Γραμμικοί Ταξινομητές - Ο αλγόριθμος Perceptron

ECE-TEL830 ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Αθανάσιος Κούτρας

Αναπληρωτής Καθηγητής

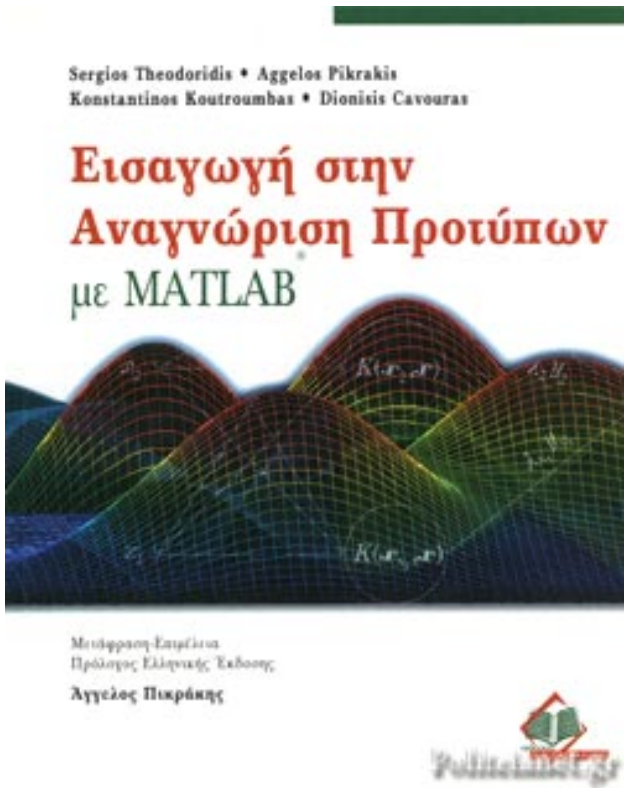
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Παν. Πελοποννήσου

2 Απριλίου 2023

Περίγραμμα διάλεξης

Υλικό μελέτης

Theodoridis S., Pikrakis, A., Koutroumbas K., Cavouras, D., **"Εισαγωγη στην αναγνώριση προτύπων με MATLAB",ΚΕΦΑΛΑΙΟ** 2



Εισαγωγή

- Στη διάλεξη αυτή θα εστιάσουμε στον κατευθείαν σχεδιασμό μιας συνάρτησης διάκρισης / επιφάνειας απόφασης η οποία θα διαχωρίζει τις κλάσεις σύμφωνα με κάποιο κριτήριο.
- στην προηγούμενη διάλεξη χρησιμοποιήσαμε τεχνικές που βασίζονται στον βέλτιστο Bayesian ταξινομητή που στηρίζεται στην εκτίμηση συναρτήσεων πυκνότητας πιθανότητας που περιγράφει την κατανομή των δεδομένων σε κάθε κλάση
- η εκτίμηση αυτή είναι τις περισσότερες φορές δύσκολη, ειδικά όταν πρόκειται για μεγάλης διάστασης χώρους.
- η προσέγγιση που ακολουθούμε είναι ο σχεδιασμός μιας επιφάνειας απόφασης που διαχωρίζει τις κλάσεις απευθείας στο σύνολο δεδομένων εκπαίδευσης, χωρίς να χρειάζεται η εκτίμηση συναρτήσεων πυκνότητας πιθανότητας.
- στην περίπτωση αυτή το πρόβλημα γίνεται ευκολότερο, αν και δεν οδηγεί στον βέλτιστο ταξινομητή.
- επειδή στην πράξη όμως το σύνολο των παραδειγμάτων που έχουμε είναι περιορισμένο, η απόδοση αυτών των ταξινομητών είναι καλύτερη από αυτή του Bayesian

Ταξινόμηση με χρήση συνάρτησης διάκρισης

- για αρχή, σχεδιάζουμε έναν γραμμικό ταξινομητή ο οποίος περιγράφεται από την εξίσωση:

$$w^T x + w_0 = 0$$

η οποία μπορεί να γραφτεί και ως:

$$w'^T x' \equiv [w^T, w_0] \begin{bmatrix} x \\ 1 \end{bmatrix} = 0$$

προκειμένου να αλλάξουμε τον χώρο που εργαζόμαστε και να απλοποιηθεί ο συμβολισμός.

- το w ονομάζεται διάνυσμα βαρών (weight vector) και το w_0 ως το κατώφλι (threshold).
- για την περίπτωση που έχουμε δύο κλάσεις, ω_1, ω_2 υπολογίζουμε την

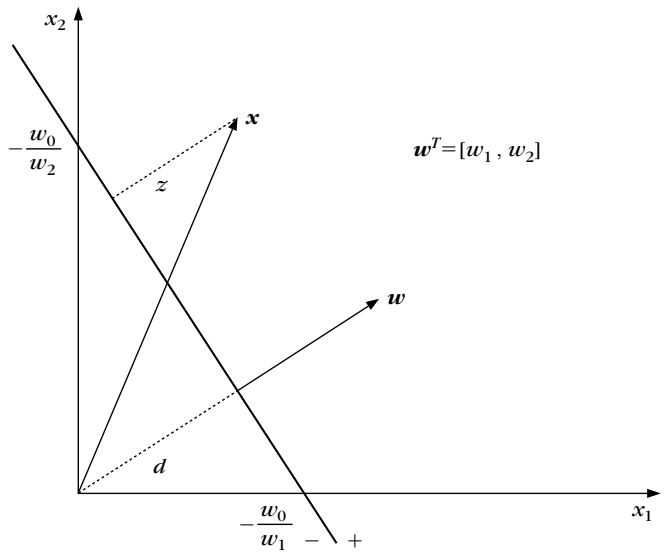
$$g(x) = w^T x + w_0 = 0 = w_1 x_1 + w_2 x_2 + \cdots + w_l x_l + w_0$$

- ας θεωρήσουμε δύο σημεία τα οποία βρίσκονται πάνω στο υπερεπίπεδο απόφασης. Για αυτά ισχύει:

$$0 = w^T x_1 + w_0 = w^T x_2 + w_0 \Rightarrow$$

$$w^T(x_1 - x_2) = 0, \forall x_1, x_2$$

- για να ισχύει η παραπάνω ισότητα από τη στιγμή που το διάνυσμα διαφοράς $x_1 - x_2$ είναι πάνω στο υπερεπίπεδο, θα πρέπει το w να είναι ορθογώνιο σε αυτό (εσωτερικό γινόμενο ίσο με μηδέν)



Σχήμα: Γεωμετρία της γραμμής απόφασης. Από τη μία ισχύει $g(x) > 0$ από την άλλη $g(x) < 0$

$$d = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2}}, z = \frac{|g(\mathbf{x})|}{\sqrt{w_1^2 + w_2^2}}$$

Ο αλγόριθμος Perceptron

- είναι κατάλληλος στην περίπτωση προβλήματος δύο κλάσεων που είναι γραμμικώς διαχωρίσιμες.
- ο αλγόριθμος αυτός υπολογίζει τις τιμές των βαρών w ενός γραμμικού ταξινομητή που διαχωρίζει τις δύο κλάσεις.
- ο αλγόριθμος αυτός είναι επαναληπτικός. Ξεκινάει με μια εκτίμηση και συγκλίνει σε μια λύση μετά από έναν πεπερασμένο αριθμό βημάτων επανάληψης.
- η λύση που προκύπτει ταξινομεί επιτυχώς όλα τα σημεία του συνόλου εκπαίδευσης
- θα προσεγγίσουμε το πρόβλημα ως πρόβλημα βελτιστοποίησης με:
  - κατάλληλη συνάρτηση κόστους
  - ένα αλγοριθμικό σχήμα για να τη βελτιστοποιήσει

Η συνάρτηση κόστους του perceptron

- η συνάρτηση κόστους που επιλέγεται είναι η:

$$J(\mathbf{w}) = \sum_{x \in Y} (\delta_x \mathbf{w}^T \mathbf{x})$$

με Y το υποσύνολο διανυσμάτων εκπαίδευσης τα οποία ταξινομήθηκαν εσφαλμένα από το υπερεπίπεδο που ορίζεται από τα $\mathbf{w}$
δ_x παίρνει τιμές ± 1 ανάλογα στην κατηγορία που ανήκει το παράδειγμα

- το άθροισμα είναι πάντοτε θετικό ή ίσο με μηδέν και η συνάρτηση κόστους είναι κατά τμήματα γραμμική.
- για την ελαχιστοποίηση της χρησιμοποιούμε ένα επαναληπτικό σχήμα της μεθόδου καθόδου προς την κατεύθυνση της παραγώγου (gradient descent)

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \rho_t \frac{\partial \mathbf{J}(\mathbf{w})}{\partial \mathbf{w}}$$

Ο αλγόριθμος Perceptron - Εκπαίδευση

- ο αλγόριθμος αυτός συγκλίνει σε μία από άπειρες δυνατές λύσεις
- ξεκινώντας από διαφορετικές αρχικές συνθήκες, προκύπτουν διαφορετικά υπερ-επίπεδα.
- η ενημέρωση στο $t-$οστό βήμα επανάληψης δίνεται από:

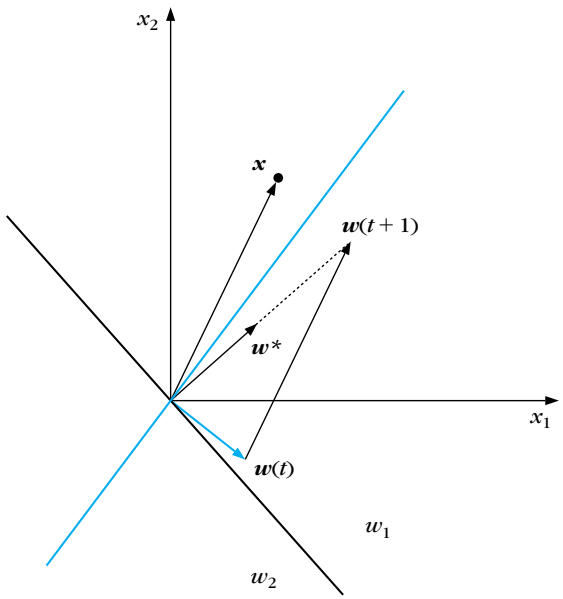
$$\mathbf{w}(t+1) = \mathbf{w}(t) - \rho_t \sum_{x \in Y} \delta_x \mathbf{x}$$

όπου $\mathbf{w}$ είναι το επαυξημένο με w_0 διάνυσμα, $\mathbf{Y}$ το σύνολο των διανυσμάτων για τα οποία έχουμε λάνθασμένη ταξινόμηση με βάση την τρέχουσα εκτίμηση $\mathbf{w}(t)$, δ_x ισούται με -1 αν $x \in \omega_1$ και $+1$ αν $x \in \omega_2$ και ρ_t είναι μια παράμετρος που ρυθμίζεται από τον χρήστη και ελέγχει την ταχύτητα σύγκλισης

- η παράμετρος αυτή μπορεί να τύχει περιορισμών από τον χρήστη για την καλύτερη απόδοση, όπως για παράδειγμα να είναι μια σταθερά.
- η σύγκλιση του αλγορίθμου επιτυγχάνεται όταν το σύνολο Y γίνει ίσο με το κενό.

Γεωμετρική ερμηνεία του αλγορίθμου perceptron

- Στο παράδειγμα υποθέτουμε ένα λανθασμένα ταξινομημένο παράδειγμα, το **x**. Ο αλγόριθμος perceptron διορθώνει το διάνυσμα βαρών προς την κατεύθυνση του **x**.
- αυτό γίνεται σε περισσότερα από ένα επαναληπτικά βήματα που εξαρτώνται από τις τιμές της παραμέτρου ρ_t



Σχήμα: Γεωμετρική ερμηνεία του αλγόριθμου perceptron

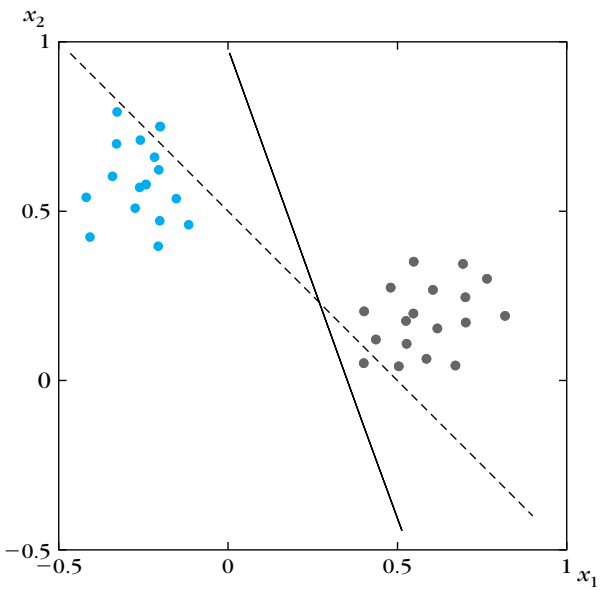
Παράδειγμα

- έστω η διακεκομένη γραμμή

$$x_1 + x_2 - 0.5$$

που αντιστοιχεί στο διάνυσμα βαρών $[1, 1, -0.5]^T$. Αυτό έχει υπολογιστεί κατά το τελευταίο βήμα επανάληψης του αλγορίθμου perceptron με $\rho_t = \rho = 0.7$

- Παρατηρούμε ότι όλα εκτός από τα $[0.4, 0.05]^T$ και $[-0.20, 0.75]^T$ ταξινομούνται σωστά.
- Να υπολογιστεί το επόμενο διάνυσμα βαρών σύμφωνα με τον αλγόριθμο perceptron. Τι αποτέλεσμα θα προκύψει ως προς την ταξινόμηση;

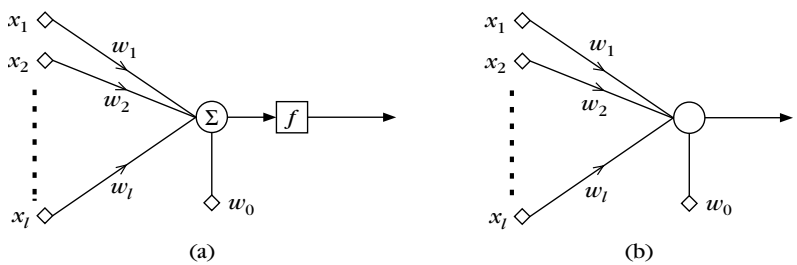


Το μοντέλο Perceptron

- μετά τον υπολογισμό του ταξινομητή (δηλαδή του διανύσματος βαρών $\mathbf{w}$ και του κατωφλίου w_0, ένα άγνωστο σημείο x ταξινομείται σε μια από τις δύο κλάσεις ανάλογα με το αποτέλεσμα της:

$$f(w^T x) = f(w_1 x(1) + w_2 x(2) + \cdots + w_l x(l) + x_0)$$

- η παραπάνω συνάρτηση $f(\cdot)$ στην απλούστερη μορφή της είναι η βηματική ή η συνάρτηση sign.
- ονομάζεται **συνάρτηση ενεργοποίησης** (activation function)
- το δίκτυο αυτό ονομάζεται perceptron ή νευρώνας. Είναι απλό παράδειγμα μηχανών μάθησης και χρησιμοποιείται για να χτίσουμε πιο πολύπλοκα δίκτυα μάθησης.



Σχήμα: Το βασικό μοντέλο perceptron (α) ένας γραμμικός συνδυαστής ακολουθούμενος από συνάρτηση ενεργοποίησης (β) συνδυαστής και συνάρτηση ενεργοποίησης συγχωνεύονται

Μορφές του αλγόριθμου perceptron

- η κλασική μορφή του αλγόριθμου που παρουσιάστηκε ονομάζεται μορφή δέσμης (batch form).
- σε κάθε επανάληψη λαμβάνονται υπόψιν όλα τα σημεία και η ενημέρωση γίνεται αφού επεξεργαστούν όλα τα σημεία δεδομένων.
- μια άλλη προσέγγιση είναι η online έκδοση.
- σε αυτή, τα δεδοομένα εξετάζονται ακολουθιακά, και η ενημέρωση πραγματοποιείται μετά από την εξέταση κάθε ενός στοιχείου.
- το σύνολο N διαδοχικών επαναλήψεων μέσω των οποίων εξετάστηκαν μια φορά ολα τα δεδομένα, ονομάζεται εποχή (epoch).
- η ενημέρωση πραγματοποιείται ως εξής:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \rho y_{(t)} \mathbf{x}_{(t)} \text{εαν } y_{(t)}(w^T(t)x_{(t)}) \leq 0$$

$$\mathbf{w}(t+1) = \mathbf{w}(t), \text{αλλιώς}$$

Εκτίμηση μέσου τετραγωνικού σφάλματος

- Αν οι κλάσεις είναι γραμμικά διαχωρίσιμες, τότε η έξοδος του perceptron θα είναι ίση με ± 1 ανάλογα με την κατηγορία.
- εαν οι κλάσεις όμως δεν είναι γραμμικά διαχωρίσιμες, τότε μπορούμε να υπολογίσουμε τα βάρη $w_1, w_2, \cdots, wo$ έτσι ώστε η διαφορά μεταξύ:
  - της πραγματικής εξόδου του ταξινομητή $\mathbf{w}^T\mathbf{x}$
  - της επιθυμητής εξόδου (για παράδειγμα ± 1) για τις δύο κατηγορίες αντίστοιχα

  να είναι ελάχιστη ακολουθώντας την έννοια του ελάχιστου μέσου τετραγωνικού σφάλματος.

Ελαχιστοποίηση μέσου τετραγωνικού σφάλματος

- η ελαχιστοποίηση του μέσου τετραγωγνικού σφάλματος πραγματοποιείται επιλέγοντας το $\mathbf{w}$ ώστε η συνάρτηση κόστους

$$J(w) = E[(y - w^T x)^2]$$

να γίνεται ελάχιστη

- η λύση που προκύπτει υπολογίζεται από την

$$\hat{w} = \underset{w}{\operatorname{argmin}} J(w)$$

η οποία μετά από υπολογισμούς καταλήγει στην:

$$\hat{\mathbf{w}} = R_x^{-1} E[\mathbf{x}y]$$

με R_x^{-1} τον γνωστό πίνακα συσχέτισης ή αυτοσυσχέτισης και είναι ίσος με τον πίνακα συνδιασποράς αν οι αντίστοιχες μέσες τιμές είναι μηδενικές και το $E[\mathbf{x}y]$ είναι γνωστό σαν ετεροσυσχέτιση ανάμεσα στην επιθυμητή έξοδο και τα διανύσματα χαρακτηριστικών (εισόδου)

- το βέλτιστο διάνυσμα βαρών προκύπτει ως λύση ενός συνόλου γραμμικών εξισώσεων, με την προϋπόθεση ότι ο πίνακας συσχέτισης είναι αντιστρέψιμος.

Γενίκευση στην περίπτωση πολλών κλάσεων

- Στην περίπτωση αυτή το πρόβλημα ανάγεται στον σχεδιασμό M γραμμικών συναρτήσεων διάκρισης
$$g_i(x) = w_i^T x$$
σύμφωνα με το κριτήριο του MSE
- οι επιθυμητές αποκρίσεις y_i επιλέγονται ώστε να είναι 1 για τα x που ανήκουν στην κατηγορία ω_i και 0 διαφορετικά.
- ορίζουμε το διάνυσμα y και τον πίνακα W ως ακολούθως:
$$y = [y_1, y_2, \cdots, y_M]^T$$
$$W =]w_1, w_2, \cdots, w_M]$$
- ο σκοπός είναι να εκτιμήσουμε τον πίνακα W ώστε
$$\hat{W} = \underset{W}{\operatorname{argmin}} E[||y - W^T x||^2] = \underset{W}{\operatorname{argmin}} E\left[\sum_{i=1}^{M} (y_i - w_i^T x)^2\right]$$
- το παραπάνω είναι ισοδύναμο με M προβλήματα MSE ελαχιστοποίησης. Δηλαδή υπολογίζει κάθε w_i έτσι ώστε η ζητούμενη έξοδος του να είναι 1 για παραδείγματα που ανήκουν στην κατηγορία ω_i και 0 για όλες τις υπόλοιπες κατηγορίες.

Ταξινομητής ελάχιστου τετραγωνικού σφάλματος

- σε αυτή την περίπτωση θέλουμε να εκτιμήσουμε το διάνυσμα παραμέτρων w ενός γραμμικού ταξινομητή.
- η διαφορά από τις προηγούμενες περιπτώσεις είναι ότι δεν απαιτείται η υπόθεση της γραμμικής διαχωρισιμότητας
- η μέθοδος που προκύπτει είναι γνωστή και ως μέθοδος των ελαχίστων τετραγώνων και προχωρά στην εκτίμηση του καλύτερου γραμμικού ταξινομητή
- ο όρος καλύτερος σημαίνει το w που ελαχιστοποιεί το κόστος

$$J(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

όπου y_i είναι οι ετικέτες κλάσεις των $x_i, i = 1, 2, \cdots, N$ και N είναι ο αριθμός των σημείων εκπαίδευσης.

LS ταξινομητής

- η LS εκτίμηση δίνεται από τη σχέση

$$\hat{w} = (X^T X)^{-1} X^T y$$

- στην παραπάνω σχέση, το μητρώο $(X^T X)^{-1} X^T$ ονομάζεται και ψευδοαντίστροφος του X και συμβολίζεται με X^+
- η μέθοδος αυτή έχει το πλεονέκτημα ότι παρουσιάζει μοναδική λύση (αντιστοιχεί στο μοναδικό ελάχιστο της $J(w)$
- η λύση αυτή λαμβάνεται λύνοντας ένα γραμμικό σύστημα εξισώσεων και αντιστοιχεί στο ελάχιστο άθροισμα των τετραγώνων του σφάλματος.

Προβλήματα του αλγόριθμου LS

- η αντιστροφή του προηγούμενου πίνακα μπορεί να οδηγήσει σε αριθμητικά προβλήματα ειδικότερα σε χώρους μεγάλης διάστασης
- έτσι σε πολλές περιπτώσεις μπορεί να προκύψει ιδιάζον (singular) πίνακας.
- για να ξεπεράσουμε το πρόβλημα μπορούμε να προσθέσουμε μια μικρή θετική σταθερά στην κύρια διαγώνιο και να λύσουμε το σύστημα:

$$\hat{w} = (X^T X + CI)^{-1} X^T y$$

- η παραπάνω εξίσωση είναι ο ελαχιστοποιητής της κανονικής (regularised) έκδοσης του κόστους της

$$J(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + C w^T w$$

Ο LS ταξινομητής για την περίπτωση πολλών κλάσεων

- Θεωρούμε ότι έχουμε στη διάθεση μας N σημεία εκπαίδευσης τα οποία ανήκουν σε $c > 2$ κλάσεις.
- το πρόβλημα που πρέπει να αντιμετωπίσουμε είναι να σχεδιάσουμε έναν ταξινομητή που αποτελείται από c γραμμικές συναρτήσεις διάκρισης (μία για κάθε κλάση):

$$g_j(x) \equiv w_j^T x + w_{j_0}, j = 1, 2, \cdots, c$$

- ο σχεδιασμός θα βασιστεί στο LS κριτήριο και ο κανόνας θα γίνει: Δοθέντος x, ταξινόμησε το στην κλάση ω_i αν:

$$g_i(x) > g_j(x), \forall j \neq i$$

LS Πολλών κλάσεων

- οι γραμμικές συναρτήσεις σχεδιάζονται ως εξής: Για κάθε x_i, ορίζεται το $c-$διάστατο διάνυσμα ετικετών κλάσης

$$y_i = [y_{i1}, y_{i2}, \cdots, y_{ic}], i = 1, 2, \cdots, N$$

- το $j-$οστό στοιχειο y_{ij} είναι 1 αν $x_i \in \omega_j$ και 0 αλλιώς.
- η εκτίμηση του w_j γίνεται έτσι ώστε να ελαχιστοποιείται το κόστος:

$$\sum_{i=1}^{N}(y_{ij} - w_j^T x_i - w_{j0})^2, j = 1, 2, \cdots, c$$

- δηλαδή πρέπει να λυθούν c προβλήματα, ένα για κάθε κλάση.
- κάθε υπερεπίπεδο w_j που προκύπτει, είναι το αποτέλεσμα εκπαίδευσης κατά τρόπο ώστε στην ιδανική περίπτωση όλα τα σημεία της κλάσης ω_j να βρίσκονται στη μια πλευρά και όλα τα υπόλοιπα σημεία να είναι στην άλλη πλευρά.

Signal & Image Processing, Pattern Recognition Group (SIPPRE)
www.sippre-group.com