυσίδες. Η ιδιαίτερη μορφολογία των αλυσίδων δεν τις καθιστά ανιχνεύσιμες ως συγκεντρώσεις από αλγορίθμους όπως ο ISODATA. Χρησιμοποιείται συχνά στις μεθόδους ιεραρχικής συσταδοποίησης. Οδηγεί στην εποπτεία της κατανομής των προτύπων σε πολυδιάστατους χώρους και μπορεί να χρησιμοποιηθεί για την εκτίμηση του πλήθους και του περιεχομένου των συγκεντρώσεων τους. Σύμφωνα με αυτήν δημιουργούμε μία διαδρομή μεταβαίνοντας από κάθε πρότυπο στο γειτονικότερό του, ξεκινώντας από κάποιο τυχαίο. Καταγράφουμε διαδοχικά τις τιμές του δείκτη του προτύπου, του δείκτη του γείτονά του και της απόστασής τους, χωρίς να μεταβούμε σε πρότυπα από τα οποία έχουμε διέλθει. Δημιουργείται έτσι μια ακολουθία $\mathbf{d}_n=[i,j,D_E(\mathbf{x}_i,\mathbf{x}_j)]$. Θεωρούμε τον όρο m της ακολουθίας με την μεγαλύτερη τιμή απόστασης, $\mathbf{d}_m=\max\{\mathbf{d}_n\}$. Το σύνολο $C_1=\{\mathbf{x}_i/\mathbf{x}_i$ έχει δείκτη την τιμή i για $\mathbf{d}_n=[i,j,D_E(\mathbf{x}_i,\mathbf{x}_j)]$ με $n\leq m\}$ αποτελεί την μία από δύο συγκεντρώσεις των του συνόλου των προτύπων. Η δεύτερη κλάση είναι το σύνολο $C_2=\{\mathbf{x}_j/\mathbf{x}_j$ έχει δείκτη την τιμή j για $\mathbf{d}_n=[i,j,D_E(\mathbf{x}_i,\mathbf{x}_j)]$ με $n\geq m\}$. Θεωρώντας την επόμενη μεγαλύτερη τιμή m' που θα αντιστοιχεί σε στοιχεία μιας από τις δύο συγκεντρώσεις μπορούμε να διαχωρίσουμε αυτήν σε δύο μέρη όπως προηγουμένως κ.ο.κ.

Η μέθοδος απαιτεί τον υπολογισμό του πίνακα των αποστάσεων των προτύπων μεταξύ τους και είναι πολυπλοκότητας $O(n^2)$.

ΠΑΡΑΔΕΙΓΜΑ

Έστω τα πρότυπα

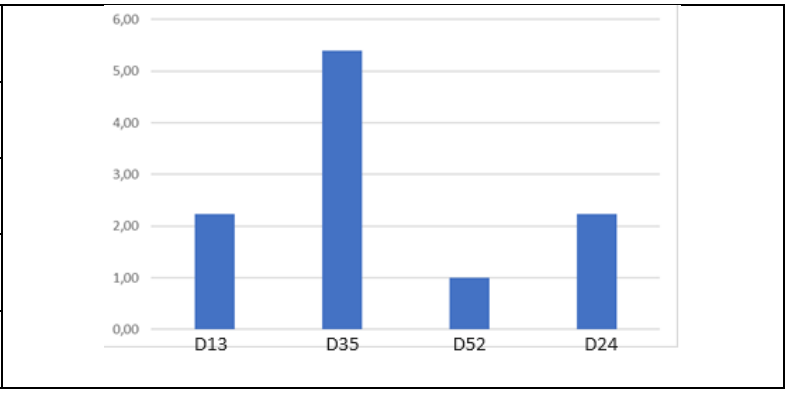
$\mathbf{x}_1 = [10, 8]^T$, $\mathbf{x}_2 = [9, 7]^T$, $\mathbf{x}_3 = [1, 10]^T$, $\mathbf{x}_4 = [2, 8]^T$, $\mathbf{x}_5 = [4, 1]^T$

Ο πίνακας των αποστάσεων τους είναι

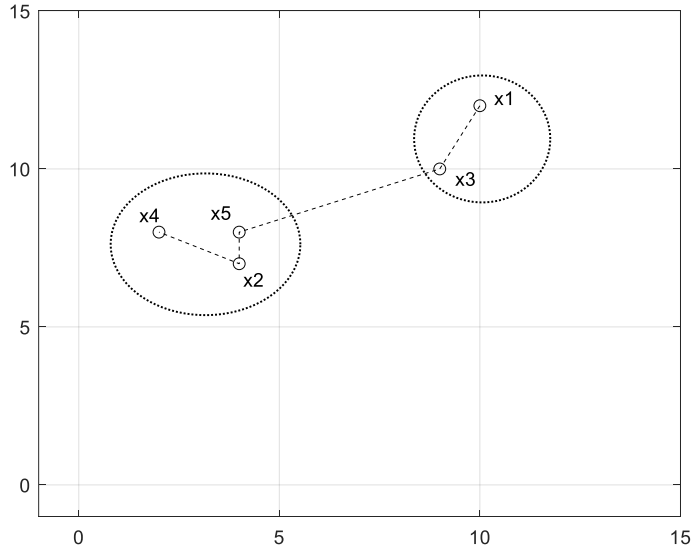
D_{12} 7.81	D_{13} 2.24	D_{14} 8.94	D_{15} 7.21
	D_{23} 5.83	D_{24} 2.24	D_{25} 1.0
		D_{34} 7.28	D_{35} 5.39
			D_{45} 2.0

Ξεκινώντας από το σημείο $\mathbf{x}_1$ μεταβαίνουμε στο γειτονικότερό του κ.ο.κ. Καταγράφεται η ακολουθία

$\mathbf{d}_n$	$\mathbf{D}_E$	i	j
$\mathbf{d}_1$	2.24	1	3
$\mathbf{d}_2$	5.39	3	5
$\mathbf{d}_3$	1.0	5	2
$\mathbf{d}_4$	2.24	2	4



Η τιμή 5.39 είναι η μεγαλύτερη στην στήλη των αποστάσεων και αφορά τον όρο m=2. Για n≤2 ο δείκτης i με τιμές 1, 3 καθορίζει τα πρότυπα της συγκέντρωσης $C_1=\{\mathbf{x}_1, \mathbf{x}_3\}$και για n≥2 ο δείκτης j με τιμές 5,2,4 καθορίζει τα πρότυπα της συγκέντρωσης $C_2=\{\mathbf{x}_5, \mathbf{x}_4, \mathbf{x}_3\}$, Σχήμα 4.2.1.



Σχήμα 4.2.1

4.3 Χάρτης απεικόνισης χαρακτηριστικών με αυτό-οργάνωση

Ο *Χάρτης Απεικόνισης Χαρακτηριστικών με Αυτό-οργάνωση* (*SOFM*: *Self organized Feature Map, SOM*) στοχεύει στην αναπαράσταση των σημείων ενός πολυδιάστατου χώρου που είναι ο αρχικός χώρος των προτύπων σε ένα χώρο λιγότερων διαστάσεων, συνήθως δύο διαστάσεων, διατηρώντας κατά το δυνατόν την τοπολογική-γεωμετρική δομή και τα στατιστικά χαρακτηριστικά των δεδομένων του αρχικού χώρου. Κατά την αναπαράσταση αυτή εκτός από την μείωση των διαστάσεων γίνεται και μείωση του πλήθους των σημείων στον τελικό χώρο που ονομάζουμε *χάρτη απεικόνισης των χαρακτηριστικών*. Βέβαια αφού τα σημεία του χάρτη είναι λιγότερα από αυτά του αρχικού είναι προφανώς ότι συχνά σε ένα σημείο του χάρτη θα απεικονίζονται πολλά σημεία του αρχικού χώρου. Εκείνο όμως που πρωτίστως πρέπει να διασφαλίζεται είναι ότι η γειτνίαση δύο σημείων του αρχικού χώρου θα συνεπάγεται και γειτνίαση των απεικονίσεων τους στο χάρτη (στο ίδιο ή σε γειτονικά σημεία). Αυτό μπορεί να διευκολύνει την οπτικοποίηση και την ανάλυση δεδομένων υψηλών διαστάσεων.

Ένας SOM είναι ένα τεχνητό νευρωνικό δίκτυο ενός επιπέδου που ονομάζεται και *επίπεδο ανταγωνισμού* (*competitive layer*). Οι SOM προτάθηκαν από τον T.Kohonen[1] και συχνά καλούνται *Χάρτες Kohonen* ή *Δίκτυα Kohonen*.

[1] Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". *Biological Cybernetics*. **43** (1): 59–69.

Τα σημεία του επιπέδου ανταγωνισμού (χάρτης) λέγονται *κόμβοι* και λειτουργούν ως νευρώνες, δηλαδή έχουν μεταβλητές-βάρη και είναι συνήθως διατεταγμένοι σε τετραγωνικό ή εξαγωνικό πλέγμα δύο διαστάσεων που λέγεται και *κάναβος (grid)*, Σχήματα 4.3.1, 4.3.2. Ένα δίκτυο Kohonen είναι μια υπολογιστική προσέγγιση που βασίζεται σε ένα μηχανισμό ανταγωνισμού που παρατηρήθηκε σε βιολογικά συστήματα νευρώνων.

Κατά την φάση της εκπαίδευσης του δικτύου Kohonen δεν επιδιώκεται η ελαχιστοποίηση κάποιας συνάρτησης κόστους. Η διαδικασία εκπαίδευσης είναι επαναληπτική με προκαθορισμένο αριθμό επαναλήψεων που καθορίζεται από μια τιμή *T*. Χαρακτηριστικά του αλγορίθμου είναι η έννοια του *νικητή νευρώνα* και της *γειτονιάς* του. Ένας νευρώνας ονομάζεται νικητής κατά την φάση της εκπαίδευσης όπως θα αναλυθεί παρακάτω. Οι γειτονικοί του νευρώνες όπως ορίζονται είναι όσοι βρίσκονται σε μία περιοχή του κανάβου που έχει κέντρο τον νικητή. Αν η περιοχή είναι τετραγωνική σε ένα τετραγωνικό κάναβο η πλευρά της γειτονιάς μπορεί να είναι $2\times d+1$ όπως φαίνεται στο σχήμα 4.3.2. Στον αλγόριθμο χρησιμοποιείται και η παράμετρος εκμάθησης α. Το πλάτος της γειτονιάς και η παράμετρος εκμάθησης μεταβάλλονται κατά την διάρκεια της εκπαίδευσης.

Η εκπαίδευση του νευρωνικού δικτύου γίνεται ως εξής:

Καθορίζεται ο κάναβος του επιπέδου ανταγωνισμού, π.χ. όπως στο Σχήμα 4.3.1 όπου το επίπεδο ανταγωνισμού είναι τοποθετημένο σε έναν κάναβο 8Χ8. Αν και ο κάναβος είναι δισδιάστατος, οι νευρώνες έχουν αριθμηθεί με ένα δείκτη k που παίρνει ακέραιες τιμές από 0 έως 63. Αν $\boldsymbol{x} \in R^N$ τότε κάθε νευρώνας k στον κάναβο έχει ένα διάνυσμα βαρών $\boldsymbol{w_k} = [w_{1k} \dots w_{nk} \dots w_{Nk}]^T$.

Εκτελούνται τα παρακάτω βήματα:

Ορίζουμε ένα μετρητή επανάληψης t=0,1,…,Τ. Αποδίδονται τυχαίες τιμές στα βάρη των νευρώνων $w_{nk}(0)$ με τιμές που προκύπτουν από την πρόσθεση μικρών τυχαίων αριθμών στη μέση τιμή των ανυσμάτων $\mathbf{x}$ όλου το συνόλου εκπαίδευσης.

Αρχικοποιείται η παράμετρος εκμάθησης α(0) με μια μικρή θετική τιμή, συνήθως μεταξύ 0.2 και 0.5.

Αρχικοποιείται η παράμετρος d(0) με την τιμή 4, που είναι ίση με το μισό του εύρους του κανάβου.

Επιλέγεται ένα τυχαίο άνυσμα $\mathbf{x}(t)$ από το σύνολο εκπαίδευσης.

Υπολογίζεται η έξοδος $o_k(t)$ του νευρώνα k σύμφωνα με τη σχέση:

$$o_k(t) = \|\mathbf{x}(t) - \boldsymbol{w_k}(t)\| = \sqrt{\sum_{n=1}^{N}(x_n(t) - w_{nk}(t))^2}. \quad (4.3.1)$$

Είναι φανερό πως η o_k είναι η Ευκλείδεια απόσταση μεταξύ των $\mathbf{x}(t)$ και $\mathbf{w}_k(t)$.

Ο νευρώνας c ανακηρύσσεται νικητής εάν ικανοποιείται η συνθήκη $o_c(t) = \min\{o_k(t)\}$. Εάν οι έξοδοι δύο νευρώνων είναι ίσες, τότε κατά σύμβαση επιλέγεται αυτός με το μικρότερο δείκτη.

Τα βάρη $\mathbf{w}_k$ ανανεώνονται σύμφωνα με τις παρακάτω σχέσεις:

$$w_{nk}(t+1) = w_{nk}(t) + \Delta w_{nk}(t) \quad (4.3.3)$$

$$\Delta w_{nk} = \begin{cases} \alpha(t) \cdot (x_n(t) - w_{nk}(t)) & \text{αν } k \in N_c \\ 0 & \text{αν } k \notin N_c \end{cases} \quad (4.3.4)$$

όπου N_c το σύνολο των δεικτών των νευρώνων που βρίσκονται μέσα σε ένα τετράγωνο του δισδιάστατου κανάβου με κέντρο το νευρώνα νικητή και πλευρά $2 \times d(t)+1$ (γειτονιά του νικητή νευρώνα).

Αυξάνεται η μεταβλητή επανάληψης κατά ένα και αποδίδονται νέες τιμές στις μεταβλητές $\alpha(t)$, $d(t)$ σύμφωνα με τις σχέσεις:

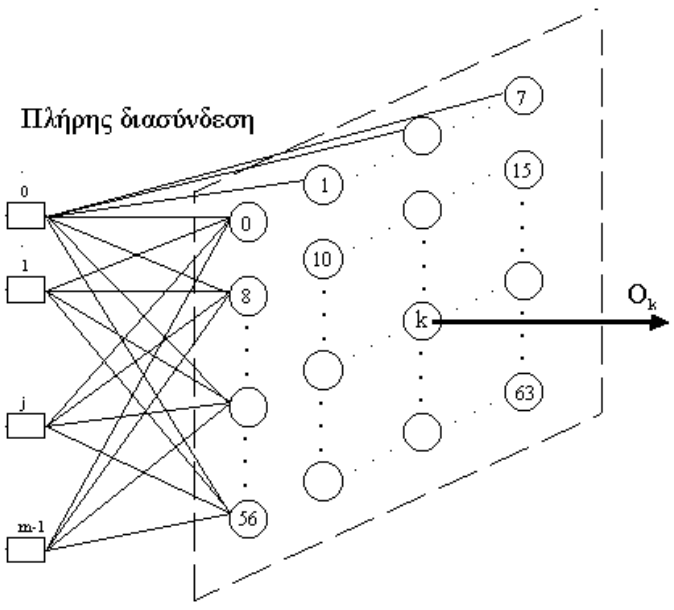
$$a(t) = a(0) \cdot (1 - \frac{t}{T}) \qquad (4.3.5)$$

$$d(t) = d(0) \cdot (1 - \frac{t}{T}) \qquad (4.3.6)$$

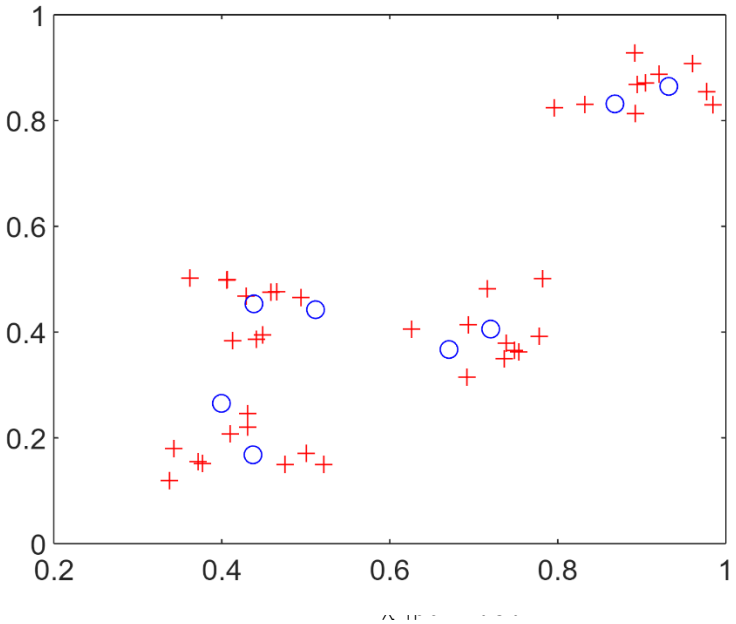
Τα βήματα 4 έως 8 επαναλαμβάνονται έως ότου η μεταβλητή *t* πάρει τη μέγιστη τελική τιμή *T*. Είναι φανερό πως οι μεταβλητές *α(t)* και *d(t)* συγκλίνουν στο μηδέν καθώς η *t* τείνει στην τιμή *T*.

Μετά την εκπαίδευση κάθε άνυσμα εισόδου του SOM αντιστοιχίζεται-απεικονίζεται από τον νευρώνα που ανακηρύσσει νικητή. Κάθε νευρώνας του επιπέδου εξόδου αντιπροσωπεύει μία ομάδα προτύπων (cluster). Πρότυπα με μεγάλη ομοιότητα (γειτονικά στον αρχικό χώρο) αντιπροσωπεύονται από τον ίδιο νευρώνα ή γειτονικούς στο επίπεδο ανταγωνισμού. Νευρώνες που αντιπροσωπεύουν μεγάλο πλήθος σημείων του αρχικού χώρου είναι σημαντικοί ως κέντρα συγκεντρώσεων (κυρίαρχοι νευρώνες), Σχήμα 4.3.3. Νευρώνες που αντιπροσωπεύουν μικρό πλήθος σημείων μπορούν να απορριφθούν. Τα βάρη των νευρώνων έχουν τις διαστάσεις του αρχικού χώρου και ως εκ τούτου οι κυρίαρχοι νευρώνες αποτελούν μία αντιπροσωπευτική υπογραφή των αρχικών πολυπληθών δεδομένων.

Αν στους νευρώνες αποδοθούν ετικέτες (Labels) τότε το νευρωνικό δίκτυο Kohonen μπορεί να είναι προ-επεξεργαστής ενός συστήματος ταξινόμησης. Τα δίκτυα *Kohonen* αποτελούν εξαιρετικό εργαλείο ανάλυσης δεδομένων σε πολλά επιστημονικά πεδία. Εξελιγμένες επεκτάσεις τους είναι οι *αναπτυσσόμενοι χάρτες αυτοοργάνωσης* (*growing self-organizing map*, GSOM), οι *σύμμορφοι χάρτες* (*conformal maps*), οι *χρονικά προσαρμοστικοί χάρτες αυτο-οργάνωσης*,(*time adaptive self-organizing map, TASOM*), οι *προσανατολισμένοι και επεκτάσιμοι χάρτες* (*oriented and scalable maps*).



Σχήμα 4.3.1



Σχήμα 4.3.3: Οκτώ νευρώνες (μπλε κύκλοι) αντιπροσωπεύουν τα δεδομένα.

ΚΕΦΑΛΑΙΟ 5

ΑΝΑΛΥΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Μια σημαντική εργασία σε ένα σύστημα αναγνώρισης είναι η ανάλυση των μετρούμενων χαρακτηριστικών των προτύπων. Με την ανάλυση των χαρακτηριστικών πετυχαίνουμε την αξιολόγηση τους, την επιλογή των καταλληλότερων για την ταξινόμηση των προτύπων και την ελάττωση του πλήθους τους. Για παράδειγμα ένα χαρακτηριστικό με κοντινές τιμές μέσα σε κάθε μία κλάση και σαφώς διαφορετικές τιμές μεταξύ των κλάσεων είναι «καλό» χαρακτηριστικό ταξινόμησης. Με κατάλληλες μεθόδους ανάλυσης χαρακτηριστικών μπορούμε να οδηγηθούμε και σε νέα χαρακτηριστικά που προκύπτουν από τον γραμμικό συνδυασμό των αρχικών και είναι καταλληλότερα από αυτά. Ακολούθως θα δούμε μεθόδους ανάλυσης χαρακτηριστικών σε συστήματα εκμάθησης με και χωρίς επόπτη.

5.1 Ανάλυση χαρακτηριστικών στην εκπαίδευση με επόπτη

Μια σημαντική εργασία σε ένα σύστημα αναγνώρισης είναι η επιλογή ενός μικρού συνόλου κατάλληλων χαρακτηριστικών από ένα πολύ μεγαλύτερο. Η επιλογή των ισχυρών χαρακτηριστικών είναι κρίσιμη για την αποτελεσματικότητα της ταξινόμησης. Για να ελαττωθεί το πλήθος των χαρακτηριστικών με την επιλογή ενός συνόλου «καλών χαρακτηριστικών» που βοηθούν την ταξινόμηση, εκτελείται μια διαδικασία αξιολόγησης με κριτήρια την *διαχωριστικότητα (separability)* και την *συσχέτιση (correlation)* των χαρακτηριστικών.

Για τον έλεγχο της ικανότητας διαχωρισμού καθορίζεται ο *παράγοντας διαχωριστικότητας (separability factor)* S_{avg}^{ℓ} μεταξύ δύο ισοπίθανων κλάσεων Α και Β για κάθε χαρακτηριστικό *ν* από τη σχέση:

$$S_\nu = \frac{|\mu_\nu^A - \mu_\nu^B|}{\sqrt{(\sigma_\nu^A)^2 + (\sigma_\nu^B)^2}} \qquad (5.1)$$

όπου μ_ν^A, μ_ν^B, είναι οι μέσες τιμές και σ_ν^A, σ_ν^B οι τυπικές αποκλίσεις σε κάθε κλάση. Μεγάλη διαχωριστικότητα σημαίνει ότι το χαρακτηριστικό έχει μεγάλη ικανότητα να διαχωρίζει τις δύο κλάσεις μεταξύ τους. Ισοδύναμα χρησιμοποιείται η ποσότητα

$$S_\nu^2 = \frac{(\mu_\nu^A - \mu_\nu^B)^2}{(\sigma_\nu^A)^2 + (\sigma_\nu^B)^2}$$

γνωστή ως *λόγος διάκρισης κατά Fisher*.

Για την ανάλυση της συσχέτισης των χαρακτηριστικών υπολογίζεται ο *παράγοντας συσχέτισης (correlation factor)* για κάθε δύο χαρακτηριστικά *ν* και *λ* προτύπων που ανήκουν στην ίδια τάξη *p*, σύμφωνα με τη σχέση

$$C_{\nu\lambda}^p = \frac{\frac{1}{K_p}\sum_{\kappa=1}^{K_p}(x_{\nu\kappa} - \mu_\nu)(x_{\lambda\kappa} - \mu_\lambda)}{\sigma_\nu \sigma_\lambda} = \frac{\sigma_{\nu\lambda}}{\sigma_\nu \sigma_\lambda} \qquad (5.2)$$

όπου K_p είναι το πλήθος των στοιχείων της κλάσης *p*, $x_{\nu\kappa}$ και $x_{\lambda\kappa}$ οι τιμές των χαρακτηριστικών, $\mu_\nu, \mu_\lambda, \sigma_\nu$ και σ_λ οι μέσες τιμές και οι τυπικές αποκλίσεις των τιμών των χαρακτηριστικών στην τάξη *p*. Ο αριθμητής $\sigma_{\nu\lambda}$ του κλάσματος λέγεται *ετεροσυσχέτιση* (*cross-correlation*) των τιμών των δύο χαρακτηριστικών. Ο παράγοντας συσχέτισης μετρά την ομοιότητα μεταξύ των δύο χαρακτηριστικών και παίρνει τιμές μεταξύ *−1* και *+1*. Μια τιμή κοντά στο *+1* ή στο *−1* σημαίνει ότι τα δύο χαρακτηριστικά συσχετίζονται πολύ ή συσχετίζονται αντίστροφα, αντίστοιχα. Μια τιμή κοντά στο μηδέν δείχνει ότι τα χαρακτηριστικά είναι κατά πολύ ασυσχέτιστα. Όταν δύο χαρακτηριστικά έχουν μεγάλη συσχέτιση μπορούμε να απορρίψουμε ένα εκ των δύο.

ΠΑΡΑΔΕΙΓΜΑ

Δίνονται οι πίνακες τεσσάρων προτύπων Π_1, Π_2, Π_3, Π_4

$$x_1^A = \begin{bmatrix}0\\1\\0\end{bmatrix}, x_2^A = \begin{bmatrix}1\\0.5\\1\end{bmatrix}, x_3^B = \begin{bmatrix}0.5\\-1\\0\end{bmatrix}, x_4^B = \begin{bmatrix}1\\0\\1\end{bmatrix}$$

Απορρίψτε το χειρότερο χαρακτηριστικό με το κριτήριο της διαχωριστικότητας.

$$\mu^A = \begin{bmatrix}(0+1)/2\\(1+0.5)/2\\(0+1)/2\end{bmatrix} = \begin{bmatrix}1/2\\3/4\\1/2\end{bmatrix}, \quad \mu^B = \begin{bmatrix}(0.5+1)/2\\(-1+0)/2\\(0+1)/2\end{bmatrix} = \begin{bmatrix}3/4\\-1/2\\1/2\end{bmatrix}$$

$$(\sigma_1^A)^2 = \frac{(0-1/2)^2 + (1-1/2)^2}{2} = 1/4 \qquad (\sigma_2^A)^2 = \frac{(1-3/4)^2 + (1/2-3/4)^2}{2} = 1/16$$

$$(\sigma_3^A)^2 = \frac{(0-{}^1\!/_2)^2 + (1-{}^1\!/_2)^2}{2} = {}^1\!/_4 \qquad (\sigma_1^B)^2 = \frac{({}^1\!/_2 - {}^3\!/_4)^2 + (1-{}^3\!/_4)^2}{2} = {}^1\!/_{16}$$

$$(\sigma_2^B)^2 = \frac{(-1+{}^1\!/_2)^2 + (0-{}^1\!/_2)^2}{2} = {}^1\!/_4 \qquad (\sigma_3^B)^2 = \frac{(0-{}^1\!/_2)^2 + (1-{}^1\!/_2)^2}{2} = {}^1\!/_4$$

$$S_{AB}^1 = \frac{|{}^1\!/_2 - {}^3\!/_4|}{\sqrt{{}^1\!/_4 + {}^1\!/_{16}}},\ S_{AB}^2 = \frac{|{}^3\!/_4 + {}^1\!/_2|}{\sqrt{{}^1\!/_{16} + {}^1\!/_4}},\ S_{AB}^3 = \frac{|{}^1\!/_2 - {}^1\!/_2|}{\sqrt{{}^1\!/_4 + {}^1\!/_4}} = 0$$

Ο συντελεστής διαχωριστικότητας για το 3ο χαρακτηριστικό είναι μηδέν και γι' αυτό μπορεί να απορριφθεί.

5.2 Ανάλυση χαρακτηριστικών στην εκπαίδευση χωρίς επόπτη

Για την ανάλυση χαρακτηριστικών στην εκπαίδευση χωρίς επόπτη, θα περιγράψουμε τον μετασχηματισμό Karhunen-Loeve (**KLT**)ή μέθοδο *ανάλυσης κύριων συνιστωσών* (*ΑΚΣ*), (*PCA*: *Principal Components Analysis).* Η ΑΚΣ είναι ένα μέσο με το οποίο «γεννώνται» νέα χαρακτηριστικά από τα υπάρχοντα, ασυσχέτιστα μεταξύ τους. Με τον μετασχηματισμό είναι δυνατή η αντιπροσώπευση του αρχικού συνόλου των χαρακτηριστικών με άλλα λιγότερα και αποτελεσματικότερα που διατηρούν σημαντικό μέρος της πληροφορίας των δεδομένων. Με τη χρήση της ΑΚΣ επιτυγχάνεται η αύξηση της διασποράς των τιμών των χαρακτηριστικών, η μείωση των διαστάσεων και η συμπίεση της πληροφορίας.

Για ένα σύνολο *K* ανυσμάτων σε N-διαστάσεων χώρο χαρακτηριστικών έστω ότι $\boldsymbol{x}_\kappa$ είναι το διάνυσμα που δίνεται από την σχέση.

$$\boldsymbol{x}_\kappa = |x_{1\kappa} \quad x_{2\kappa} \quad \dots \quad x_{\nu\kappa} \quad \dots \quad x_{N\kappa}|^T \qquad (5.2.1)$$

Με μέσο όρο $\boldsymbol{\mu}_x = E[\boldsymbol{x}_\kappa] = \frac{1}{K}\sum_{\kappa=1}^{K} \boldsymbol{x}_\kappa = \boldsymbol{0}$ (5.2.2)

όπου *E[.]* η μαθηματική προσδοκία. Αν ο μέσος όρος $\boldsymbol{\mu}_x$ δεν είναι το μηδενικό διάνυσμα τότε τροποποιούμε τις τιμές αφαιρώντας από αυτές τον μέσο όρο τους ώστε οι τιμές που θα προκύψουν να έχουν μηδενικό μέσο όρο. Ο πίνακας συνδιασποράς $\boldsymbol{C_x}$ όλων των ανυσμάτων του συνόλου εκπαίδευσης δίνεται από τη σχέση (με την προϋπόθεση ότι ο μέσος όρος είναι μηδέν).

$$\boldsymbol{C}_x = E[\boldsymbol{x}\boldsymbol{x}^T] = \frac{1}{K}\sum_{\kappa=1}^{K} \boldsymbol{x}_\kappa \boldsymbol{x}_\kappa{}^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1\nu} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2\nu} & \cdots & \sigma_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{\lambda 1} & \sigma_{\lambda 2} & \cdots & \sigma_{\lambda\nu} & \cdots & \sigma_{\lambda N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{N\nu} & \cdots & \sigma_N^2 \end{bmatrix} \qquad (5.2.3)$$

Κάθε διαγώνια τιμή σ_ν^2 του πίνακα συνδιασποράς εκφράζει τη διασπορά των τιμών $x_{\nu\kappa}$ (άξονας ν), ενώ οι υπόλοιπες τη συνδιασπορά $\sigma_{\nu\kappa}$ μεταξύ δύο διαφορετικών

μεταβλητών $x_{μκ}$ και $x_{νκ}$. Ο πίνακας $\mathbf{C}_x$ είναι συμμετρικός με θετικές τιμές. Είναι φανερό πως για τη διαδικασία της ταξινόμησης είναι επιθυμητές μεγάλες τιμές των διαγωνίων τιμών σ_v^2 διότι μαρτυρούν ένα μεγάλο άπλωμα (spreading) των δεδομένων, ενώ είναι επιθυμητές μηδενικές τιμές για τις συνδιασπορές ώστε οι μεταβλητές να είναι μεταξύ τους ασυσχέτιστες. Για να το επιτύχουμε αυτό αναζητούμε ένα τέτοιο μετασχηματισμό $\boldsymbol{W}$ που θα εκφράζεται από ένα πίνακα $N×N$ ώστε τα διανύσματα $\boldsymbol{y} = \boldsymbol{W}^T \cdot \boldsymbol{x}$ να έχουν πίνακα συνδιασποράς $\boldsymbol{C}_y$ με μηδενικές ετεροσυσχετίσεις. Θα ισχύει:

$$\mathbf{C}_y=E[\mathbf{y}\mathbf{y}^T]=E[\mathbf{W}^T\mathbf{x}(\mathbf{W}^T\mathbf{x})^T]= E[\mathbf{W}^T\mathbf{x}\mathbf{x}^T\mathbf{W}]= \mathbf{W}^TE[\mathbf{x}\mathbf{x}^T]\mathbf{W}= \mathbf{W}^T\mathbf{C}_x\mathbf{W} \qquad (5.2.4)$$

Ο πίνακας $\mathbf{C}_x$ είναι συμμετρικός θετικά ορισμένος και έχει ιδιοτιμές $\lambda_1>\lambda_2>\ldots>\lambda_v>\ldots>\lambda_N>0$ με αντίστοιχα μοναδιαία ιδιοδιανύσματα $\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_v,\ldots,\mathbf{e}_N$. Αν $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$ και $\mathbf{Q}=[\mathbf{q}_1,\ldots,\mathbf{q}_v,\ldots,\mathbf{q}_N]$ ισχύει: $\mathbf{Q}^T\cdot\mathbf{Q}=\mathbf{I}$ και $\mathbf{Q}^T \cdot \mathbf{C}_x \cdot \mathbf{Q} = \boldsymbol{\Lambda}$. (Δες Παράρτημα)
Αν ο πίνακας $\mathbf{W}=\mathbf{Q}$ τότε από την (5.2.4) συνεπάγεται ότι $\mathbf{C}_y=\boldsymbol{\Lambda}$.

$$\boldsymbol{C}_y = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} = \boldsymbol{\Lambda} \ , \quad \lambda_1 > \lambda_2 > \ldots > \lambda_v > \ldots > \lambda_N \qquad (5.2.5)$$

Τα διανύσματα $\boldsymbol{y_\kappa} = |y_{1\kappa} \ \ y_{2\kappa} \ \ \ldots \ \ y_{v\kappa} \ \ \ldots \ \ y_{N\kappa}|^T$ που προκύπτουν από την σχέση $\boldsymbol{y} = \boldsymbol{Q}^T \cdot \boldsymbol{x}$ θα είναι $\boldsymbol{y_\kappa} = |\boldsymbol{x_\kappa^T} \cdot \boldsymbol{q_1}, \ \boldsymbol{x_\kappa^T} \cdot \boldsymbol{q_2} \ \ldots \ \boldsymbol{x_\kappa^T} \cdot \boldsymbol{q_v} \ \ldots \ \boldsymbol{x_\kappa^T} \cdot \boldsymbol{q_N}|^T$. Τα μοναδιαία ιδιονύσματα $\mathbf{q}_v$ του $\mathbf{C}_x$ είναι κάθετα μεταξύ τους και αποτελούν ένα ορθογώνιο σύστημα συντεταγμένων οι άξονες του οποίου ονομάζονται *Κύριες Συνιστώσες* (*Principal Components*). Κάθε μοναδιαίο ίδιοδιάνυσμα $\mathbf{q}_v$ περιγράφει έναν άξονα επί του οποίου προβάλλεται κάθε διάνυσμα $\mathbf{x}_\kappa$, η προβολή του είναι $y_{v\kappa} = \boldsymbol{x_\kappa^T} \cdot \boldsymbol{q_v}$. Οι τιμές των προβολών επί των κύριων αξόνων είναι αναλλοίωτες (αμετάβλητες) ως προς την περιστροφή των δεδομένων από κάποιο μετασχηματισμό στροφής.

Η τιμή $\lambda_v=\sigma'^2_v$ είναι η διασπορά των τιμών $y_{v\kappa}$ κατά τον ν άξονα. Το ιδιοδιάνυσμα $\mathbf{q}_v$ είναι ένα "αποτελεσματικό" χαρακτηριστικό ενώ η τιμή λ_v αξιολογεί τη σπουδαιότητά του. Εκτός του ότι τα διανύσματα $\mathbf{y}_\kappa$ περιγράφουν τα πρότυπα με τιμές ασυσχέτιστες, με τον μετασχηματισμό μπορούμε να μειώσουμε και το πλήθος των τιμών που τα περιγράφουν αποδεχόμενοι μικρή απώλεια πληροφορίας (συμπίεση). Προς τούτο χρησιμοποιούμε τις προβολές y_v που αντιστοιχούν σε ιδιοδιανύσματα με τις μεγαλύτερες αντίστοιχες ιδιοτιμές, αποδεχόμενοι ένα μέσο τετραγωνικό σφάλμα $\overline{\varepsilon^2}(M) = \sum_{v=M+1}^{N} \lambda_v$, όπου Μ ο δείκτης του ιδιοδιανύσματος μετά από το οποίο διαγράφουμε τις προβολές. Για παράδειγμα αν $\lambda_1>\lambda_2>\lambda_3>\lambda_4>\lambda_5$ οι ιδιοτιμές $\mathbf{q}_1,\mathbf{q}_2,\mathbf{q}_3,\mathbf{q}_4,\mathbf{q}_5$ τα αντίστοιχα ιδιοδιανύσματα και $\mathbf{y}=[y_1,y_2,y_3,y_4,y_5]^T$ διάνυσμα των προβολών που προέκυψαν από τον μετασχηματισμό $\boldsymbol{y} = \boldsymbol{Q}^T \cdot \boldsymbol{x}$, περιγράφουμε τα δεδομένα

με διανύσματα $\mathbf{y}'=[y_1, y_2, y_3]^T$ αποδεχόμενοι ένα μέσο τετραγωνικό σφάλμα $\overline{\varepsilon^2}(3) = \sum_{\nu=3+1}^{N} \lambda_\nu = \lambda_4 + \lambda_5$.

Ο KLT μας προσφέρει ασυσχέτιστα χαρακτηριστικά και δυνατότητα μείωσης του πλήθους τους. Μας προσφέρει επιπρόσθετα ανεξαρτησία ως προς την περιστροφή και την μετατόπιση διότι οι τιμές των προβολών επί των κύριων αξόνων είναι αναλλοίωτες (αμετάβλητες) ως προς την περιστροφή των δεδομένων από μετασχηματισμό στροφής και από μετατόπιση. Η ιδιότητα αυτή είναι εξαιρετικά χρήσιμη για δεδομένα μορφών σε δυαδικές ψηφιακές εικόνες (π.χ. χαρακτήρες, γράμματα).

ΠΑΡΑΔΕΙΓΜΑ

$x_1 = \begin{bmatrix}1\\0\end{bmatrix}$, $x_2 = \begin{bmatrix}1\\1\end{bmatrix}$, $x_3 = \begin{bmatrix}2\\0\end{bmatrix}$, $x_4 = \begin{bmatrix}4\\3\end{bmatrix}$, $x_5 = \begin{bmatrix}5\\2\end{bmatrix}$

$$\mu = \begin{bmatrix}(1+1+2+4+5)/5\\(0+1+0+3+2)/5\end{bmatrix} = \begin{bmatrix}\frac{13}{5}\\\frac{6}{5}\end{bmatrix} = \begin{bmatrix}2{,}6\\1{,}2\end{bmatrix}$$

$$\sigma_1^2 = \frac{(1\text{-}2.6)^2 + (1\text{-}2.6)^2 + (2\text{-}2.6)^2 + (4\text{-}2.6)^2 + (5\text{-}2.6)^2}{5} = 2.64$$

$$\sigma_2^2 = \frac{(0\text{-}1.2)^2 + (1\text{-}1.2)^2 + (0\text{-}1.2)^2 + (3\text{-}1.2)^2 + (2\text{-}1.2)^2}{5} = 1.36$$

$$\sigma_{21} = \frac{(0\text{-}1.2)(1\text{-}2.6) + (1\text{-}1.2)(1\text{-}2.6) + (0\text{-}1.2)(2\text{-}2.6) + (3\text{-}1.2)(4\text{-}2.6) + (2\text{-}1.2)(5\text{-}2.6)}{5} = 1.48$$

$\sigma_{12} = \sigma_{21} = 1.48$

$$C_x = \begin{vmatrix}\sigma_1^2 & \sigma_{12}\\\sigma_{21} & \sigma_{22}^2\end{vmatrix} = \begin{bmatrix}2.64 & 1.48\\1.48 & 1.36\end{bmatrix}$$

$$\det(C_x - \lambda \cdot I) = \det\begin{bmatrix}2.64-\lambda & 1.48\\1.48 & 1.36-\lambda\end{bmatrix} = (2.64-\lambda)\cdot(1.36-\lambda) - 1.48^2 =$$

$1.36 \cdot 2.64 - 1.36 \cdot \lambda - 2.64 \cdot \lambda + \lambda^2 - 1.48^2 =$

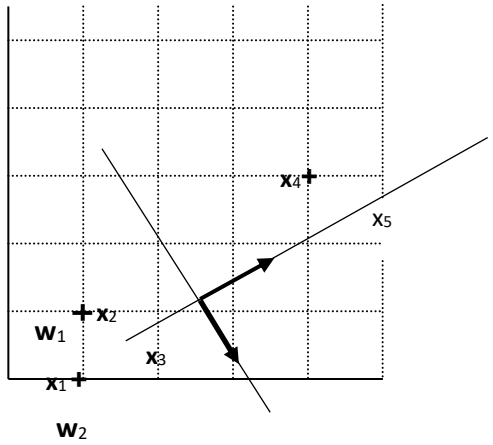
$\lambda^2 - 4 \cdot \lambda - 1.4$ που είναι το χαρακτηριστικό πολυώνυμο.

Για $\lambda^2 - 4 \cdot \lambda + 1.4 = 0$ προκύπτουν οι ρίζες $\lambda_1 = 3.61$, $\lambda_2 = 0.39$ που είναι οι ιδιοτιμές του C_x. Τα αντίστοιχα ιδιοδυανύσματα $\mathbf{w}_1$, $\mathbf{w}_2$ ικανοποιούν τις σχέσεις $(\mathbf{C}_x - \lambda_1 \cdot \mathbf{I}) \cdot \mathbf{q}_1 = 0$ και $(\mathbf{C}_x - \lambda_2 \cdot \mathbf{I}) \cdot \mathbf{q}_2 = 0$.

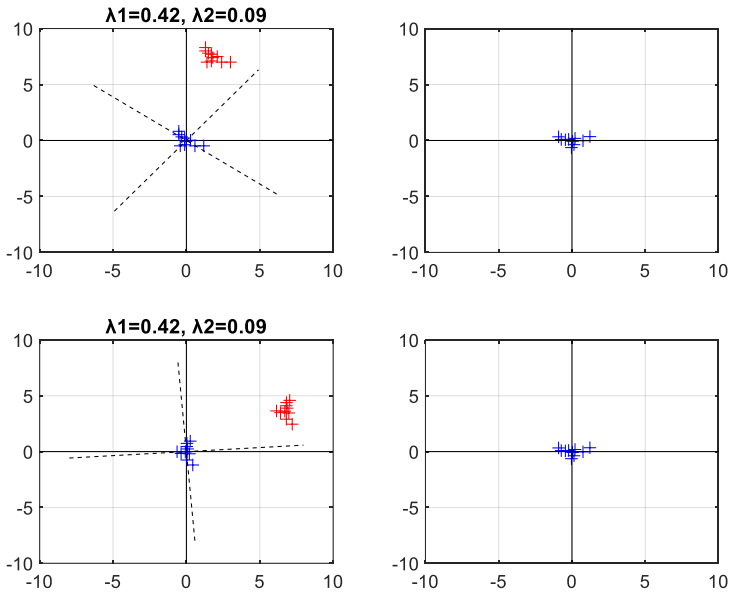
Για το $\mathbf{q}_1$ θα έχουμε αναλυτικά

$$\begin{vmatrix}2.64-3.61 & 1.48\\1.48 & 1.36-3.61\end{vmatrix} \cdot \begin{vmatrix}q_{11}\\q_{21}\end{vmatrix} = 0 \Rightarrow \begin{matrix}-0.97 \cdot q_{11} + 1.48 \cdot q_{21} = 0\\1.48 \cdot q - 2.25 \cdot q_{21} = 0\end{matrix}$$

από το παραπάνω αόριστο σύστημα λύση για $q_{11}=\alpha$ είναι $\mathbf{q}_1=[\alpha, \alpha \cdot 0.65]^T$. Ακόμη για να ισχύει $\|\boldsymbol{q}_1\|=1 \Leftrightarrow \sqrt{\alpha^2 + (\alpha \cdot 1.65)^2} = 1 \Leftrightarrow \alpha = \pm 1/\sqrt{1+1.65^2} \approx .84$, άρα ένα μοναδιαίο ιδιοδιάνυσμα είναι το $\mathbf{q}_1=[0.84, 0.55]^T$. Όμοια για την ιδιοτιμή λ_2 προκύπτει $\mathbf{q}_2=[0.55, -0.84]^T$. Τα διανύσματα είναι κάθετα μεταξύ τους όπως αναμενόταν. Στο Σχ.5.2.1 φαίνονται γεωμετρικά τα αποτελέσματα της μεθόδου.



ΠΑΡΑΔΕΙΓΜΑ ΑΝΕΞΑΡΤΗΣΙΑΣ ΑΠΟ ΣΤΡΟΦΗ



Ανάλυση Κύριων Συνιστωσών με Νευρωνικό Δίκτυο.

Ο KLT είναι μία αναλυτική διαδικασία για την εκτίμηση των ιδιοτιμών και των ιδιοδιανυσμάτων από τον πίνακα συνδιασποράς των δεδομένων $\boldsymbol{C_x}$. Ειδικά, εάν το πλήθος των διαστάσεων του χώρου των ανυσμάτων είναι μεγάλο ο υπολογισμός και ο χειρισμός του πίνακα $\boldsymbol{C_x}$ είναι ανέφικτος.

Για τον παραπάνω λόγο αντί του KLT χρησιμοποιούμε στο πρώτο μέρος του ταξινομητή ένα γραμμικό νευρωνικό δίκτυο ενός επιπέδου (single-layer linear feed forward neural net), το οποίο επιτελεί την ΑΚΣ (Σχ. 5.2.1). Το δίκτυο αυτό εκπαιδεύεται χωρίς επόπτη (unsupervised) από τον γενικευμένο αλγόριθμο του Hebb που βασίζεται στον κανόνα εκμάθησης του Hebb. Με τον αλγόριθμο αυτό το νευρωνικό δίκτυο συγκλίνει με πιθανότητα ένα στα ιδιοδιανύσματα ενός πίνακα συνδιασποράς $\boldsymbol{C_x}$. Ο υπολογισμός του $\boldsymbol{C_x}$ δεν είναι αναγκαίος επειδή τα ιδιοδιανύσματα προκύπτουν κατευθείαν από τα δεδομένα. Η εκπαίδευση του νευρωνικού δικτύου γίνεται ως εξής:

Έστω ότι

— $t=1,2,3,\ldots$ μία μεταβλητή για τη μέτρηση της επανάληψης της διαδικασίας,

— $\boldsymbol{x}(t)$ το διάνυσμα εισόδου στο νευρωνικό δίκτυο τη χρονική στιγμή t, με συνιστώσες $x_\nu(t)$, $\nu=0,1,\ldots,N$,

— K το πλήθος των νευρώνων ($K \leq N$) που ως εκ τούτου είναι και το πλήθος των επιθυμητών κύριων συνιστωσών,

— κ ένας δείκτης που αποδίδεται στους νευρώνες $\kappa=1,\ldots,K$,

— $w_{\kappa\nu}(t)$ η τιμή του βάρους της σύναψης που συνδέει τον κ νευρώνα με την ν είσοδο κατά την επανάληψη t,

— $\boldsymbol{W}_{K \times N}(t)$ ο πίνακας των $w_{\kappa\nu}(t)$ και

— $\boldsymbol{y}(t)$ το άνυσμα εξόδου του δικτύου κατά την επανάληψη t, με συνιστώσες $y_\kappa(t)$.

Εκτελούνται τα ακόλουθα βήματα υπολογισμών:

Βήμα 1. Αφαιρείται πρώτα το διάνυσμα $\boldsymbol{\mu_x}$ της μέσης τιμής από κάθε στοιχείο του συνόλου εκπαίδευσης. Με τον τρόπο αυτό οι τιμές που προκύπτουν έχουν μηδενικό διάνυσμα μέσης τιμής.

Βήμα 2. Αποδίδονται αρχικά ($t=0$) στα βάρη $w_{\kappa\nu}(0)$ των συνάψεων μικρές τυχαίες τιμές και στην παράμετρο γ του ρυθμού εκπαίδευσης μικρή θετική τιμή (π.χ. $\gamma=0.007$).

Βήμα 3. Υπολογίζεται το διάνυσμα εξόδου $\boldsymbol{y}(t)$ με συνιστώσες $y_\kappa(t)$ για $\nu=1,\ldots,N$, $\kappa=1,\ldots,K$ από τη σχέση

$$y_k(t) = \sum_{\nu=1}^{N} w_{\kappa\nu}(t) \cdot x_\nu(t), \quad \text{ή} \quad \boldsymbol{y}(t)=\boldsymbol{W}(t)\ \boldsymbol{x}(t) \qquad (63)$$

και η ποσότητα $\Delta w_{\kappa\nu}(t)$ από την σχέση

$$\Delta w_{\kappa\nu}(t) = \gamma\left(\underbrace{y_\kappa(t)\cdot x_\nu(t)}_{(\alpha)} - \underbrace{y_\kappa(t)\cdot\sum_{\lambda=1}^{\kappa} w_{\lambda\nu}(t)\cdot y_\lambda(t)}_{(\beta)}\right) \qquad (64)$$

Ο όρος (α) εκφράζει τον απλό κανόνα του Hebb, που λέει ότι εάν δύο νευρώνες που βρίσκονται στα άκρα μιας σύναψης (σύνδεσης), ενεργοποιούνται συγχρόνως, τότε η ισχύς της σύναψης αυξάνεται. Αλλιώς

εξασθενεί ή εκφυλίζεται. Ο όρος (β) επιβάλλει ένα όριο στην αύξηση της σύναψης. Η τιμή του $w_{κν}(t+1)$ προσαρμόζεται σύμφωνα με τη σχέση

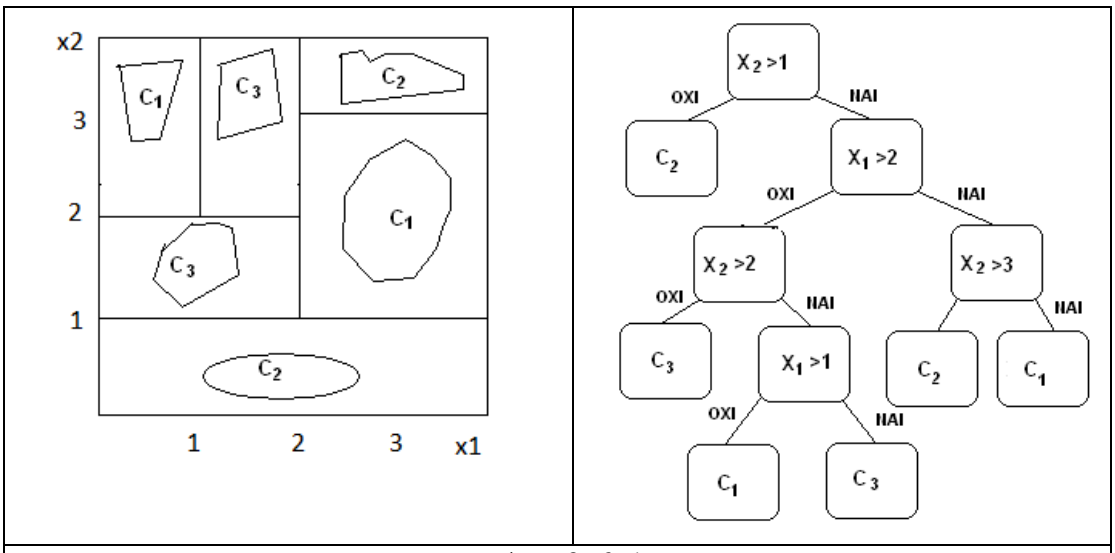
$$W_{κν}(t+1) = w_{κν}(t) + \Delta w_{κν} \qquad (5.2.7)$$

Βήμα 4. Αύξηση της μεταβλητής t κατά ένα και επανάληψη από το *βήμα 3* έως ότου τα βάρη των συνάψεων $w_{κν}$ φθάσουν στη σταθερή κατάσταση.

Μετά τη φάση της εκπαίδευσης τα βάρη $w_{κν}$ του $κ$ νευρώνα συγκλίνουν στην $ν$ συνιστώσα του ιδιοδιανύσματος που αντιστοιχεί στην $κ$ ιδιοτιμή του πίνακα συνδιασποράς $\boldsymbol{C_x}$. Με άλλα λόγια οι γραμμές του πίνακα $\boldsymbol{W}$ έχουν προσεγγίσει τα πρώτα K ιδιοδιανύσματα του $\boldsymbol{C_x}$, ταξινομημένες σε φθίνουσα σειρά.

3.6 Δένδρα απόφασης

Τα δένδρα απόφασης (decision trees) αποτελούν μια ευρεία κατηγορία τεχνικών μη γραμμικών ταξινομητών στην οποία ο χώρος των χαρακτηριστικών διαιρείται σε περιοχές που αντιστοιχούν στις επιθυμητές κλάσεις. Η απόφαση της ταξινόμησης ενός προτύπου προκύπτει από τις απαντήσεις σε ερωτήματα που υποβάλλονται σύμφωνα με μία δενδρική δομή. Θα παρουσιάσουμε ακολούθως μια αντιπροσωπευτική και δημοφιλή τεχνική των δένδρων απόφασης κατά την οποία ο χώρος χωρίζεται σε υπερ-παραλληλόγραμμα. Για ένα υπό ταξινόμηση πρότυπο κάθε ερώτημα αφορά ένα χαρακτηριστικό x_v και είναι της μορφής «ισχύει $x_v < t$» όπου t κατάλληλη τιμή κατωφλίου. Τα δένδρα απόφασης αυτής της μορφής ονομάζονται συνήθη δυαδικά δένδρα ταξινόμησης - ΣΔΔΤ(ordinary binary classification trees - OBCT). Η λειτουργία ενός ΣΔΔΤ γίνεται εύκολα αντιληπτή με ένα παράδειγμα δύο χαρακτηριστικών και κλάσεις κατανεμημένες όπως στο Σχ.[3.6.1]. Με απλή γεωμετρική παρατήρηση μπορούμε να οδηγηθούμε στο δένδρο απόφασης του που φαίνεται στο ίδιο σχήμα.



Σχήμα 3.6.1

Ο οπτικός διαχωρισμός βέβαια δεν είναι δυνατός σε χώρους υψηλότερης του τρία διάστασης και σε κάθε περίπτωση η εκτέλεση από Υπολογιστές απαιτεί την αλγοριθμική και μαθηματικολογική διατύπωση. Παρατηρώντας τον χώρο των προτύπων του παραδείγματος και τις ορθογώνιες περιοχές που περικλείουν τις κλάσεις, σε σχέση με το αντίστοιχο δένδρο απόφασης διαπιστώνουμε ότι σε κάθε κόμβο t τίθεται ως ερώτημα μια συνθήκη η ισχύς της οποίας χωρίζει ένα υποσύνολο X_t του χώρου τον προτύπων σε δύο μέρη. Ο ριζικός κόμβος αφορά όλο σύνολο εκπαίδευσης. Η απάντηση του ερωτήματος μπορεί να είναι θετική (Ναι) ή αρνητική (Όχι) και από αυτή καθορίζονται τα δύο υποσύνονα X_{tN}, X_{tO} που ικανοποιούν τις σχέσεις

$$X_{tN} \cup X_{tO} = X_t \quad \text{και} \quad X_{tN} \cap X_{tO} = \emptyset$$

Πρέπει ακόμη να καθορισθούν κριτήρια ώστε:

- ο διαχωρισμός να είναι βέλτιστος
- να τερματίζεται ο διαχωρισμός ενός κόμβου
- να αντιστοιχίζεται σε ένα κόμβο-φύλλο μία κλάση

Μια ερώτηση κόμβου είναι της μορφής $x_v>\alpha$, όπου x_v η τιμή του ν χαρακτηριστικού ενός προτύπου και α μια τιμή κατωφλίου. Το α μπορεί να πάρει απεριόριστες τιμές στο πεδίο μεταβολής του, εν τούτοις αν Λ_t είναι το πλήθος των προτύπων που ανήκουν στον X_t, αρκεί ένα πεπερασμένο πλήθος τιμών $\alpha_{v\lambda}$, λ=1,2,…, Λ_t , για να διαχωριστεί ο χώρος. Η τιμή $\alpha_{v\lambda}$ θα επιλεγεί ώστε να ικανοποιείται το κριτήριο του βέλτιστου διαχωρισμού.

Καθορισμός του κριτηρίου διαχωρισμού

Ο διαχωρισμός του X_t πρέπει να είναι τέτοιος ώστε σε ένα τουλάχιστον από τα δύο μέρη του να κυριαρχούν πληθυσμιακά τα πρότυπα μιας μόνο κλάσης, να είναι δηλαδή «καθαρό» σύμφωνα με την ορολογία των δυαδικών δένδρων απόφασης. Ως μέτρο της καθαρότητας ενός κόμβου μπορεί να χρησιμοποιηθεί η εντροπία (μέση πληροφορία) γνωστή από την θεωρία της πληροφορίας. Για έναν κόμβο t η εντροπία $I(t)$ δίνεται από την σχέση

$$I(t) = -\sum_{i=1}^{M} P(C_i|t) \cdot log_2 P(C_i|t)$$

Όπου $P(C_i|t)$ η πιθανότητα ένα πρότυπο του χώρου $X(t)$ να ανήκει στην κλάση C_i. Υπενθυμίζεται ότι $0\cdot\log_2 0=0$ και ότι $I(t)$ μεγιστοποιείται αν οι τιμές $P(C_i|t)$είναι ίσες μεταξύ τους. Οι πιθανότητες $P(C_i|t)$ εκτιμώνται με βάση τα ποσοστά Λ_t^i/Λ_t όπου Λ_t το πλήθος των προτύπων του κόμβου και Λ_t^i το πληθος όσων εξ αυτών ανήκουν στην κλάση C_i. Με τον διαχωρισμό του κόμβου η καθαρότητά του μεταβάλλεται κατά $\Delta I(t)$ σύμφωνα με τη σχέση

$$\Delta I(t) = I(t) - \frac{\Lambda_{tN}}{\Lambda_t} \cdot I(t_N) - \frac{\Lambda_{tO}}{\Lambda_t} \cdot I(t_O)$$

Η ποσότητα $\Delta I(t)$ υπολογίζεται για όλα τα χαρακτηριστικά και τις κατάλληλες τιμές κατωφλίου των προτύπων που ανήκουν στους χώρους X_t, X_{tN}, X_{tO}. Ακολούθως επιλέγεται το χαρακτηριστικό και το κατώφλι που μεγιστοποιούν την $\Delta I(t)$ και προκύπτει το ερώτημα του κόμβου t.

Καθορισμός του κριτηρίου τερματισμού του διαχωρισμού

Ο διαχωρισμός του χώρου X_t, είναι δυνατόν να τερματίσει όταν η μέγιστη τιμή $\Delta I(t)$ είναι μικρότερη από ένα προκαθορισμένο όριο. Εναλλακτικά μπορεί να τερματιστεί αν ο πληθυσμός Λ_t είναι αρκετά μικρός ή όταν τα πρότυπα ανήκουν σε μία κλάση ($\Delta I(t)=0$).

Καθορισμός κριτηρίου αντιστοίχισης μιας κλάσης στον κόμβο-φύλλο

Σύμφωνα με τα παραπάνω μετά τον τερματισμό στο κόμβο t αποδίδεται ως όνομα το όνομα της κλάσης των προτύπων που υπερτερούν πληθυσμιακά στον χώρο X_t.

3.7 Ταξινομητές με βάση τον κανόνα πιθανοτήτων του Bayes

Στις προσεγγίσεις που παρουσιάσαμε στα προηγούμενα κεφάλαια δεν λάβαμε υπόψη την πιθανότητα εμφάνισης των προτύπων του συνόλου εκπαίδευσης. Στο κεφάλαιο αυτό θα ασχοληθούμε με ταξινομητές που βασίζονται σ αυτή την πιθανότητα και στην ελαχιστοποίηση του σφάλματος ταξινόμησης που προκύπτει από αυτήν.

Θα ξεκινήσουμε την παρουσίαση με το πρόβλημα των δύο κλάσεων C_1, C_2. Στόχος μας αρχικά είναι ο υπολογισμός της πιθανότητας το πρότυπο να ανήκει στη κλάση C_1 (ή C_2) δεδομένου ότι έχει την τιμή **x**, δηλαδή να βρούμε τις υπό συνθήκη πιθανότητες $P(C_1|\mathbf{x})$ και $P(C_2|\mathbf{x})$. Αν $P(C_1|\mathbf{x}) > P(C_2|\mathbf{x})$ το πρότυπο ανήκει στη κλάση C_1, διαφορετικά στη C_2. Ακολούθως αναζητούμε εκείνες της περιοχές του χώρου των προτύπων για τις οποίες ελαχιστοποιείται το σφάλμα ταξινόμησης με βάση την παραπάνω θεώρηση. Από το κανόνα του Bayes ισχύει ότι

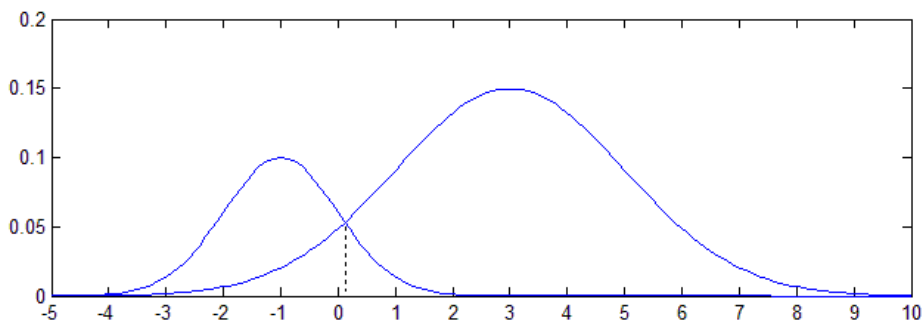
$$P(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)\cdot P(C_1)}{p(\boldsymbol{x})} \quad \text{και} \quad P(C_2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_2)\cdot P(C_2)}{p(\boldsymbol{x})} \qquad [3.7.1]$$

Όπου $P(C_1)$ και $P(C_2)$ οι πιθανότητες το πρότυπο **x** να ανήκει στις κλάσεις C_1, C_2 αντίστοιχα και $p(\mathbf{x}|C_1)$, $p(\mathbf{x}|C_2)$ συναρτήσεις πυκνότητας πιθανότητας (σ.π.π.) για την γενική περίπτωση που το **x** είναι διανυσματική συνεχής τυχαία μεταβλητή. Για τις σ.π.π. θα χρησιμοποιούμε το μικρό p και για τις πιθανότητες κεφαλαίο P.

Άρα το **x** ανήκει

στη C_1 αν $p(\mathbf{x}|C_1)\cdot P(C_1) > p(\mathbf{x}|C_2)\cdot P(C_2)$, στη C_2 αν $p(\mathbf{x}|C_1)\cdot P(C_1) < p(\mathbf{x}|C_2)\cdot P(C_2)$
[3.7.2]

Στα προηγούμενα σημειώνεται ότι οι πιθανότητες $P(C_1)$ και $P(C_2)$ υπολογίζονται προσεγγιστικά από τις σχέσεις $P(C_1) = N_1/N$, $P(C_2) = N_2/N$ με N_1, N_2 το πλήθος των προτύπων, οι ποσότητες $p(\mathbf{x}|C_1)$, $p(\mathbf{x}|C_2)$ θεωρούνται γνωστές ή εκτιμώνται από το σύνολο εκπαίδευσης. Από την σχέση [3.7.2] βρίσκουμε ακολούθως τις περιοχές R_1 και R_2 που διαχωρίζουν τον χώρο των προτύπων. Για παράδειγμα για δύο κλάσεις με $P(C_1) = 0.25$ και $P(C_2)=0.75$, $p(\mathbf{x}|C_1)$, $p(\mathbf{x}|C_2)$ κανονικές κατανομές με μέσους όρους $\mu_1=-1$, $\mu_2=3$ και διασπορές $\sigma_1=1$, $\sigma_2=2$ οι περιοχές R_1 και R_2 θα είναι όπως στο Σχ._



Σχ.[]

Παρακάτω θα δείξουμε ότι με βάση τα παραπάνω το λάθος ταξινόμησης ελαχιστοποιείται. Αν R_1 η περιοχή που περιλαμβάνει τις τιμές του $\boldsymbol{x}$ που ταξινομούνται στη κλάση C_1 και R_2 η περιοχή που περιλαμβάνει τις τιμές του $\boldsymbol{x}$ που ταξινομούνται στη κλάση C_2 το σφάλμα θα συμβαίνει αν το $\boldsymbol{x}$ ανήκει στη R_1 και το πρότυπο στη κλάση C_2 ή αν το $\boldsymbol{x}$ ανήκει στη R_2 και το πρότυπο στη κλάση C_1 και η πιθανότητα P_e να συμβεί θα είναι

$$P_e = P(\boldsymbol{x} \in R_2, C_1) + P(\boldsymbol{x} \in R_1, C_2) =$$

$$P(C_1) \cdot P(\boldsymbol{x} \in R_2 | C_1) + P(C_2) \cdot P(\boldsymbol{x} \in R_1 | C_2) =$$

$$P(C_1) \cdot \int_{R_2} p(\boldsymbol{x}|C_1) d\boldsymbol{x} + P(C_2) \cdot \int_{R_1} p(\boldsymbol{x}|C_2) d\boldsymbol{x} =$$

$$\int_{R_2} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_1} p(C_2|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} \qquad (3.7.3)$$

Ισχύει ακόμη ότι

$$\int_{-\infty}^{+\infty} p(\boldsymbol{x}|C_1) d\boldsymbol{x} = 1 \Rightarrow \int_{R_1} p(\boldsymbol{x}|C_1) d\boldsymbol{x} + \int_{R_2} p(\boldsymbol{x}|C_1) d\boldsymbol{x} = 1 \Rightarrow$$

$$\frac{1}{P(C_1)} \int_{R_1} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} + \frac{1}{P(C_1)} \int_{R_2} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} = 1 \Rightarrow$$

$$\int_{R_1} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_2} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} = P(C_1) \Rightarrow$$

$$\int_{R_2} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} = P(C_1) - \int_{R_1} p(C_1|\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} \qquad (3.7.4)$$

Από τις σχέσεις (3.7.1) και (3.7.2) προκύπτει

$$P_e = P(C_1) - \int_{R_1} \left(P(C_1|\boldsymbol{x}) - P(C_2|\boldsymbol{x})\right) p(\boldsymbol{x}) d\boldsymbol{x} \qquad \textit{έλεγξε το πρόσημο}$$

Που σημαίνει ότι το σφάλμα ελαχιστοποιείται όταν η περιοχή R_1 είναι τέτοια ώστε το ολοκλήρωμα να είναι θετικό δηλαδή $P(C_1|\boldsymbol{x}) > P(C_2|\boldsymbol{x})$. Το αυτό αντίστοιχα ισχύει και για την R_2.

Στην περίπτωση των πολλών κλάσεων το πρότυπο ανήκει στην κλάση C_i για την οποία ισχύει $P(C_i|\boldsymbol{x}) > P(C_j|\boldsymbol{x})$ $\mu\varepsilon\ i \neq j$

Συναρτήσεις διάκρισης και διαχωριστικές επιφάνειες

Στην περίπτωση που απόφαση λαμβάνεται με κριτήριο την ελαχιστοποίηση του σφάλματος ταξινόμησης για δύο κλάσεις η σχέση μπορούμε σύμφωνα με τα προηγούμενα να ισχυριστούμε ότι το πρότυπο ανήκει στην κλάση K_1 αν $P(C_1|\boldsymbol{x}) - P(C_2|\boldsymbol{x}) > 0$ διαφορετικά ανήκει στην κλάση K_2. Θα παρουσιάσουμε ακολούθως την δυνατότητα

ταξινόμησης του προτύπου μέσω του προσδιορισμού μιας διαχωριστικής επιφάνειας του χώρου που θα ελαχιστοποιεί το σφάλμα ταξινόμησης στην περίπτωση που η συνάρτηση πυκνότητας πιθανότητας (σ.π.π., probability density function, pdf) είναι η κανονική κατανομή (Gaussian).

Για μονοδιάστατη τυχαία μεταβλητή η σ.π.π. της κανονική μεταβλητής δίνεται από τον τύπο

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Όπου μ και σ η μέση τιμή και η διασπορά της κατανομής.

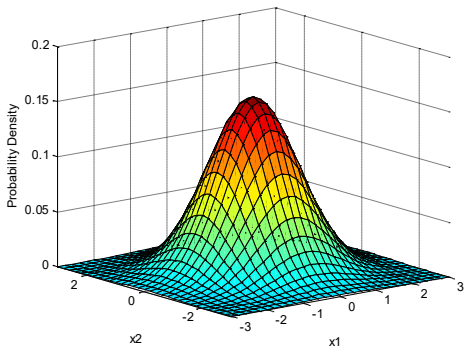
Στην γενική περίπτωση που η τυχαία μεταβλητή είναι διάνυσμα $\mathbf{x}$ ενός N διάστατου χώρου αντίστοιχη σ.π.π. δίδεται από τη σχέση

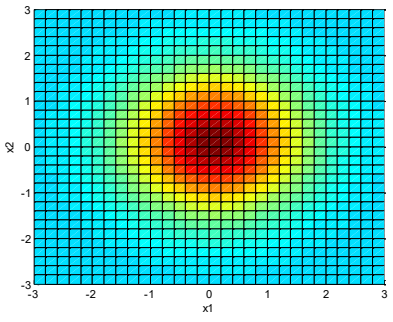
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{C}|^{1/2}} e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}}$$

Όπου $\boldsymbol{\mu}$ μέση τιμή και $\boldsymbol{C}$ ό πίνακας συνδιασποράς της σ.π.π. και $|\boldsymbol{C}|$ η ορίζουσά του. Προσοχή η χρήση του συμβόλου $\boldsymbol{C}$ δεν αναφέρεται πλέον στην κλάση Κ αλλά στο πίνακα συνδιασποράς.
Για λόγους γεωμετρικής απεικόνισης θα θεωρήσουμε την περίπτωση που η τυχαία μεταβλητή ανήκει στον δισδιάστατο χώρο. Ο πίνακας συνδιασποράς αποτελείται από τις διασπορές και τις ετεροσυσχετίσεις των γραμμών της διανυσματικής μεταβλητής $\mathbf{x}=[x_1, x_2]^T$ και είναι της μορφής

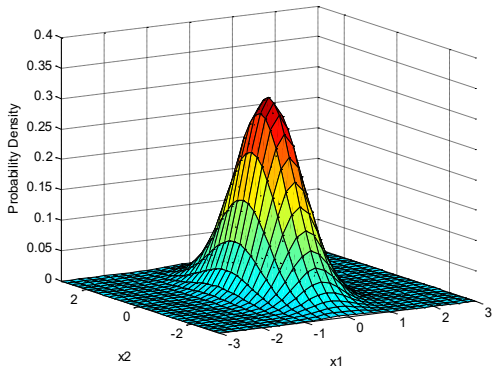
$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

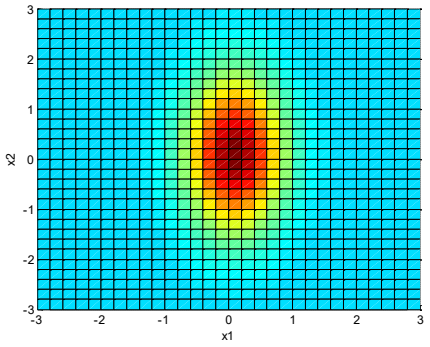
Στα ακόλουθα σχήματα βλέπουμε τις σ.π.π κανονικών κατανομών για διαφορετικούς πίνακες συνδιασποράς και μέσους όρους.



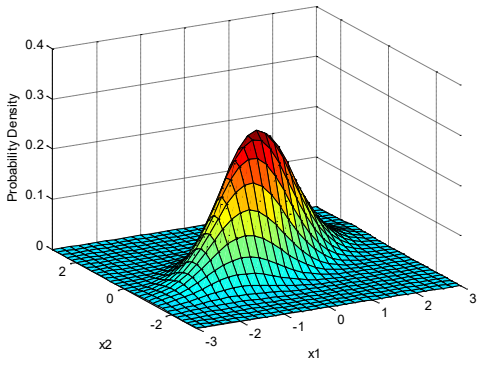


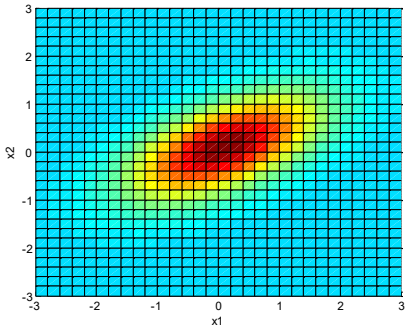
Η σ.π.π. κανονικής κατανομής για $\boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$





Η σ.π.π. κανονικής κατανομής για $\boldsymbol{C} = \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix}$





Η σ.π.π. κανονικής κατανομής για $\boldsymbol{C} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}$

Λόγω της εκθετικής μορφής της σ.π.π. της εκθετικής κατανομής μπορούμε να χρησιμοποιήσουμε για απλούστευση των υπολογισμών μία μονότονη λογαριθμική συνάρτηση,δηλαδή η πρόταση

το $\mathbf{x}$ ανήκει στη C_1 αν $p(\mathbf{x}|K_1)\cdot P(K_1) > p(\mathbf{x}|K_2)\cdot P(K_2)$, στη K_2 αν $p(\mathbf{x}|K_1)\cdot P(K_1) < p(\mathbf{x}|K_2)\cdot P(K_2)$ διατυπώνεται το $\mathbf{x}$ ανήκει

στη K_1 αν $g_1(\mathbf{x}) > g_2(\mathbf{x})$, στη K_2 αν $g_1(\mathbf{x}) < g_2(\mathbf{x})$

με

$$g_i(\boldsymbol{x}) = ln(p(\boldsymbol{x}|K_i)\cdot P(K_i)) = ln(p(\boldsymbol{x}|K_i)) + \ln(P(K_i)) =>$$

$$g_i(\boldsymbol{x}) = -\frac{(\boldsymbol{x}-\boldsymbol{\mu_i})^T \boldsymbol{C_i^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_i})}{2} + \ln\big(P(K_i)\big) + c_i =>$$

$$g_i(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{C_i^{-1}}\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu_i}^T(\boldsymbol{C_i^{-1}} + (\boldsymbol{C_i^{-1}})^T)\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_i}^T \boldsymbol{C_i^{-1}}\boldsymbol{\mu_i} + \ln\big(P(K_i)\big) + c_i$$

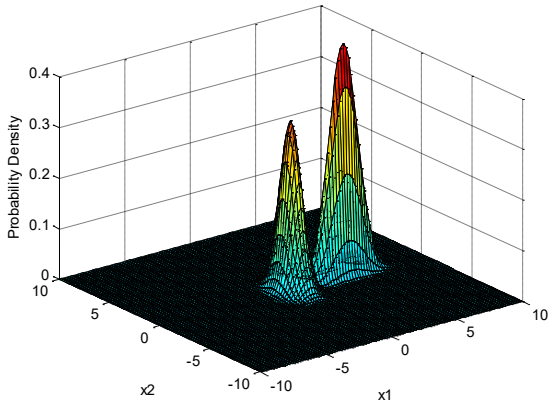
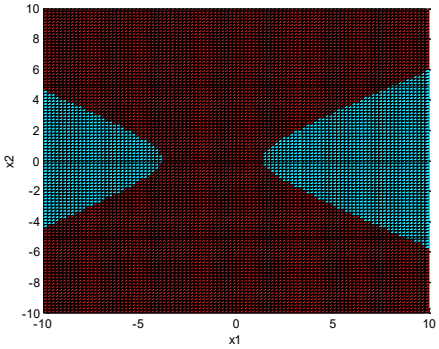
Και $c_i = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|\mathbf{C}_i|)$ σταθερά ποσότητα

Γενικά αυτή είναι μια μη γραμμική τετραγωνική μορφή. Τα σημεία του χώρου για τα οποία

$g_i(\boldsymbol{x}) = \boldsymbol{min}\{g_m(\boldsymbol{x})\}, \forall m \neq i$ καθορίζουν την περιοχή της κλάσης K_i. Στην περίπτωση των δύο κλάσεων η διαχωριστική τους καμπύλη ικανοποιεί την σχέση

$$d(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = 0$$

Για $g_1(\boldsymbol{x})$ και $g_2(\boldsymbol{x})$ συναρτήσεις απόφασης δύο κλάσεων Ακολουθούν παραδείγματα για διαφορετικές σ.π.π.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Οι σ.π.π. για $\boldsymbol{\mu}_1=[0, 0]^T$, $\boldsymbol{C}_1 = \begin{bmatrix} 3 & 0 \\ 0 & 0.75 \end{bmatrix}$ και $\boldsymbol{\mu}_2=[4,0]^T$, $\boldsymbol{C}_2 = \begin{bmatrix} 1.2 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$</td>
<td>Η διαχωριστική επιφάνεια των κλάσεων για ισοπίθανες κλάσεις</td>
</tr>
</table>

ΠΑΡΑΡΤΗΜΑ Α: ΒΑΣΙΚΕΣ ΠΡΑΞΕΙΣ ΓΡΑΜΜΙΚΗΣ ΑΛΓΕΒΡΑΣ

Ο ανάστροφος πίνακα $\mathbf{A}^{\mathrm{T}}$ ενός πίνακα **A** είναι ένας πίνακας που έχει ως γραμμές τις στήλες του **A** με ίδιο δείκτη. Αν ο $\mathbf{A}_{\mathrm{M\times N}}$ έχει Μ γραμμές και Ν στήλες, τότε ο $\mathbf{A}^{\mathrm{T}}{}_{\mathrm{N\times M}}$ θα έχει Ν γραμμές και Μ στήλες. Η ν γραμμή του $\mathbf{A}^{\mathrm{T}}$ είναι η ν στήλη του **A** και η μ γραμμή του η μ στήλη του **A**. Για παράδειγμα αν

$$\boldsymbol{A}_{3\times 2}=\begin{vmatrix}1 & 0.5\\ -1 & 3\\ 2 & -2\end{vmatrix}\Leftrightarrow \boldsymbol{A}^{T}_{2\times 3}=\begin{vmatrix}1 & -1 & 2\\ 0.5 & 3 & -2\end{vmatrix}.$$

Πολλαπλασιασμός πινάκων. Έστω οι πίνακες $\mathbf{A}_{\mathrm{N\times K}}$ και $\mathbf{B}_{\mathrm{K\times M}}$, το γινόμενό τους είναι ένας πίνακας $\mathbf{\Gamma}_{\mathrm{N\times M}}$ για τον οποίο γραφουμε $\mathbf{\Gamma}=\mathbf{A}\cdot\mathbf{B}$ και έχει στοιχεία $\Gamma_{\mu\kappa}$ που δίνονται από την σχέση

$$\Gamma_{\nu\mu}=\sum_{\kappa=1}^{K}A_{\nu\kappa}\cdot B_{\kappa\mu}$$

Για παράδειγμα αν

$$A_{3\times 2}=\begin{vmatrix}1 & 2\\ 0.5 & 4\\ 3 & 1\end{vmatrix}\ \text{και}\ B_{2\times 3}=\begin{vmatrix}3 & -1 & 1\\ 1 & 0.5 & -0.5\end{vmatrix}$$

$$A\cdot B=\Gamma_{3\times 3}=\begin{vmatrix}1\cdot 3+2\cdot 1 & 1\cdot(-1)+2\cdot 0.5 & 1\cdot 1+2\cdot(-0.5)\\ 0.5\cdot 3+4\cdot 1 & 0.5\cdot(-1)+4\cdot 0.5 & 0.5\cdot 1+4\cdot(-0.5)\\ 3\cdot 3+1\cdot 1 & 3\cdot(-1)+1\cdot 0.5 & 3\cdot 1+1\cdot(-0.5)=\end{vmatrix}=\begin{vmatrix}5 & 0 & 0\\ 7.5 & 1.5 & -1.5\\ 10 & -2.5 & 2.5\end{vmatrix}$$

ΠΑΡΑΡΤΗΜΑ Γ: ΣΤΑΤΙΣΤΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΟΥ ΧΩΡΟΥ ΤΩΝ ΠΡΟΤΥΠΩΝ.

Για ένα πλήθος Κ προτύπων που περιγράφονται σε ένα χώρο με Ν διαστάσεις η μέση τιμή μ είναι ένας πίνακας-στήλης που ορίζεται από τη σχέση

$\mu=\frac{1}{K}\sum_{\kappa=1}^{K}x_k=[\mu_1,\mu_2,\ldots,\mu_\nu\ldots\mu_N]^T$ όπου $\mu_\nu=\frac{1}{K}\sum_{\kappa=1}^{K}x_{\nu\kappa}$ (2.3.1)

Η μέση τιμή μ αναφέρεται και ως η μαθηματική προσδοκία Ε(x) όπου x μεταβλητή για τους πίνακες-στήλης των προτύπων. Αν υπάρχουν λ πρότυπα που οι τιμές των πινάκων-στήλης τους είναι ίσες, λαμβάνονται υπόψη λ φορές στο άθροισμα της (9) και ως εκ τούτου δεν χρησιμοποιείται στον τύπο (9) η συνάρτηση συχνότητας εμφάνισης f(x) κάθε τιμής της μεταβλητής x. Οι τιμές μν όπου ν=1…Ν ορίζουν το μέσο διάνυσμα $\vec{\mu}$ των Κ προτύπων στο Ν-διάστατο χώρο τους. Ο πίνακας συμμεταβλητότητας ή πίνακας συνδιασποράς Cov ορίζεται από τη σχέση

$$C_{ov} = \frac{1}{K}\sum_{\kappa=1}^{K}(x_k-\mu)\cdot(x_k-\mu)^T = \frac{1}{K}\sum_{\kappa=1}^{K}\begin{vmatrix} x_{1\kappa}-\mu_1 \\ x_{2\kappa}-\mu_2 \\ \vdots \\ x_{\nu\kappa}-\mu_\nu \\ \vdots \\ x_{\kappa N}-\mu_N \end{vmatrix}\cdot[x_{1\kappa}-\mu_1, x_{2\kappa}-\mu_2, \ldots, x_{\nu\kappa}-\mu_\kappa, \ldots x_{N\kappa}-\mu_N] =$$

$$\frac{1}{K}\sum_{\kappa=1}^{K}\begin{vmatrix} (x_{1\kappa}-\mu_1)^2 & (x_{1\kappa}-\mu_1)\cdot(x_{2\kappa}-\mu_2) & \cdots & (x_{1\kappa}-\mu_1)\cdot(x_{\nu\kappa}-\mu_\nu) & \cdots & (x_{1\kappa}-\mu_1)\cdot(x_{N\kappa}-\mu_N) \\ (x_{2\kappa}-\mu_2)\cdot(x_{1\kappa}-\mu_1) & (x_{2\kappa}-\mu_2)^2 & \cdots & (x_{2\kappa}-\mu_2)\cdot(x_{\nu\kappa}-\mu_\nu) & \cdots & (x_{2\kappa}-\mu_2)\cdot(x_{N\kappa}-\mu_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ (x_{\lambda\kappa}-\mu_\lambda)\cdot(x_{1\kappa}-\mu_1) & (x_{\lambda\kappa}-\mu_\lambda)\cdot(x_{2\kappa}-\mu_2) & \cdots & (x_{\lambda\kappa}-\mu_\lambda)\cdot(x_{\nu\kappa}-\mu_\nu) & \cdots & (x_{\lambda\kappa}-\mu_\lambda)\cdot(x_{N\kappa}-\mu_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ (x_{N\kappa}-\mu_N)\cdot(x_{1\kappa}-\mu_1) & (x_{N\kappa}-\mu_N)\cdot(x_{2\kappa}-\mu_2) & \cdots & (x_{N\kappa}-\mu_N)\cdot(x_{\nu\kappa}-\mu_\nu) & \cdots & (x_{N\kappa}-\mu_N)^2 \end{vmatrix}$$

$$= \begin{vmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1\nu} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2\nu} & \cdots & \sigma_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{\lambda 1} & \sigma_{\lambda 2} & \cdots & \sigma_{\lambda\nu} & \cdots & \sigma_{\lambda N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{N\nu} & \cdots & \sigma_N^2 \end{vmatrix} = E(\ (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\ ) \quad (2.3.2)$$

$$\text{όπου } \sigma_{\lambda\nu} = \frac{1}{K}\sum_{\kappa=1}^{K}(x_{\lambda\kappa}-\mu_\lambda)\cdot(x_{\nu\kappa}-\mu_\nu)$$

Η ποσότητα σν2 είναι η μεταβλητότητα ή διασπορά (variance) των μετρήσεων του ν-οστού χαρακτηριστικού των προτύπων. Η ποσότητα $\sigma_{\lambda\nu}=\sigma_{\nu\lambda}$ λέγεται συμμεταβλητότητα ή συνδιασπορά των τιμών των λ και ν χαρακτηριστικών των προτύπων και χρησιμοποιείται για τον υπολογισμό του συντελεστή συσχέτισης Ρλν δύο χαρακτηριστικών μ,ν σύμφωνα με τη σχέση

$$P_{\lambda\nu} = \frac{\sigma_{\lambda\nu}}{\sigma_\lambda\cdot\sigma_\nu} \qquad (2.3.3)$$

Η ποσότητα $\sigma_\nu = \sqrt{\sigma_\nu{}^2}$ λέγεται τυπική απόκλιση. Οι παραπάνω ποσότητες χρησιμοποιούνται για την αξιολόγηση και ανάλυση των μετρούμενων χαρακτηριστικών.

ΠΑΡΑΡΤΗΜΑ Γ:ΤΑ ΙΔΙΟΔΙΑΝΥΣΜΑΤΑ ΤΟΥ ΠΙΝΑΚΑ ΣΥΝΔΙΑΣΠΟΡΑΣ

Σε ένα συμμετρικό πίνακα $\boldsymbol{C_{NXN}}$, $\boldsymbol{C^T = C}$, με ιδιοτιμές $\lambda_\mu, \lambda_\nu, \lambda_\mu \neq \lambda_\nu$ και αντίστοιχα μοναδιαία ιδιοδιανύσματα $\boldsymbol{w_\mu \neq w_\nu}$ ισχύει ότι $\boldsymbol{w_\mu^T \cdot w_\nu = 0}$. Δηλαδή τα ιδιοδιανύσματα είναι κάθετα μεταξύ τους.

Απόδειξη: Επειδή λ_μ, λ_ν ιδιοτιμές και $\boldsymbol{w_\mu, w_\nu}$ ιδιοδιανύσματα ισχύουν οι σχέσεις

$$\left.\begin{aligned} \boldsymbol{C\cdot w_\mu} &= \lambda_\mu \cdot \boldsymbol{w_\mu} \\ \boldsymbol{C\cdot w_\nu} &= \lambda_\nu \cdot \boldsymbol{w_\nu} \end{aligned}\right\} \Rightarrow \begin{aligned} \boldsymbol{w_\nu^T\cdot C\cdot w_\mu} &= \boldsymbol{w_\nu^T}\cdot\lambda_\mu\cdot\boldsymbol{w_\mu} = \lambda_\mu\cdot\boldsymbol{w_\nu^T\cdot w_\mu}\ (A) \\ \boldsymbol{w_\mu^T\cdot C\cdot w_\nu} &= \boldsymbol{w_\mu^T}\cdot\lambda_\nu\cdot\boldsymbol{w_\nu} = \lambda_\nu\cdot\boldsymbol{w_\mu^T\cdot w_\nu}\ (B) \end{aligned}$$

$$\boldsymbol{w_\nu^T\cdot C\cdot w_\mu} = \left(\boldsymbol{w_\mu^T\cdot C\cdot w_\nu}\right)^T \quad (\Gamma)$$

Υπενθυμίζεται ότι $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T$.

Από τις σχέσεις (Α), (Β), (Γ) συνεπάγεται

$$\lambda_\mu \cdot \boldsymbol{w}_\nu^T \cdot \boldsymbol{w}_\mu = \lambda_\nu \cdot \boldsymbol{w}_\nu^T \cdot \boldsymbol{w}_\mu \Leftrightarrow (\lambda_\mu - \lambda_\nu)\boldsymbol{w}_\nu^T \cdot \boldsymbol{w}_\mu = 0$$

επειδή $\lambda_\mu \neq \lambda_\nu$ έπεται ότι $\boldsymbol{w}_\nu^T \cdot \boldsymbol{w}_\mu = \boldsymbol{w}_\mu^T \cdot \boldsymbol{w}_\nu = 0$.

Από τα προηγούμενα συνεπάγεται ότι για τον πίνακα $\boldsymbol{W} = [\boldsymbol{w}_1, \ \boldsymbol{w}_2, \ \cdots, \ \boldsymbol{w}_N]$ ισχύει ότι

$$\boldsymbol{W}^T \cdot \boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_N^T \end{bmatrix} [\boldsymbol{w}_1, \ \boldsymbol{w}_2, \ \cdots, \ \boldsymbol{w}_N] = \begin{bmatrix} \boldsymbol{w}_1^T \cdot \boldsymbol{w}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{w}_2^T \cdot \boldsymbol{w}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{w}_N^T \cdot \boldsymbol{w}_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ & 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I}$$

Άρα αντίστροφος του $\boldsymbol{W}$ είναι ο $\boldsymbol{W}^{-1} = \boldsymbol{W}^T$.

Για τα ιδιοδιανύσματα του $\boldsymbol{C}$ ισχύει εξ' ορισμού ότι

$$\left.\begin{array}{r} \boldsymbol{C} \cdot \boldsymbol{w}_1 = \lambda_1 \cdot \boldsymbol{w}_1 \\ \boldsymbol{C} \cdot \boldsymbol{w}_2 = \lambda_2 \cdot \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{C} \cdot \boldsymbol{w}_N = \lambda_N \cdot \boldsymbol{w}_N \end{array}\right\} \Rightarrow \boldsymbol{C} \cdot [\boldsymbol{w}_1, \ \cdots, \ \boldsymbol{w}_N] = [\boldsymbol{w}_1, \ \cdots, \ \boldsymbol{w}_N] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} = \boldsymbol{W} \cdot \boldsymbol{\Lambda}$$

όπου $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$

Αφού $\boldsymbol{C} \cdot \boldsymbol{W} = \boldsymbol{W} \cdot \boldsymbol{\Lambda} \Rightarrow \boldsymbol{W}^{-1} \cdot \boldsymbol{C} \cdot \boldsymbol{W} = \boldsymbol{W}^{-1} \cdot \boldsymbol{W} \cdot \boldsymbol{\Lambda} = \boldsymbol{I} \cdot \boldsymbol{\Lambda} = \boldsymbol{\Lambda} \Leftrightarrow \boldsymbol{W}^T \cdot \boldsymbol{C} \cdot \boldsymbol{W} = \boldsymbol{\Lambda}$.

ΠΑΡΑΡΤΗΜΑ Δ: ΕΝΤΟΛΕΣ MATLAB & TOOLBOX

Mprabo bravo