Data Mining
Classification: Alternative Techniques

Lecture Notes for Chapter 5

(cont.)

Classification Problem

- Πρόβλημα μάθησης με επίβλεψη (**Supervised learning**)
- Δεδομένα του συνόλου εκπαίδευσης $X = \{(x_n, y_n)\}_{n=1}^{N}$ αποτελούμενα από ζεύγη
  - σημείων σε ένα χώρο ($x_n \in R^d$) και
  - ετικέτας ή πρόβλεψης y_n .
- **Στόχος** να φτιάξουμε μία συνάρτηση ***f(x)*** *:*

$$\forall \; x_n \in X \quad f(x_n) = y_n$$

Μηχανές Διανυσματικής Στήριξης (Support Vector Machines - SVMs)

- Πρόβλημα μάθησης με επίβλεψη (**Supervised learning**)

- Δεδομένα του συνόλου εκπαίδευσης $X = \{(x_n, y_n)\}_{n=1}^{N}$ αποτελούμενα από ζεύγη σημείων σε ένα χώρο ($x_n \in R$) και ετικέτας ή πρόβλεψης y_n .

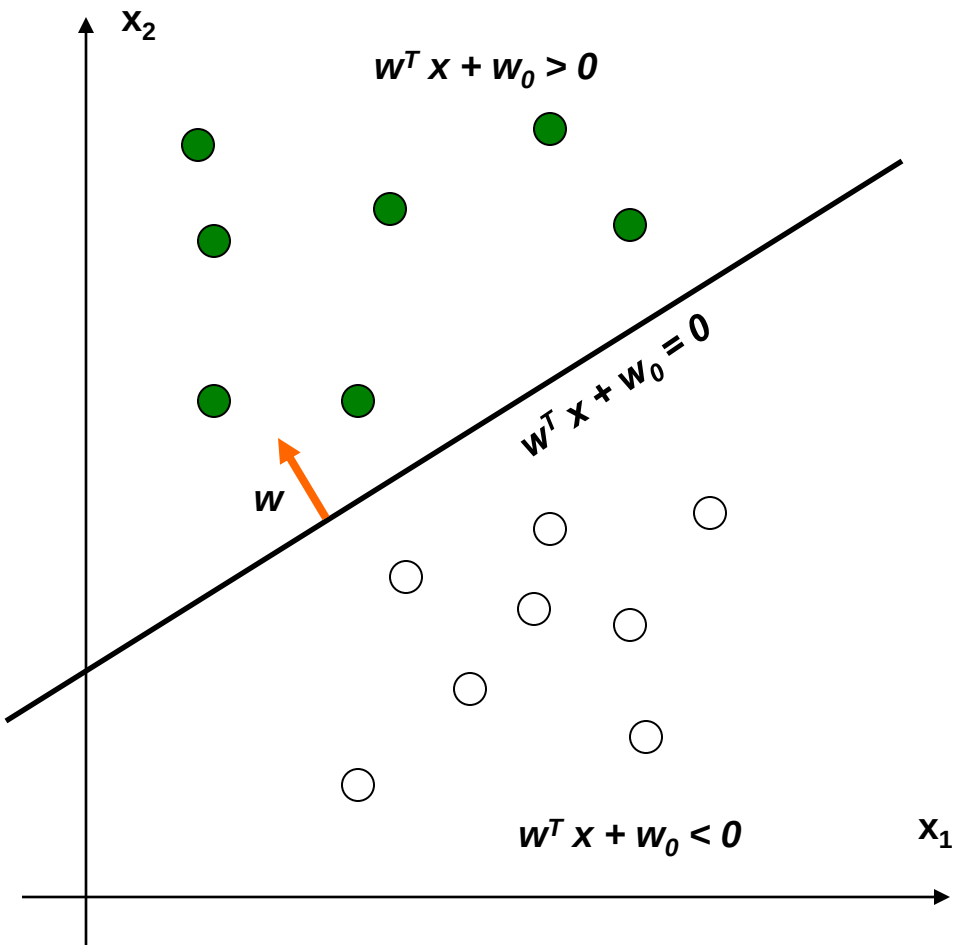
- Στόχος να φτιάξουμε μία συνάρτηση *f(x)* που να προβλέπει την κατηγορία y των σημείων.

Γραμμική ταξινόμηση

- *f(x)* γραμμική συνάρτηση:

$$f(x) = w^T x + w_0$$

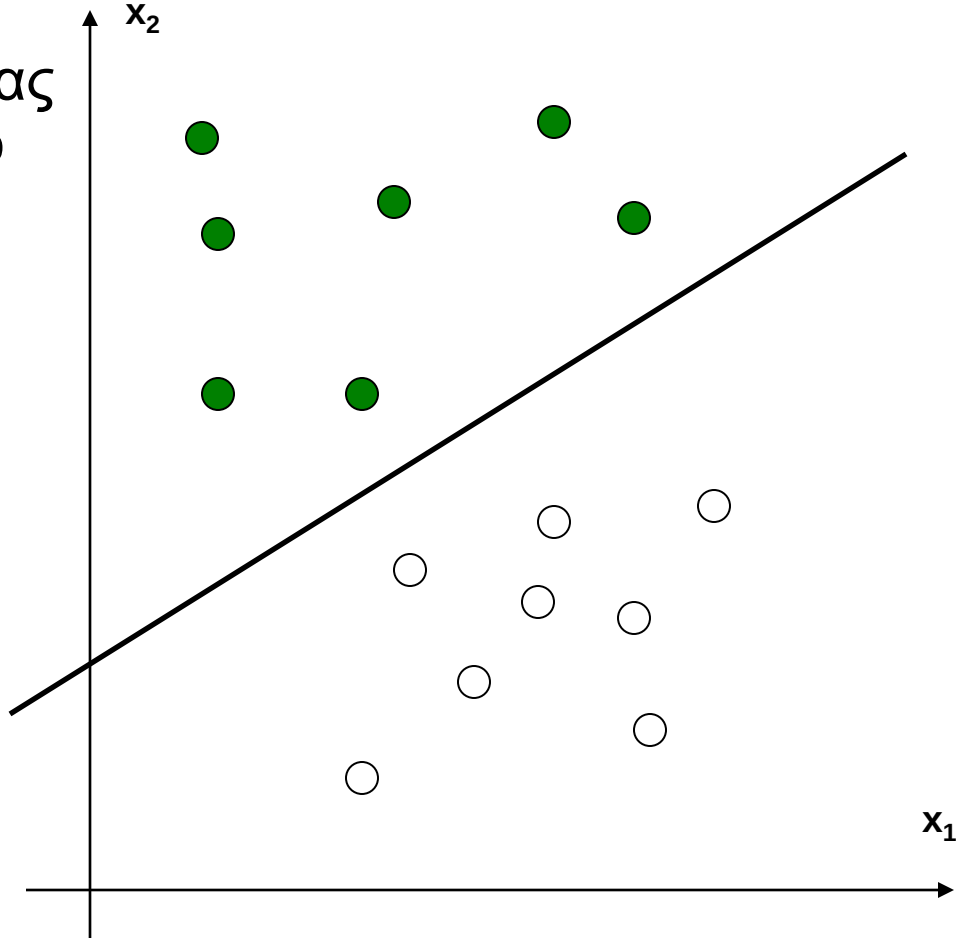
- **Ορίζει ένα υπερεπίπεδο ως διαχωριστική επιφάνεια στον χώρο των προτύπων**



Linear Discriminant Function

- Ποια είναι η βέλτιστη θέση της διαχωριστικής επιφάνειας ώστε να ελαχιστοποιηθεί το σφάλμα διάκρισης των 2 κατηγοριών;

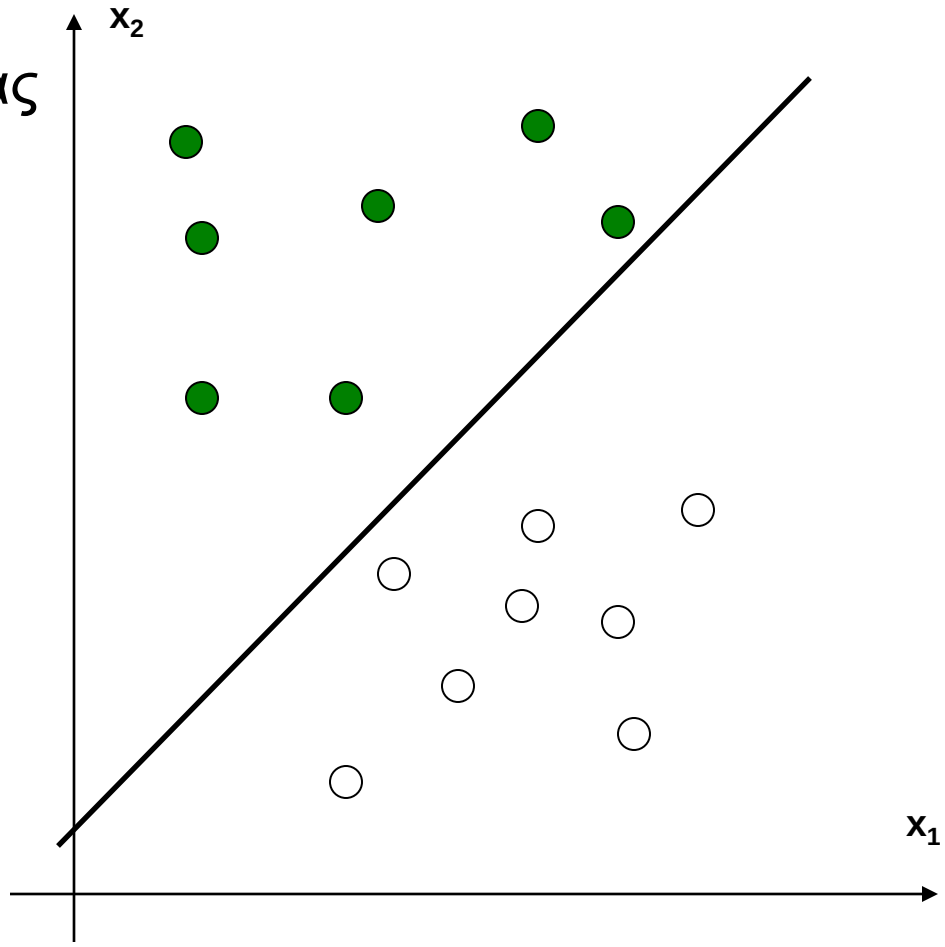
- **Άπειρος αριθμός λύσεων!**



Linear Discriminant Function

- Ποια είναι η βέλτιστη θέση της διαχωριστικής επιφάνειας ώστε να ελαχιστοποιηθεί το σφάλμα διάκρισης των 2 κατηγοριών;

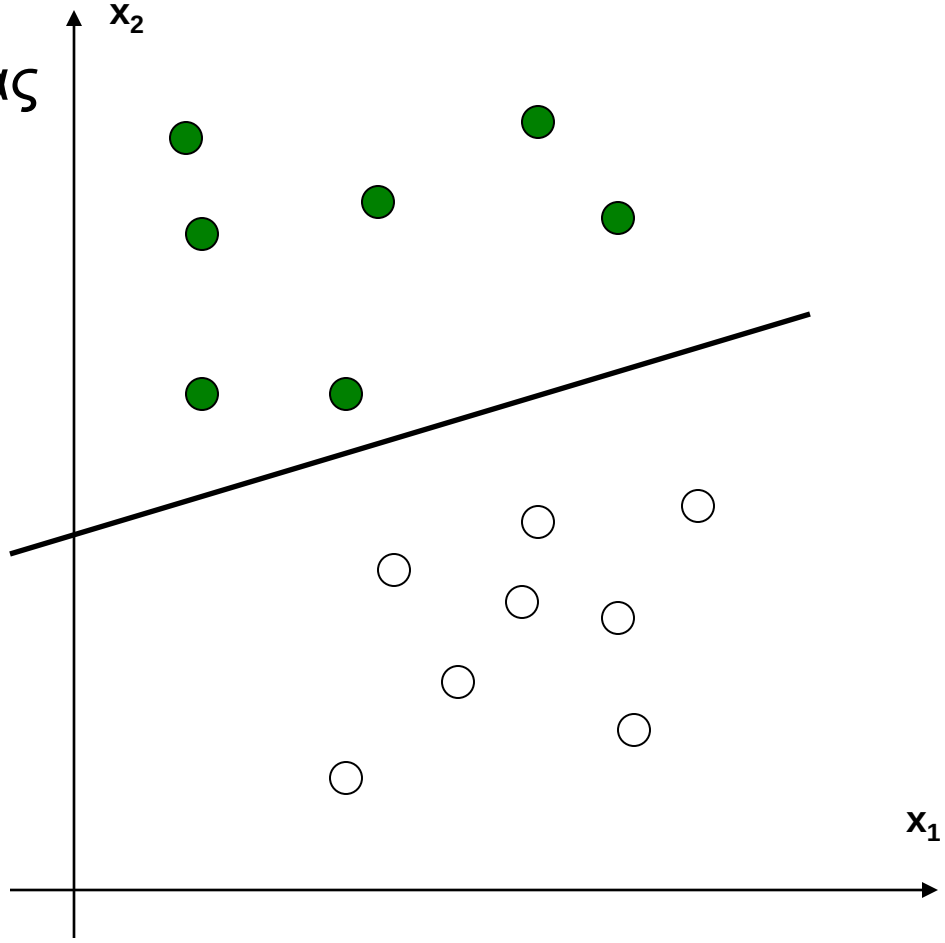
- **Άπειρος αριθμός λύσεων!**



Linear Discriminant Function

- Ποια είναι η βέλτιστη θέση της διαχωριστικής επιφάνειας ώστε να ελαχιστοποιηθεί το σφάλμα διάκρισης των 2 κατηγοριών;

- **Άπειρος αριθμός λύσεων!**

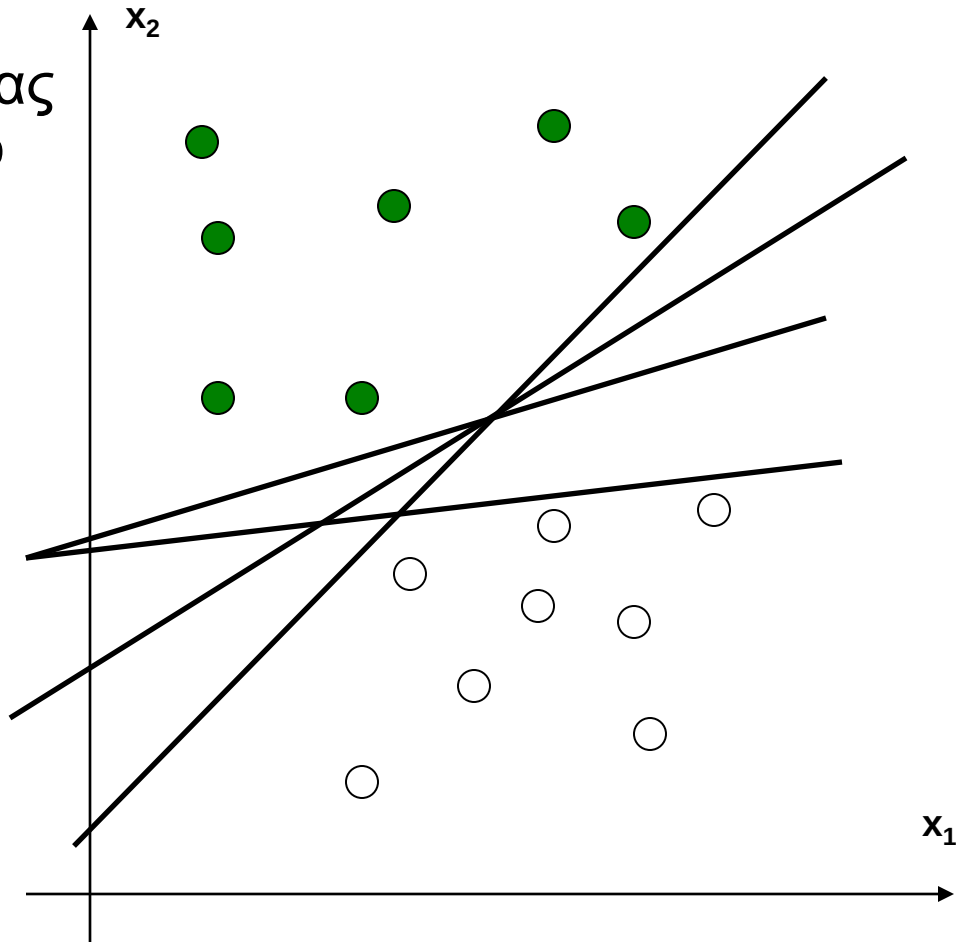


Linear Discriminant Function

- Ποια είναι η βέλτιστη θέση της διαχωριστικής επιφάνειας ώστε να ελαχιστοποιηθεί το σφάλμα διάκρισης των 2 κατηγοριών;

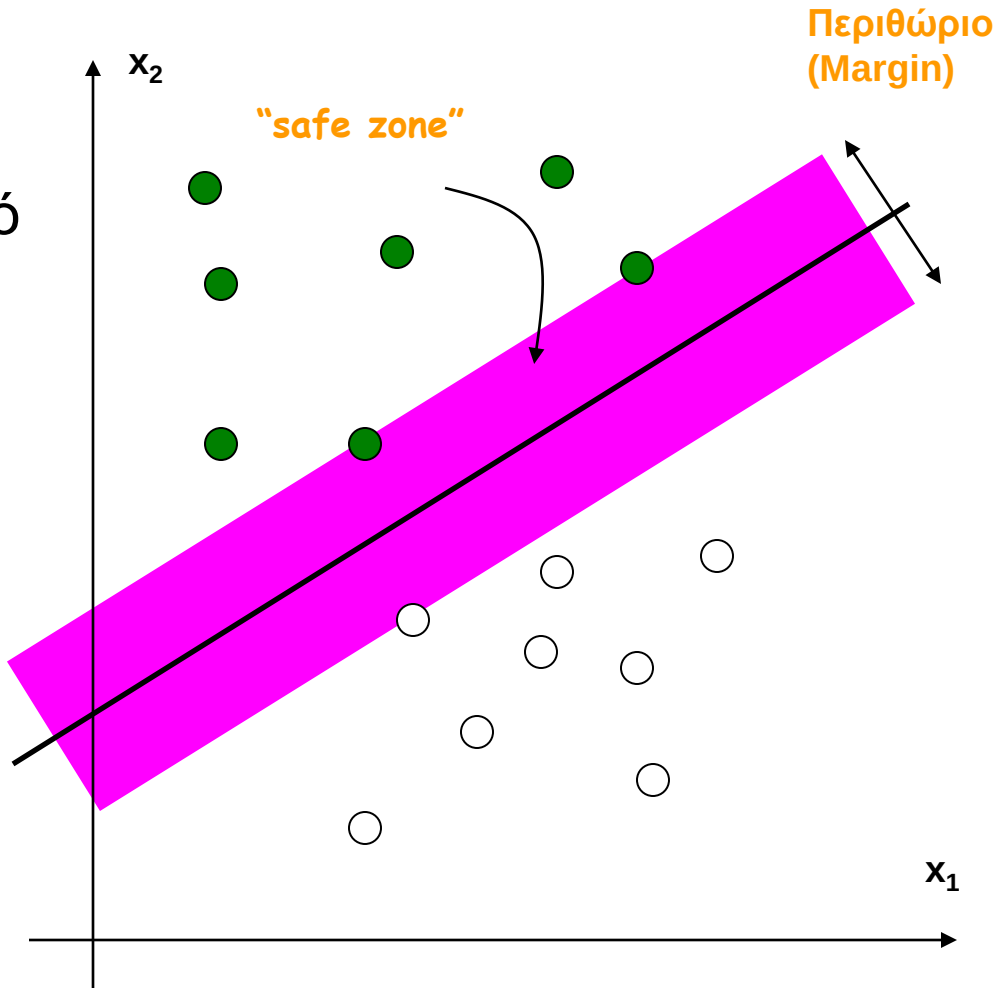
- **Άπειρος αριθμός λύσεων!**

- **Ποια είναι η βέλτιστη λύση?**



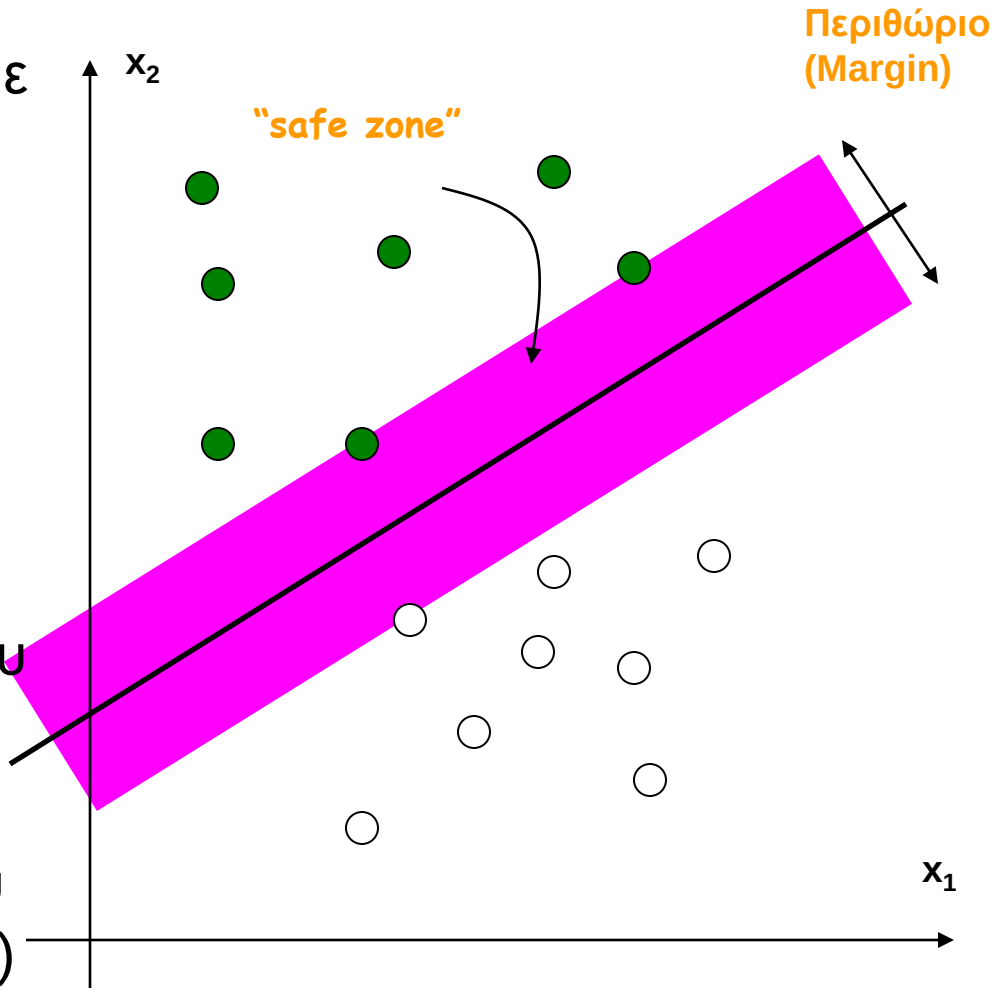
Λύση: Μεγιστοποίηση Περιθωρίου

- **Ορισμός 1**: Περιθώριο (Margin) είναι η μικρότερη απόσταση των σημείων από την διαχωριστική επιφάνεια
- **Ορισμός 2: Περιθώριο (Margin) είναι το πλάτος του ορίου γύρω από την διαχωριστική επιφάνεια που μπορεί να επεκταθεί χωρίς να καλύψει κάποιο σημείο.**
- **Γιατί είναι η βέλτιστη λύση?**



Μεγιστοποίηση Περιθωρίου (συν.)

- Η λογική είναι να επιλέξουμε να τοποθετήσουμε την διαχωριστική επιφάνεια σε τέτοια θέση ώστε να **μεγιστοποιηθεί το περιθώριο** ανάμεσα στις 2 κατηγορίες.
- Έτσι θα **ελαχιστοποιηθεί το ρίσκο** της απόφασης του ταξινομητή.
- Παράλληλα **αυξάνεται η γενικευτική ικανότητα** του ταξινομητή (**Vapnick, 1963**)



Μεγιστοποίηση Περιθωρίου (συν.)

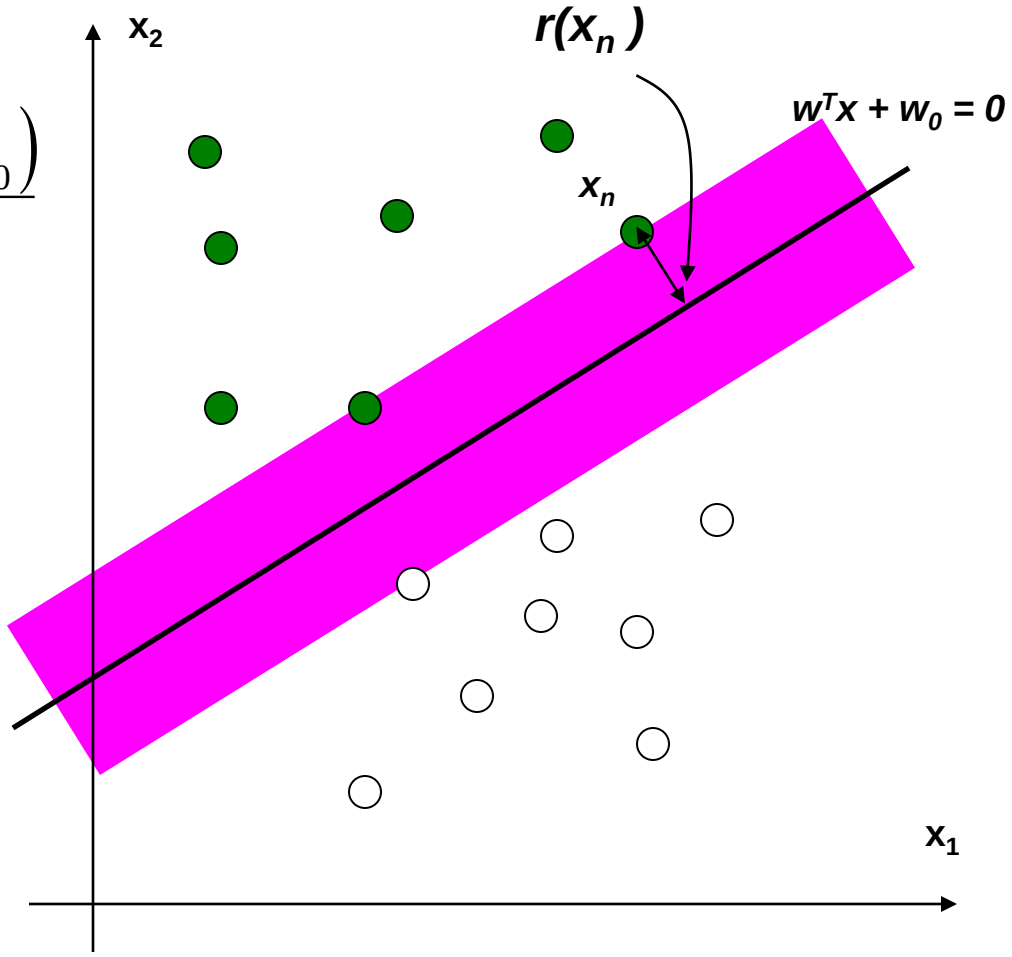
- Απόσταση σημείου x_n :

$$r(x_n) = \frac{|w^T x_n + w_0|}{\|w\|^2} = \frac{y_n(w^T x_n + w_0)}{\|w\|^2}$$

- Έτσι το περιθώριο ορίζεται:

$$\text{margin} = \min_{x_n}\left\{2\frac{y_n(w^T x_n + w_0)}{\|w\|^2}\right\}$$

$$= \frac{2}{\|w\|^2}\min_{x_n}\left\{y_n(w^T x_n + w_0)\right\}$$



Μεγιστοποίηση Περιθωρίου (συν.)

- **Πρόβλημα** μεγιστοποίησης περιθωρίου (maximum margin)

$$\{\hat{w}, \hat{w}_0\}: \max_{w, w_0} \left\{ \frac{2}{\|w\|^2} \min_{x_n} \left\{ y_n \left(w^T x_n + w_0 \right) \right\} \right\}$$

- Η επίλυση του είναι δύσκολη. Καταφεύγουμε σε ένα **τρυκ**:

- Μπορούμε να πολ/σουμε με μια σταθερά ***k*** και η απόσταση $r(x_n)$ να παραμείνει σταθερή (scaling factor).
- Ετσι, μπορούμε να διαλέξουμε μία τιμή του k έτσι ώστε για το σημείο $\boldsymbol{x}^*$ με την μικρότερη απόσταση να ισχύει:

$$y_n \left(w^T x^* + w_0 \right) = 1$$

Μεγιστοποίηση Περιθωρίου (συν.)

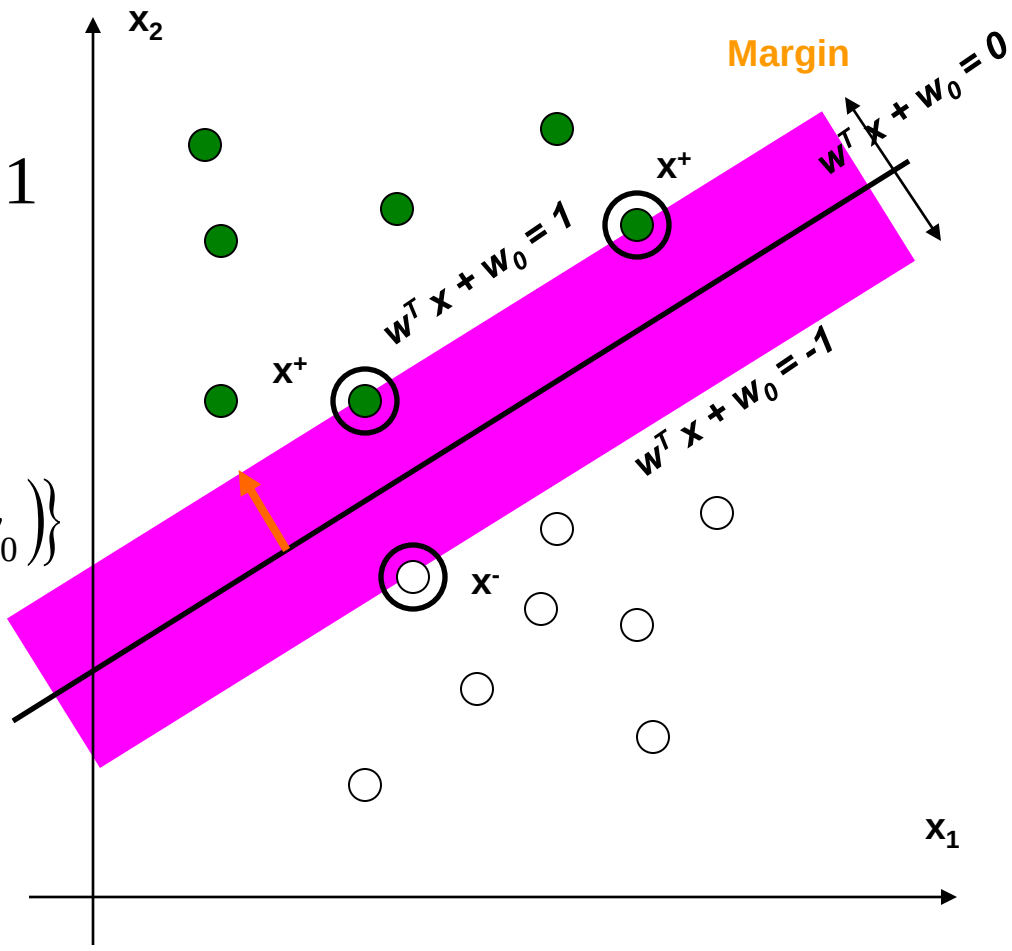
- Επομένως ισχύει ότι:

$$\forall x_n \in X \quad : \quad y_n\left(w^T x_n + w_0\right) \geq 1$$

- Το περιθώριο είναι:

$$\text{margin} = \frac{2}{\|w\|^2} \min_{x_n} \left\{ y_n\left(w^T x_n + w_0\right) \right\}$$

$$= \frac{2}{\|w\|^2}$$



Το πρόβλημα μεγιστοποίησης περιθωρίου

$$\{\hat{w}, \hat{w}_0\} : \max_{w, w_0} \left\{ \frac{2}{\|w\|^2} \right\} \quad \textbf{s.t.} \; y_n \left(w^T x_n + w_0 \right) \geq 1 \;\; \forall n$$

$$\{\hat{w}, \hat{w}_0\} : \min_{w, w_0} \left\{ \frac{1}{2} \|w\|^2 \right\} \quad \textbf{s.t.} \; y_n \left(w^T x_n + w_0 \right) \geq 1 \;\; \forall n$$

πρόβλημα **κυρτού τετραγωνικού προγραμματισμού** (***convex quadratic programming***) με ανισοτικούς περιορισμούς.

Πολλαπλασιαστές Lagrange

- Πρόβλημα ελαχιστοποίησης με **ισοτικούς** περιορισμούς

$$\min_{\theta}\{f(\theta)\} \quad \text{s.t.} \quad g(\theta)=c \Rightarrow g(\theta)-c=0$$

- Λύση με **πολλαπλασιαστές Lagrange**: Κατασκευή συνάρτησης ενσωματώνοντας τους περιορισμούς:

$$L(\theta,\lambda)= f(\theta)-\lambda(g(\theta)-c)$$

- ✓ *λ:* ο πολλαπλασιαστής Lagrange
- ✓ Ισχύει ότι: $\min_{\theta} L(\theta,\lambda)=\min_{\theta} f(\theta)$ και $g(\hat{\theta})-c=0$

$$L(\hat{\theta},\lambda)= f(\hat{\theta})$$

Πολλαπλασιαστές Lagrange (συν.)

- Πρόβλημα ελαχιστοποίησης με **ανισοτικούς** περιορισμούς

$$\min_{\theta}\{f(\theta)\} \text{ s.t. } g(\theta)\geq c \Rightarrow g(\theta)-c\geq 0$$

- Λύση με **πολλαπλασιαστές Lagrange**: Κατασκευή συνάρτησης ενσωματώνοντας τους περιορισμούς:

$$L(\theta,\lambda)= f(\theta)-\lambda(g(\theta)-c)$$

- ✓ Ισχύουν οι **Karush-Khun-Tucker (KKT) συνθήκες** στην λύση:

$$\lambda \geq 0$$

$$g(\theta)-c\geq 0$$

$$\lambda(g(\theta)-c)=0$$

Λύση του προβλήματος μεγιστοποίησης περιθωρίου με Lagrange

Quadratic programming with linear constraints

$$\min_{w,w_0}\left\{\frac{1}{2}\|w\|^2\right\}$$

$$\text{s.t. } y_n\left(w^T x_n + w_0\right)\geq 1 \ \forall n$$

Συνάρτηση Lagrange

$$L\left(w, w_0, a\right) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n\left(y_n\left(w^T x_n + w_0\right) - 1\right)$$

$$\{\hat{w}, \hat{w}_0\}: \min_{w, w_0}\left\{\frac{1}{2}\|w\|^2\right\} \text{ s.t. } y_n\left(w^T x_n + w_0\right) \geq 1 \ \forall n$$

$$L\left(w, w_0, a\right) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n\left(y_n\left(w^T x_n + w_0\right) - 1\right)$$

ΚΚΤ συνθήκες

$$a_n \geq 0$$

$$y_n\left(w^T x_n + w_0\right) - 1 \geq 0$$

$$a_n\left(y_n\left(w^T x_n + w_0\right) - 1\right) = 0$$

$$\text{minimize } L(w, w_0, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left(y_n \left(w^T x_n + w_0 \right) - 1 \right)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \hat{w} = \sum_{n=1}^{N} a_n y_n x_n$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{n=1}^{N} a_n y_n = 0$$

Prime problem

$$\text{minimize } L(w, w_0, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left(y_n \left(w^T x_n + w_0 \right) - 1 \right)$$

Συνάρτηση Lagrange του Δυϊκού (Dual) προβλήματος

$$\text{maximize } L_D(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m y_n y_m x_n^T x_m$$

$$\textbf{s.t. } a_n \geq 0 \quad , \quad \sum_{n=1}^{N} a_n y_n = 0$$

Χρήσιμες παρατηρήσεις της λύσης

1. Το **αρχικό πρόβλημα** έχει *d+1* πλήθος αγνώστων παραμέτρων για τους γραμμικούς συντελεστές {*w*, w_0}, όπου *d* η διάσταση των δεδομένων.

   Το **δϋικό πρόβλημα** έχει *N* πλήθος αγνώστων για τους πολ/στές Lagrange { a_n }, όπου *N* το πλήθος των δεδομένων του συνόλου εκπαίδευσης.

   Έτσι η μέθοδος SVM βολεύει για **πολυδιάστατα** δεδομένα όπου το *d*>>*N* καθώς κατασκευάζει ένα (δϋικό) χώρο αναζήτησης αρκετά μικρότερο από τον αρχικό χώρο.

Χρήσιμες παρατηρήσεις της λύσης (συν.)

2. Ο κανόνας απόφασης γίνεται:

$$\left.\begin{aligned} f(x) &= w^T x + w_0 \\ \hat{w} &= \sum_{n=1}^{N} a_n y_n x_n \end{aligned}\right| \Rightarrow f(x) = \sum_{n=1}^{N} a_n y_n x_n^T x + w_0$$

δηλ. ένας γραμμικός συνδυασμός εσωτερικών γινομένων του υπό-εξέταση σημείου x με όλα τα σημεία x_n του συνόλου εκπαίδευσης, με συντελεστές που σχετίζονται με τους πολ/στές Lagrange.

Χρήσιμες παρατηρήσεις της λύσης (συν.)

3. Οι **ΚΚΤ συνθήκες ορίζονται** ως εξής:

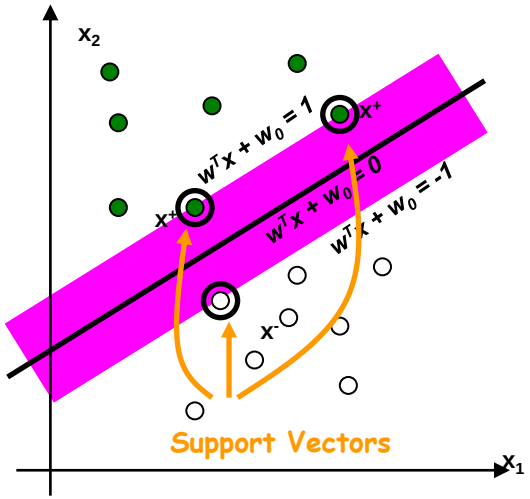
$$a_n \geq 0$$
$$y_n\left(w^T x_n + w_0\right) - 1 \geq 0$$
$$a_n\left(y_n\left(w^T x_n + w_0\right) - 1\right) = 0$$

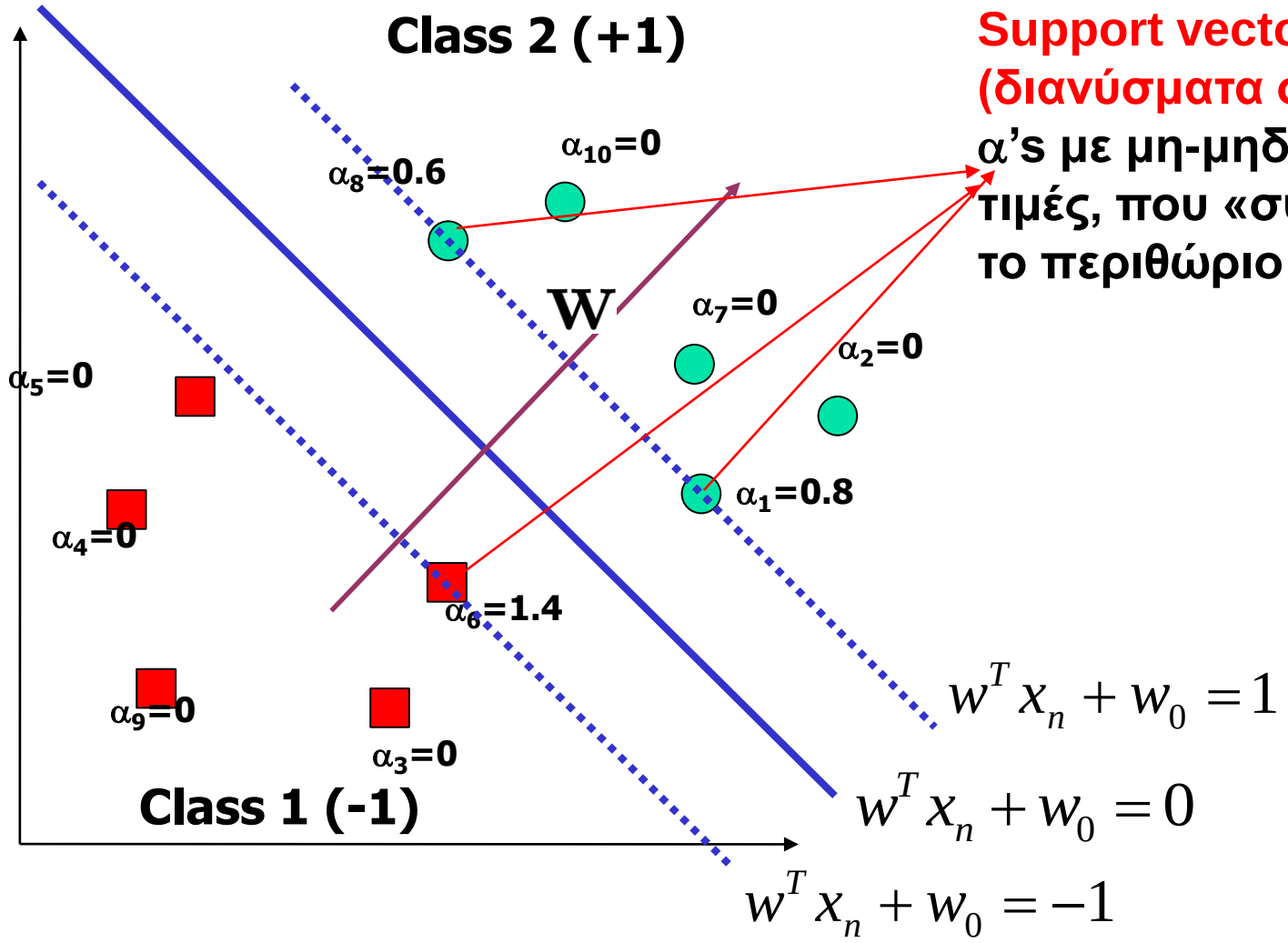
Έτσι: $a_n = 0$ **ή** $y_n\left(w^T x_n + w_0\right) - 1 = 0$

δηλ. τα σημεία που βρίσκονται πάνω στο περιθώριο (margin)

Χρήσιμες παρατηρήσεις της λύσης (συν.)

- Όσα σημεία βρίσκονται εκτός του περιθωρίου, δηλ. $a_n=0$, δεν παίζουν κάποιο ενεργό ρόλο καθώς δεν χρησιμοποιούνται στον κανόνα απόφασης της μεθόδου.
- Όσα σημεία βρίσκονται στο περιθώριο όπου ισχύει $y_n\left(w^T x_n + w_0\right) - 1 = 0$ έχουν $a_n > 0$. Αυτά ονομάζονται ***support vectors*** (διανύσματα στήριξης) και παίζουν ενεργό ρόλο στην απόφαση. Μάλιστα, εκφράζουν τον βαθμό σημαντικότητας των δεδομένων.

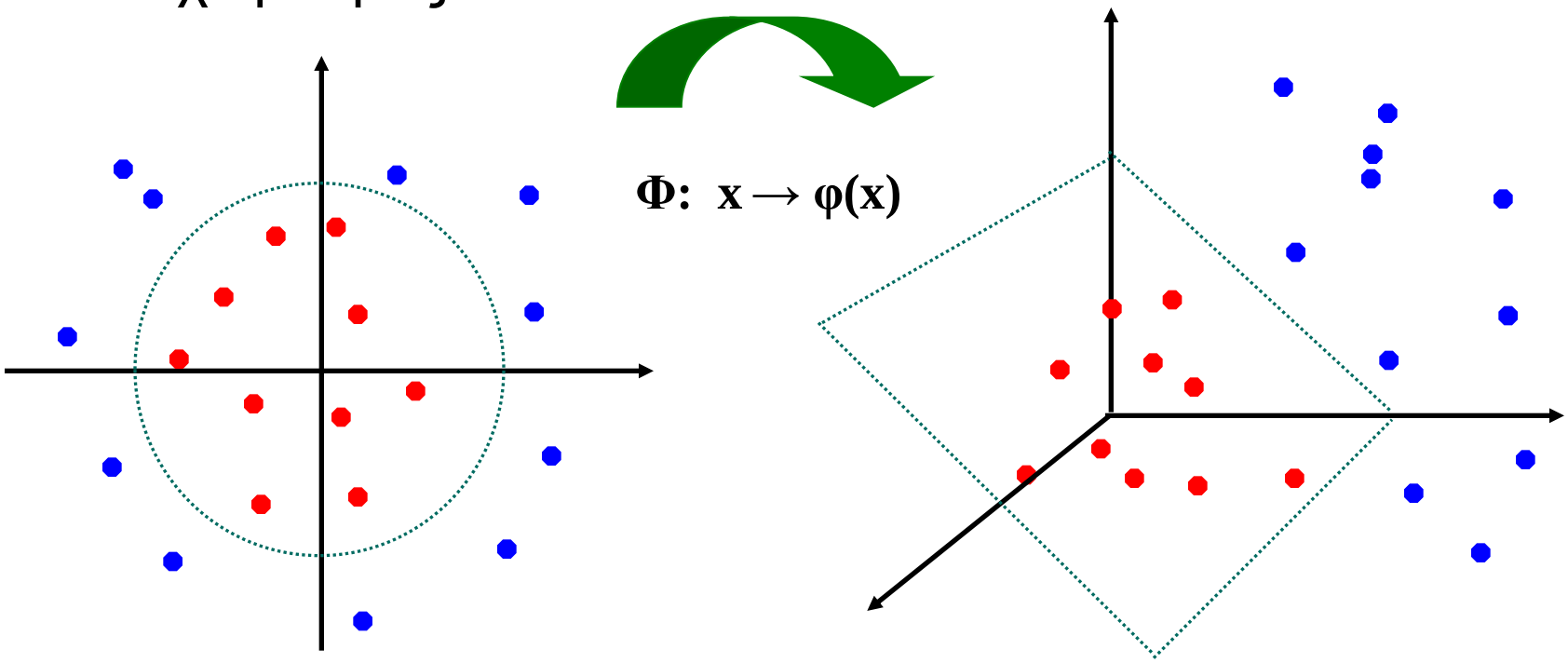



Class 2 (+1)
Support vectors
(διανύσματα στήριξης)
α's με μη-μηδενικές
τιμές, που «συγκρατούν»
το περιθώριο
α10=0
α8=0.6
W
α7=0
α2=0
α5=0
α1=0.8
α4=0
α6=1.4
α9=0
α3=0
Class 1 (-1)
$w^T x_n + w_0 = 1$
$w^T x_n + w_0 = 0$
$w^T x_n + w_0 = -1$

Χρήσιμες παρατηρήσεις της λύσης (συν.)

4. **Kernel trick.** Χρήση συνάρτησης φ(x).

- **Ιδέα:** Ο χώρος των δεδομένων μετασχηματίζεται σε έναν μεγαλύτερης διάστασης χώρο που είναι γραμμικά διαχωρίσιμος



4. Kernel trick

- Όλα τα εσωτερικά γινόμενα παίρνουν την μορφή:

$$x_i^T x_j \quad \rightarrow \quad \phi(x_i)^T \phi(x_j) \equiv K(x_i, x_j)$$

Kernel function

- Συνάρτηση πυρήνα (***Kernel function***)**:** συνάρτηση που ορίζεται ως το εσωτερικό γινόμενο δύο διανυσμάτων του μετασχηματισμένου χώρου (εκφράζει ομοιότητα).
- Έτσι ο κανόνας απόφασης μετατρέπεται:

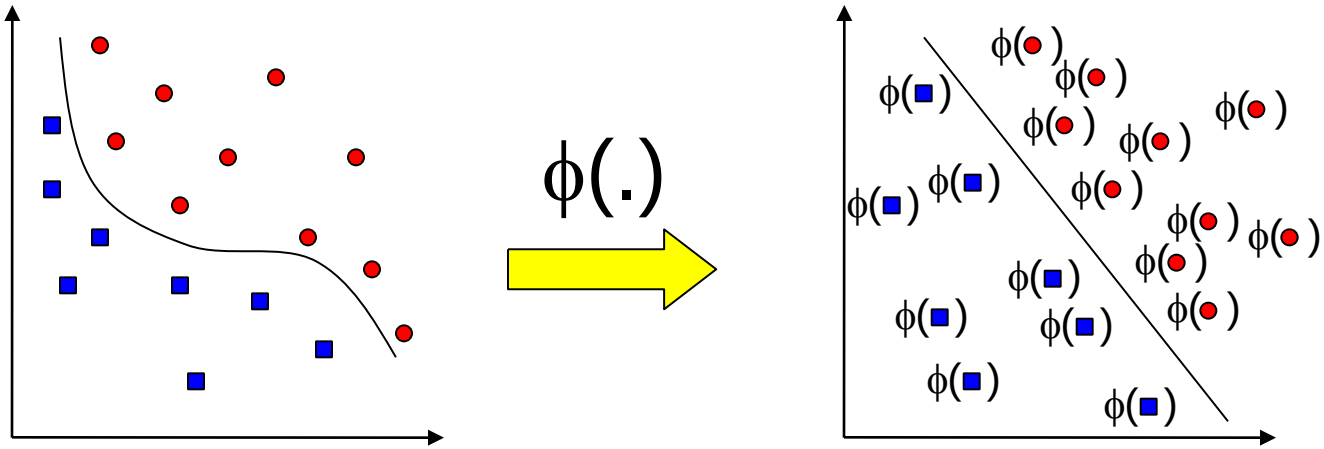
$$f(x) = \sum_{n=1}^{N} a_n y_n x_n^T x + w_0 \quad \rightarrow \quad f(x) = \sum_{n=1}^{N} a_n y_n \phi(x_n)^T \phi(x) + w_0$$

$$f(x) = \sum_{n=1}^{N} a_n y_n K(x_n, x) + w_0$$

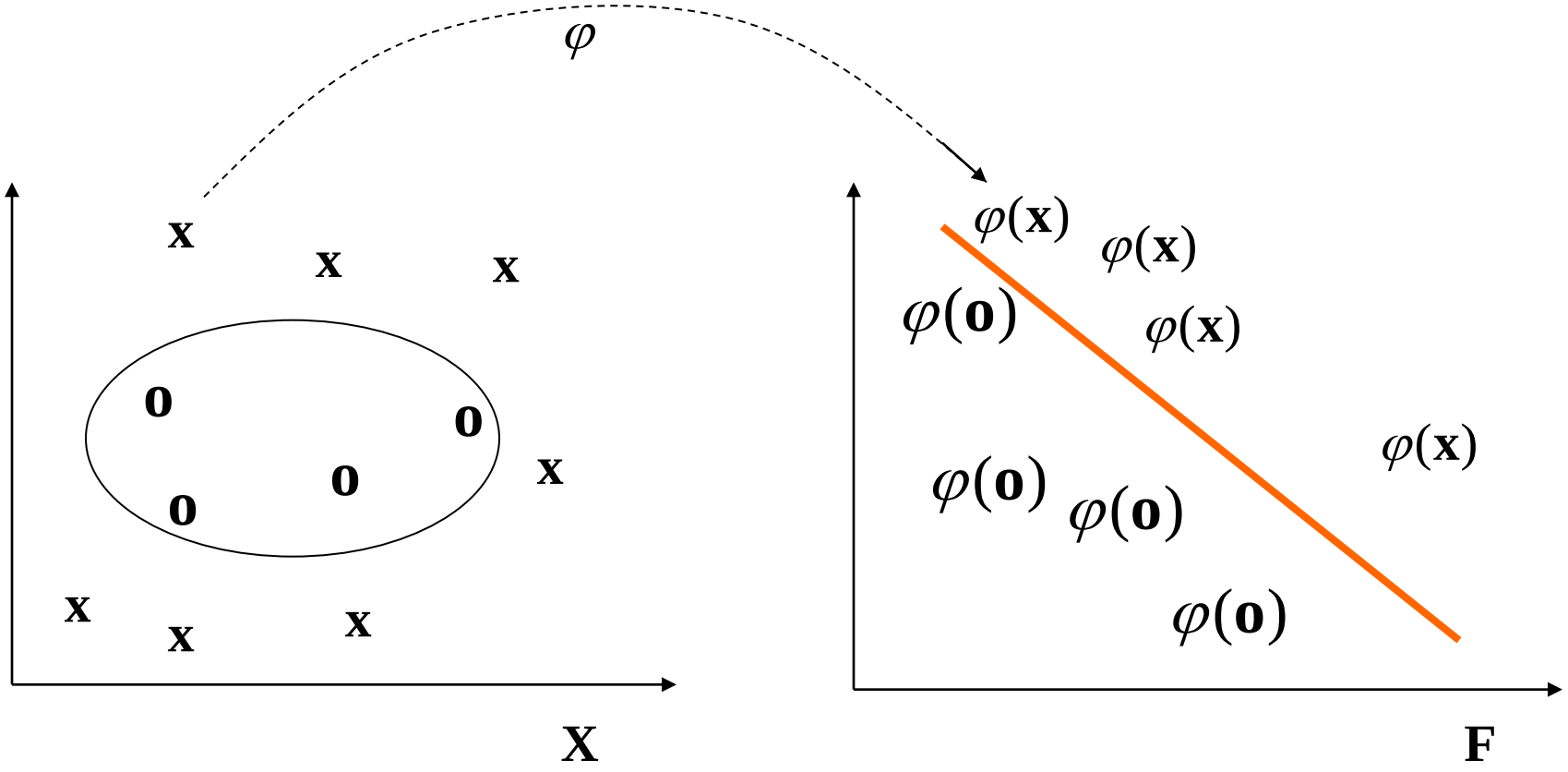
Παραδείγματα συναρτήσεων πυρήνα $K(x_i, x_j)$

- **Linear Kernel** $K(x_i, x_j) = x_i^T x_j$
- **Polynomial Kernel** $K(x_i, x_j) = (x_i^T x_j + 1)^p$
- **Gaussian ή RBF Kernel** $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2o^2}}$
- **Cosine** $K(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\|^2 \|x_j\|^2}$
- **Sigmoid** $K(x_i, x_j) = \tanh(\beta_0 x_i^T x_j + \beta_1)$
- **……**

- **Κατασκευή ενός γραμμικού χώρου χαρακτηριστικών (feature space) μέσω της φ(x)**



- **Κατασκευή ενός γραμμικού χώρου χαρακτηριστικών (feature space) μέσω της φ(x)**



- Κατασκευή συναρτήσεων πυρήνα ανάλογα με το πρόβλημα και του τύπου δεδομένων

- Παράδειγμα: Συνάρτηση πυρήνα δεδομένων που περιγράφονται με ιστογράμματα, π.χ. εικόνες

$$K(h_1, h_2) = \sum_{i=1}^{N} \min(h_1(i), h_2(i))$$

Χρήσιμες παρατηρήσεις της λύσης (συν.)

5. Εύρεση του σταθερού όρου w_0

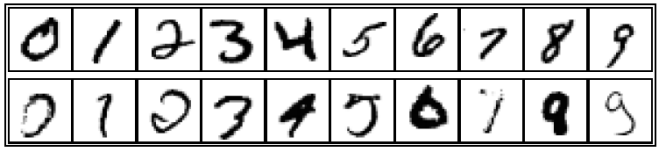
- Σύνολο των support vectors $S = \left\{ x_n : y_n \left(w^T \phi(x_n) + w_0 \right) = 1 \right\}$
- Αλλά καθώς: $\hat{w} = \sum_{n=1}^{N} a_n y_n \phi(x_n)$
- Αντικαθιστώντας
$$y_n \left(\sum_{x_m \in S} a_m y_m \phi(x_m)^T \phi(x_n) + w_0 \right) = 1 \quad \forall x_n \in S$$
- Και αθροίζοντας παίρνουμε τελικά:
$$\sum_{x_n \in S} y_n \left(\sum_{x_m \in S} a_m y_m K(x_m, x_n) + w_0 \right) = N_s = |S|$$
Πληθάριθμος S

- Δοκιμάζουμε διάφορες τιμές του w_0 ώστε να πετύχουμε την ισότητα

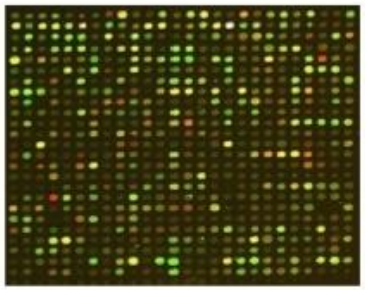
Εφαρμογές των SVMs

Κυρίως σε περιπτώσεις πολυδιάστατων δεδομένων

- **Βιοπληροφορική (Bioinformatics) – gene expression data analysis**
- **Text categorization – mining**
- **Handwritten character recognition**
- **Machine Vision**
- **Time series analysis**



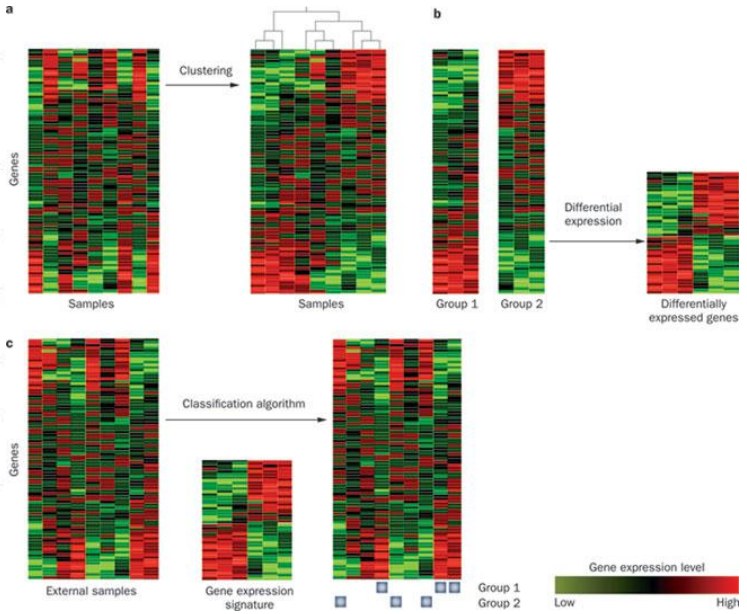
- **Βιοπληροφορική (Bioinformatics) – gene expression data**

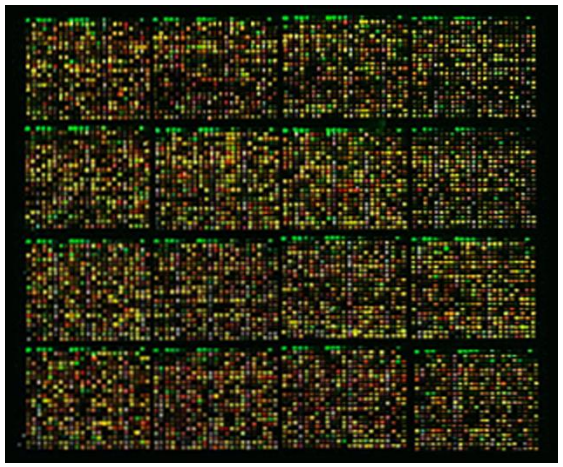


Microarray Chip

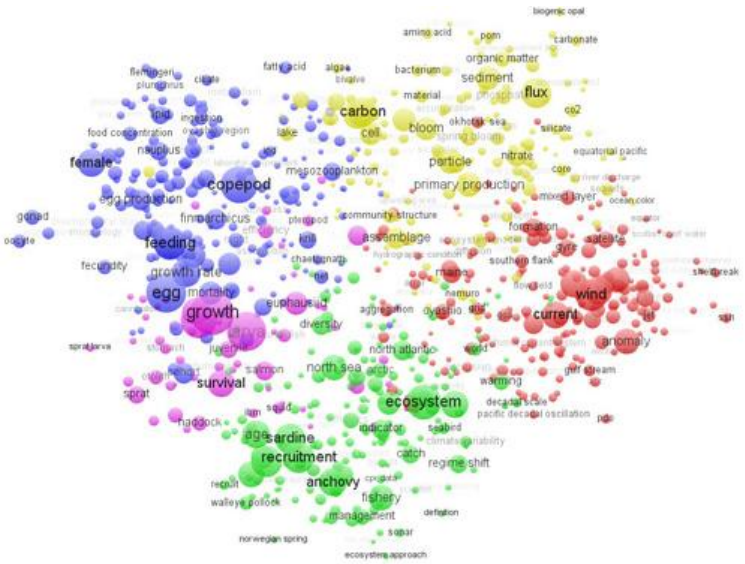
n Samples

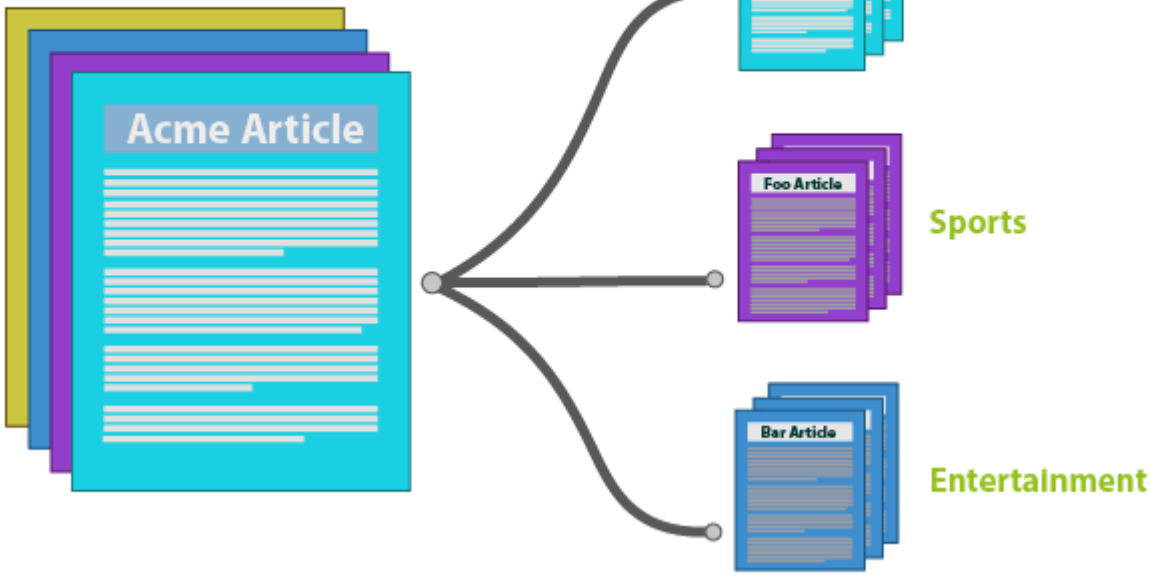
Gene	Gene Expression
a	
b	
c	
...	
n	





- **Text categorization – mining**



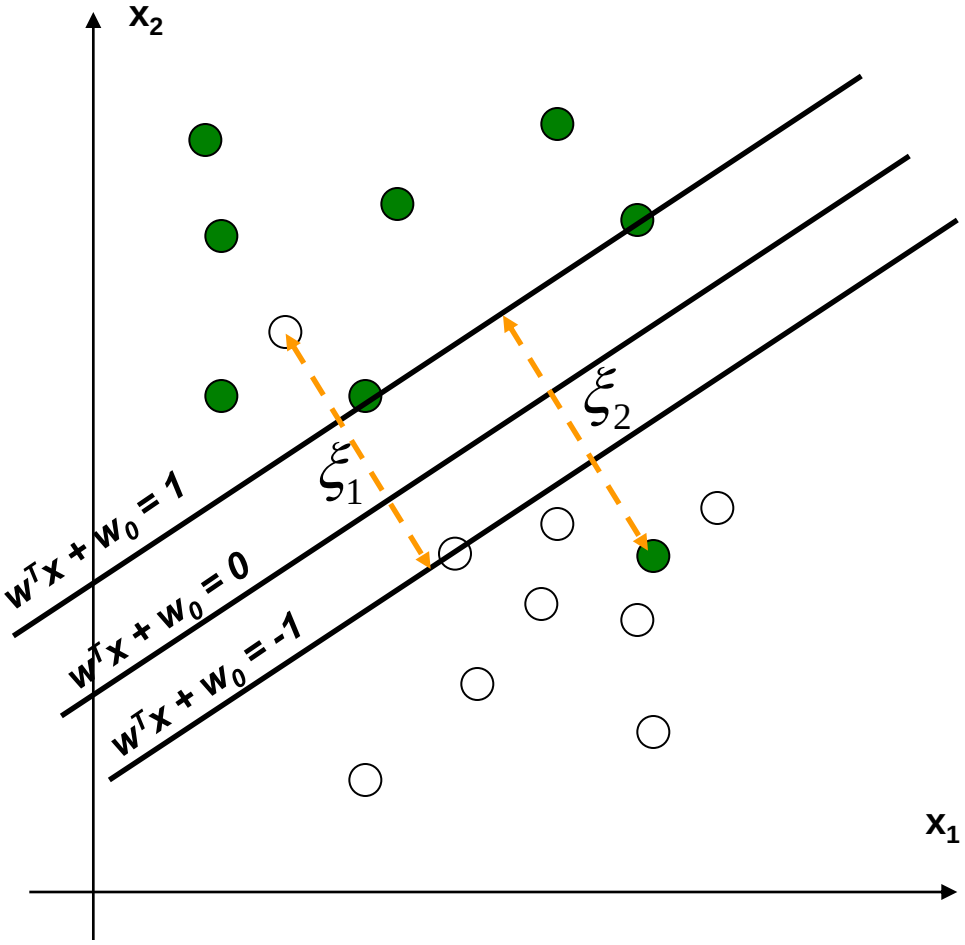


Bag of words (lexicon)

Nonlinear SVM
Μη-γραμμικά διαχωρίσιμη περίπτωση

- Τι γίνεται στην περίπτωση όπου τα δεδομένα είναι μη-γραμμικά διαχωρίσιμα?

- Εισαγωγή **βοηθητικών μεταβλητών** ξ_i που επιτρέπουν σφάλματα, δηλ. σημεία να βρίσκονται σε λάθος πλευρά του περιθωρίου.

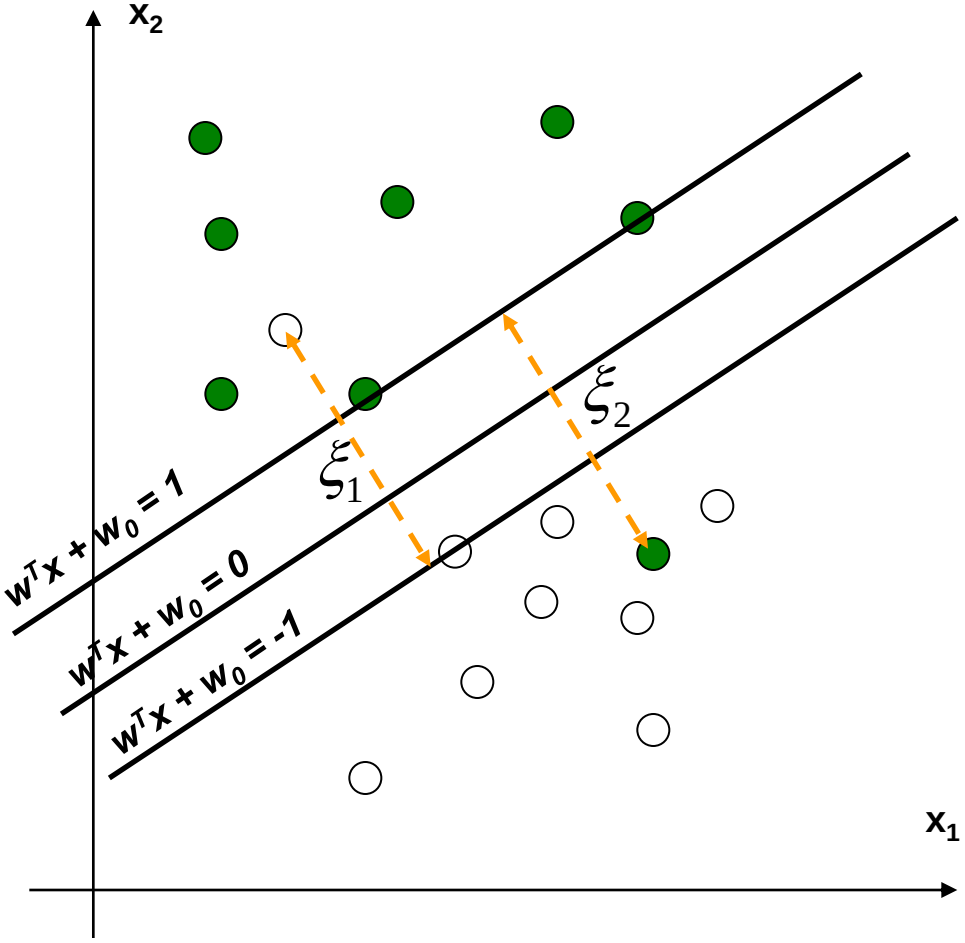


Nonlinear SVM
Μη-γραμμικά διαχωρίσιμη περίπτωση

- Αν το σημείο x_n βρίσκεται στη σωστή πλευρά (όχι σφάλμα) τότε $\xi_n = 0$.
- Αν υπάρχει σφάλμα, τότε:

$$\xi_n = \left| y_n - f\left(x_n\right) \right|$$

- Έτσι, αν βρίσκεται πάνω στην $w^T x + w_0 = 0$ τότε $\xi_n = 1$.
- Αν είναι λάθος ταξινομημένο τότε $\xi_n > 1$

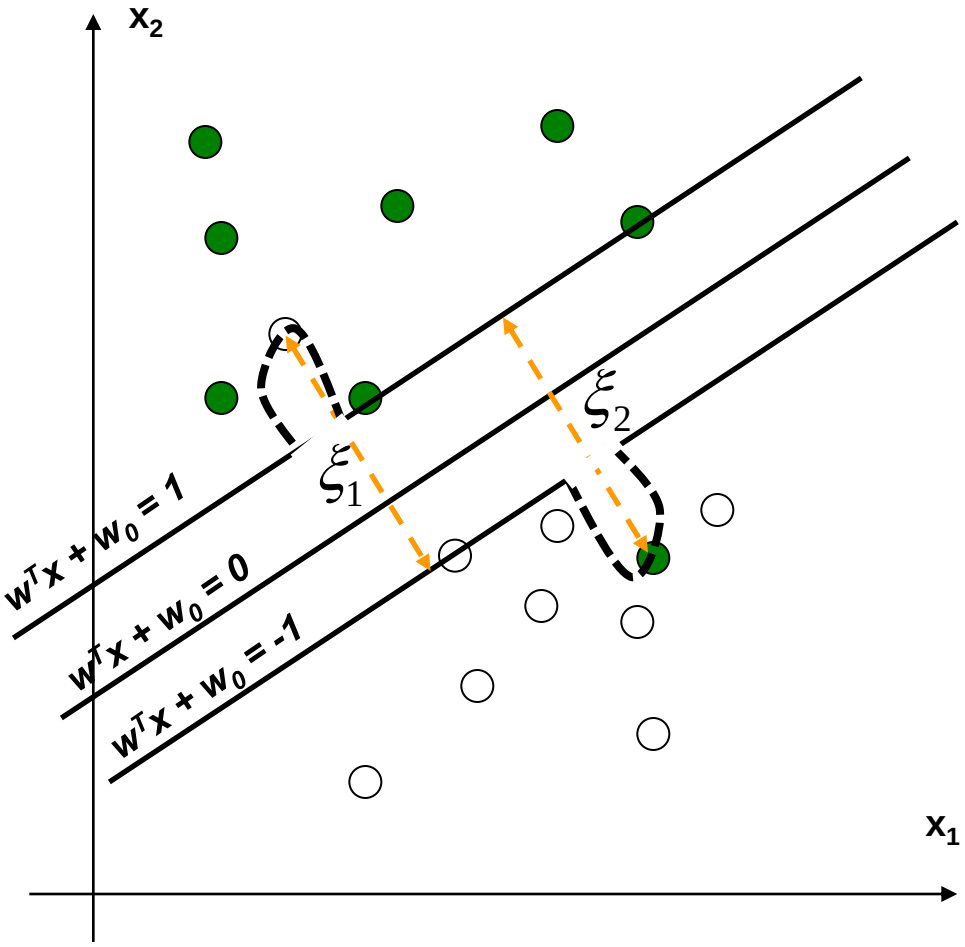


Nonlinear SVM
Μη-γραμμικά διαχωρίσιμη περίπτωση

- Έτσι θέλουμε $\forall\ n$

$$y_n\left(w^T x_n + w_0\right) \geq 1 - \xi_n$$

- Δηλ. το ξ_n παρέχει την **ανοχή στο σφάλμα** για κάθε σημείο x_n και άρα το περιθώριο τοπικά δύναται να εισέλθει στο χώρο της άλλης κατηγορίας.



Nonlinear SVM
Κατασκευή της αντικειμενικής συνάρτησης

- Το $\sum_{n=1}^{N} \xi_n$ εκφράζει το συνολικό σφάλμα

- **Πρόβλημα:**

$$\min_{w, w_0, \xi} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} \xi_n \right\}$$

$$\textbf{s.t.} \;\; y_n \left(w^T x_n + w_0 \right) \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0$$

***C*:** ελεύθερη παράμετρος

- **Πρόβλημα:**

$$\min_{w,w_0,\xi}\left\{\frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n\right\}$$

$$\textbf{s.t.}\ \ y_n\left(w^T x_n + w_0\right) \geq 1-\xi_n \quad \forall n$$

$$\xi_n \geq 0$$

Συνάρτηση Langrange

$$L\left(w, w_0, \xi\right) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}a_n\left(y_n\left(w^T x_n + w_0\right) - 1 - \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

$$\text{minimize } L(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\left(y_n\left(w^T x_n + w_0\right) - 1 - \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \hat{w} = \sum_{n=1}^{N} a_n y_n x_n$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{n=1}^{N} a_n y_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow a_n = C - \mu_n$$

$$\text{minimize } L(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\left(y_n\left(w^T x_n + w_0\right) - 1 - \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

ΚΚΤ συνθήκες

$$a_n \geq 0$$

$$y_n\left(w^T x_n + w_0\right) - 1 + \xi_n \geq 0$$

$$a_n\left(y_n\left(w^T x_n + w_0\right) - 1 + \xi_n\right) = 0$$

ή

$$a_n = 0 \qquad y_n\left(w^T x_n + w_0\right) - 1 + \xi_n = 0$$

$$\mu_n \geq 0$$

$$\xi_n \geq 0$$

$$\mu_n\xi_n = 0$$

ή

$$\mu_n = 0 \qquad \xi_n = 0$$

$$\text{minimize} \;\; L(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\left(y_n\left(w^T x_n + w_0\right) - 1 - \xi_n\right) - \sum_{n=1}^{N}\mu_n\xi_n$$

Συνάρτηση Langrange του Δυϊκού (Dual) προβλήματος

$$\text{maximize} \;\; L_D(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m x_n^T x_m$$

$$\textbf{s.t.} \quad 0 \le a_n \le C \quad , \quad \sum_{n=1}^{N} a_n y_n = 0$$

- Δυϊκό πρόβλημα (Dual problem)

$$\text{maximize } L_D(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m x_n^T x_m$$

$$\textbf{s.t.} \quad 0 \le a_n \le C \quad , \quad \sum_{n=1}^{N} a_n y_n = 0$$

- Αν $\alpha_n > 0$ τότε τα x_n είναι τα support vectors και ισχύει:

$$y_n\left(w^T x_n + w_0\right) - 1 + \xi_n = 0$$

- Αν $\alpha_n < C$ τότε $\mu_n > 0$ και άρα $\xi_n = 0$. Ισχύει:

$$y_n\left(w^T x_n + w_0\right) - 1 = 0$$

- Δυϊκό πρόβλημα (Dual problem) (συν.)

$$\text{maximize } L_D(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m x_n^T x_m$$

$$\textbf{s.t.} \quad 0 \le a_n \le C \quad , \quad \sum_{n=1}^{N} a_n y_n = 0$$

- Αν $\alpha_n = C$ τότε $\mu_n = 0$ και άρα $\xi_n > 0$. Έτσι το σημείο x_n βρίσκεται μέσα στο περιθώριο.
  - Αν $\xi \le 1$ τότε το σημείο είναι “σωστά” ταξινομημένο,
  - Αν $\xi_n > 1$ τότε είναι “εσφαλμένα” ταξινομημένα

Statistical Classifiers

- **Υπόθεση:**

  Τα πρότυπα που ανήκουν σε μια κατηγορία ω_j είναι **δείγματα** ή **παρατηρήσεις** που προέρχονται από μια κατανομή, δηλ. από έναν μηχανισμό γέννησης:

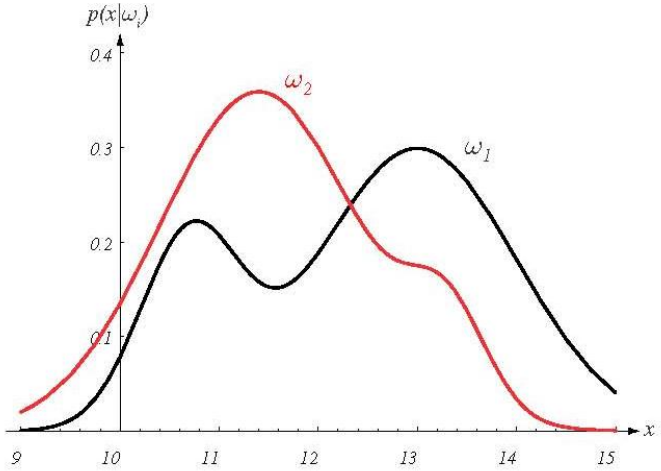
$$p(x \mid \omega_j)$$

- Επίσης κάθε κατηγορία ω_j (j=1,…,K) έχει μία πιθανότητα $P(\omega_j)$, όπου ισχύει ότι:

$$\sum_{j=1}^{K} P(\omega_j) = 1$$

Statistical Classifiers

- ***P(ω_j)*** : εκ των προτέρων (***prior***) πιθανότητα να εμφανιστεί ένα δεδομένο της κατηγορίας ω_j



- ***p(x|ω_j)***: η υπό-συνθήκη (ή δεσμευμένη) συνάρτηση πυκνότητας πιθανότητας της κατηγορίας ω_j (***class conditional pdf - likelihood***)
  - πόσο συχνά θα εμφανιστεί ένα πρότυπο με χαρακτηριστικά του *x* αν υποθέσουμε ότι ανήκει στην ω_j

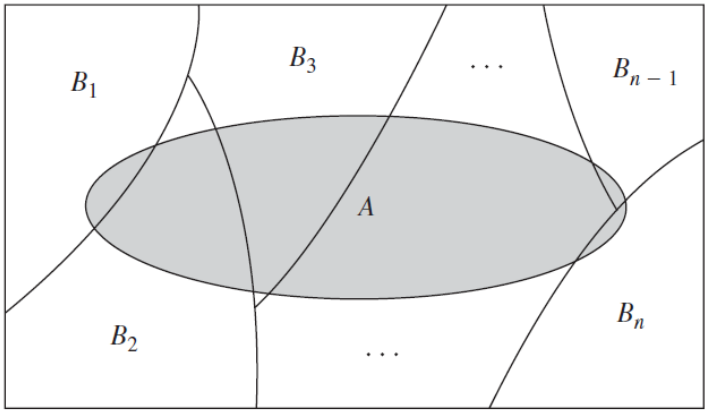
- ***P(ω_j \|x)***: η υπό-συνθήκη πιθανότητα της κατηγορίας ω_j (***posterior probability***)
  - Η εκ των υστέρων πιθανότητα ότι ένα πρότυπο με χαρακτηριστικά όπως του *x* ανήκει στην κατηγορία ω_j .
  - Το ποσό βεβαιότητας ότι ένα πρότυπο με χαρακτηριστικά όπως του x είναι πρότυπο της κατηγορίας ω_j .

- **Στόχος**: να υπολογίζουμε τις εκ των υστέρων πιθανότητες ώστε να **αποφασίζουμε** με βάση αυτές.

Ο Κανόνας (ή Νόμος) του Bayes (Thomas Bayes – 1702-61)

Έστω **δ.χ. Ω** με Κ ασυμβίβαστα ενδεχόμενα:

- $B_1 \cup B_2 \cup \ldots \cup B_K = \Omega,\ B_i \cap B_j = \varnothing \quad \forall i,j.$
- $P(B_1)+P(B_2)+\ldots+P(B_K)=P(\Omega)=1$



Έστω ότι εμφανίζεται το **ενδεχόμενο *A***.

Ψάχνουμε να βρούμε κατά **πόσο μεταβλήθηκε** η πιθανότητα του B_j **μετά** την εμφάνιση του ***A*** (***εκ των υστέρων***).

$$P(B_j \mid A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A \mid B_j)P(B_j)}{\sum_{k=1}^{K} P(A \mid B_k)P(B_k)}$$

Μπεϋζιανή Ταξινόμηση
Εφαρμογή του Bayes στο πρόβλημα της ταξινόμησης

- Έστω δ.χ. προτύπων χωρισμένος σε *K* κατηγορίες { ω_j }.
- Κάθε κατηγορία έχει μία αδέσμευτη πιθανότητα $P(\omega_j)$,
- και μία υπό-συνθήκη κατανομή $p(x|\omega_j)$.
- Ένα πρότυπο x ανήκει στην κατηγορία ω_j με πιθανότητα:

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

που είναι η εκ των υστέρων πιθανότητα $x \in \omega_j$

όπου $p(x) = \sum_{k=1}^{K} p(x \mid \omega_k)P(\omega_k)$ έτσι ώστε $\sum_{j=1}^{K} P(\omega_j \mid x) = 1$

- Απόφαση με βάση την **μέγιστη** εκ των υστέρων πιθανότητα (***maximum posterior rule***):

$$x \in \omega_{j^*} \;:\; \max_{j=1,\ldots,K} \left\{ P\left(\omega_j \mid x\right) \right\} = \max_{j=1,\ldots,K} \left\{ \frac{p\left(x \mid \omega_j\right) P\left(\omega_j\right)}{p\left(x\right)} \right\}$$

$$= \max_{j=1,\ldots,K} \left\{ \frac{p\left(x \mid \omega_j\right) P\left(\omega_j\right)}{\sum_{k=1}^{K} p\left(x \mid \omega_k\right) P\left(\omega_k\right)} \right\}$$

$$= \max_{j=1,\ldots,K} \left\{ p\left(x \mid \omega_j\right) P\left(\omega_j\right) \right\}$$

- **Για K=2 κατηγορίες {ω_1, ω_2}**

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

or

Decide ω_1 if $p(x \mid \omega_1)P(\omega_1) > p(x \mid \omega_2)P(\omega_2)$; otherwise decide ω_2

or

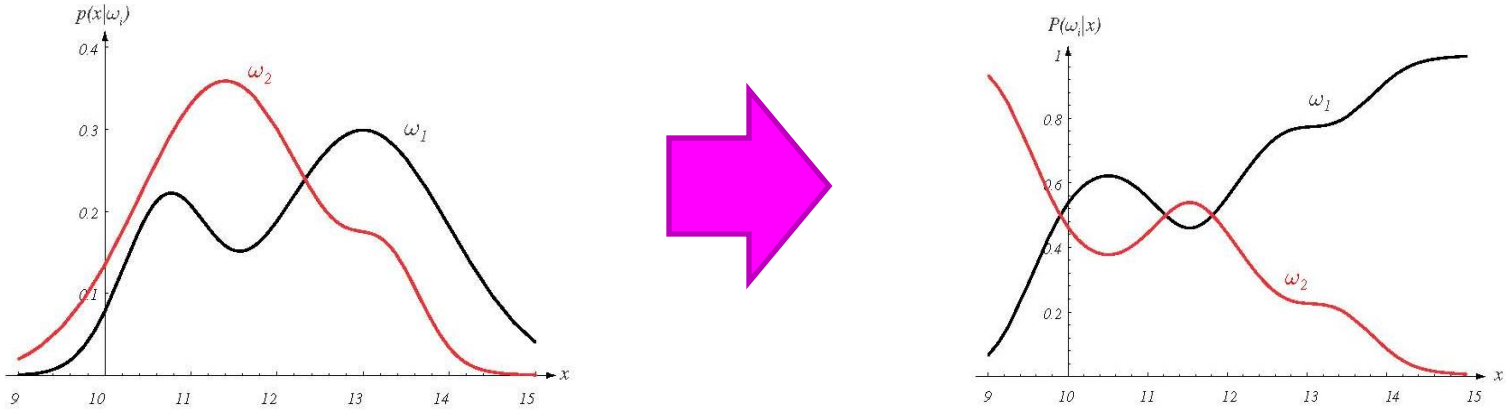
Decide ω_1 if $\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$; otherwise decide ω_2

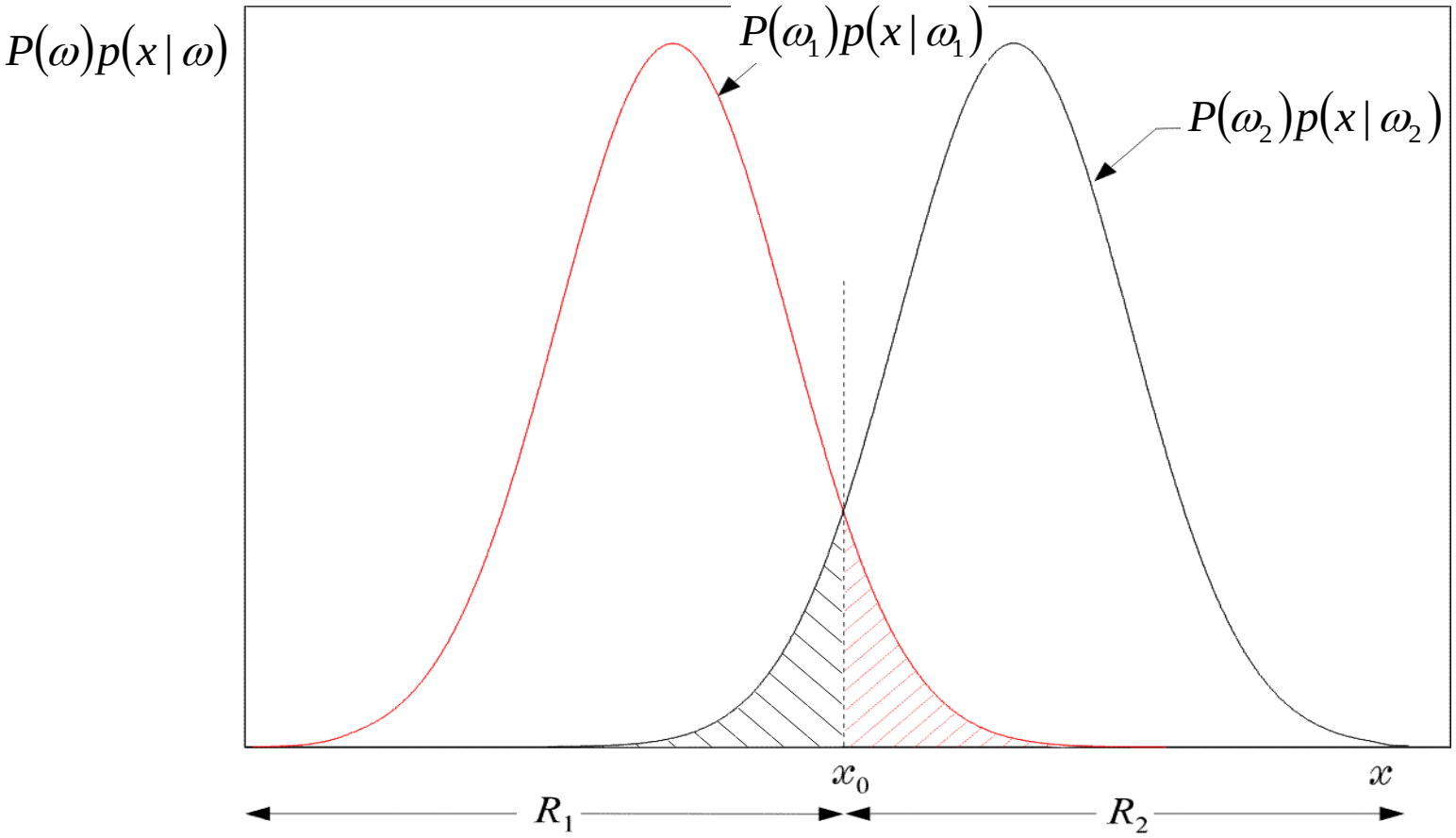
likelihood ratio **threshold**

- Παράδειγμα για K=2 κατηγορίες $\{\omega_1, \omega_2\}$

$$P(\omega_1) = \frac{2}{3} \qquad P(\omega_2) = \frac{1}{3}$$

p(x/ω$_j$) ***P(ω$_j$/x)***




$P(\omega)p(x|\omega)$
$P(\omega_1)p(x|\omega_1)$
$P(\omega_2)p(x|\omega_2)$
x_0
x
R_1
R_2
Περιοχή απόφασης της ω_1
Περιοχή απόφασης της ω_2

Σφάλμα ταξινόμησης – Πιθανότητα σφάλματος

- Πότε κάνουμε σφάλμα στην απόφαση;
  - όταν για κάποιο $x \in \omega_1$ ο μηχανισμός αποφασίζει ότι είναι κατηγορίας ω_2 $P(\omega_1)p(x|\omega_1) < P(\omega_2)p(x|\omega_2)$

    περιοχή απόφασης R_2

    ή
  - Όταν για κάποιο $x \in \omega_2$ ο μηχανισμός αποφασίζει ότι είναι κατηγορία $P(\omega_1)p(x|\omega_1) > P(\omega_2)p(x|\omega_2)$

    περιοχή απόφασης R_1

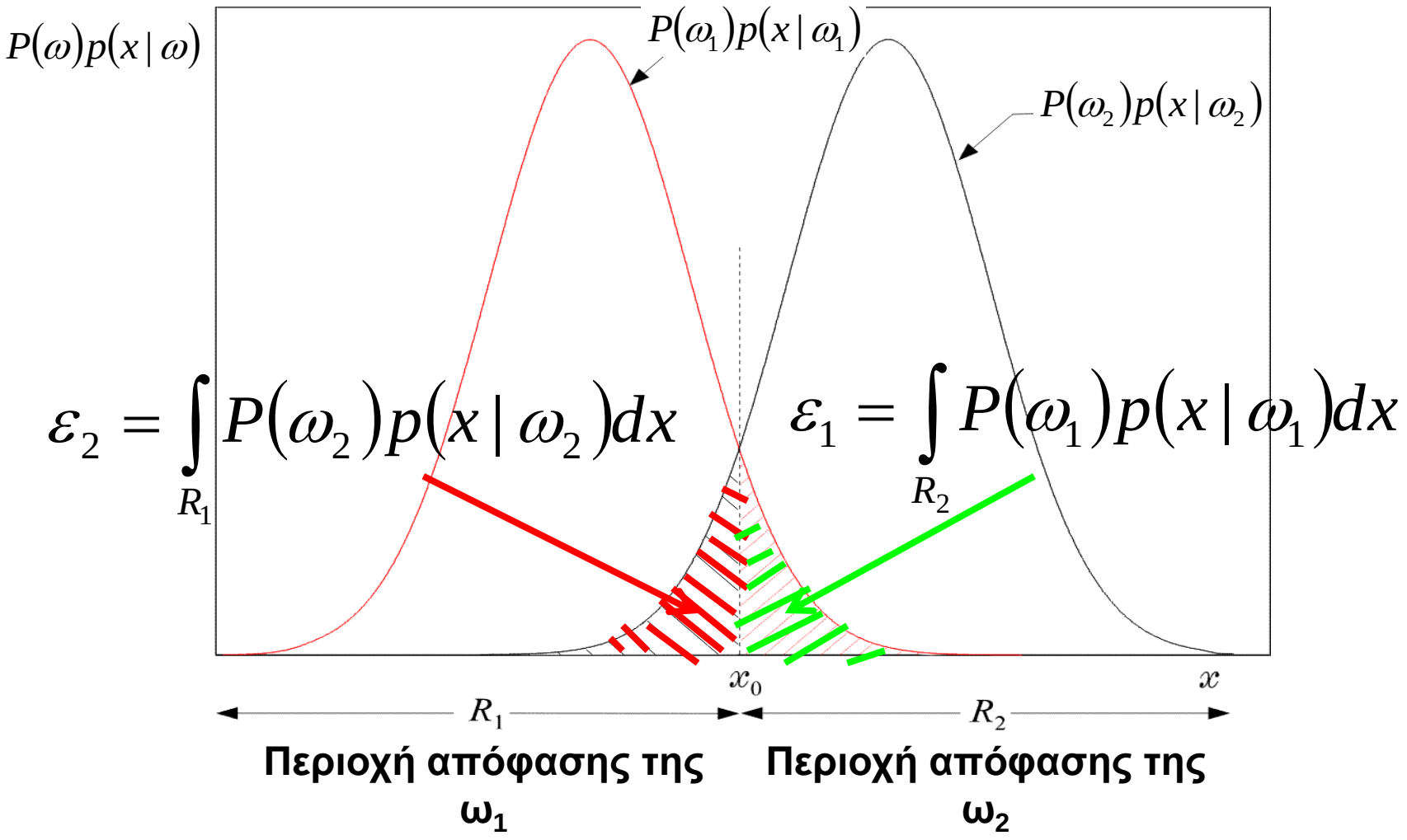
- Πιθανότητα σφάλματος

$$P(error) = P(x \in \omega_1, x \in R_2) + P(x \in \omega_2, x \in R_1) =$$
$$= P(x \in R_2 \mid \omega_1) P(\omega_1) + P(x \in R_1 \mid \omega_2) P(\omega_2) =$$
$$= \left[\int_{R_2} p(x \mid \omega_1) dx \right] P(\omega_1) + \left[\int_{R_1} p(x \mid \omega_2) dx \right] P(\omega_2)$$

έτσι

$$P(error) = \int_{R_2} P(\omega_1) p(x \mid \omega_1) dx + \int_{R_1} P(\omega_2) p(x \mid \omega_2) dx$$

- **Bayes classifier ≡ minimum error classifier**



Ο «αφελής» Ταξινομητής Bayes
Naïve Bayes Classifier

- Βασική υπόθεση **ανεξαρτησίας** μεταξύ των χαρακτηριστικών των προτύπων $x = (x_1, x_2, \ldots, x_d)$

- Έτσι η υπό-συνθήκη πυκνότητα γράφεται ως:

$$p(x \mid \omega_j) = p(x_1, x_2, \ldots, x_d \mid \omega_j) =$$

$$= p(x_1 \mid \omega_j) p(x_2 \mid \omega_j) \cdots p(x_d \mid \omega_j) = \prod_{i=1}^{d} p(x_i \mid \omega_j)$$

- **Πλεονεκτήματα**
  - Ευκολία στην εφαρμογή και χρήση της μεθόδου
    - Μικρή πολυπλοκότητα (αριθμός παραμέτρων)
    - Γρήγορη διαδικασία εκπαίδευσης (fast training)
    - Γρήγορη διαδικασία απόφασης (fast decision)
  - Διευκόλυνση στον χειρισμό χαρακτηριστικών μικτού τύπου (συνεχή, διακριτά)

- **Μειονεκτήματα**
  - Η υπόθεση της ανεξαρτησίας χαρακτηριστικών, αν και υποχρεωτική, είναι δυνατόν να επιφέρει μειωμένη απόδοση στον ταξινομητή

Παράδειγμα εφαρμογής του Naïve Bayes Classifier στο πρόβλημα της ταξινόμησης κειμένων

- Διανυσματική αναπαράσταση κειμένων (bag of words):
  - Υποθέτουμε λεξικό αποτελούμενο από ***d*** όρους.
  - Κάθε κείμενο: $\boldsymbol{x_n = (x_{n1}, x_{n2}, \ldots, x_{nd})}$, όπου $\boldsymbol{x_{ni}}$ η συχνότητα εμφάνισης της **i-**οστής λέξης στο λεξικό.

- Κάθε κατηγορία $\boldsymbol{\omega_j}$ περιγράφεται από ένα διάνυσμα πιθανοτήτων $\boldsymbol{\theta_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jd})}$ δηλ. **πολυωνυμική** κατανομή. Τότε:

$$p(x \mid \omega_j) = p(x \mid \theta_j) = \prod_{i=1}^{d} p(x_i \mid \theta_{ji}) = \prod_{i=1}^{d} \theta_{ji}^{x_i}$$

- Κανόνας απόφασης για K=2 κατηγορίες ($P(\omega_1)=\rho$) :

$$\frac{p(x\mid\omega_1)}{p(x\mid\omega_2)}\underset{>}{<}\frac{P(\omega_2)}{P(\omega_1)}\Rightarrow\frac{\prod_{i=1}^{d}\theta_{1i}^{x_i}}{\prod_{i=1}^{d}\theta_{2i}^{x_i}}\underset{>}{<}\frac{1-p}{p}\Rightarrow\prod_{i=1}^{d}\left(\frac{\theta_{1i}}{\theta_{2i}}\right)^{x_i}\underset{>}{<}\frac{1-\rho}{\rho}$$

ή

$$\sum_{i=1}^{d}x_i\ln\left(\frac{\theta_{1i}}{\theta_{2i}}\right)-\ln\left(\frac{1-\rho}{\rho}\right)\underset{>}{<}0$$

- Ο κανόνας απόφασης: $\sum_{i=1}^{d} x_i \ln\left(\frac{\theta_{1i}}{\theta_{2i}}\right) - \ln\left(\frac{1-\rho}{\rho}\right) \begin{matrix} < \\ > \end{matrix} 0$
- γράφεται ως εξής:

$$\sum_{i=1}^{d} w_i x_i + w_0 \begin{matrix} < \\ > \end{matrix} 0 \Rightarrow \mathbf{w}^T \mathbf{x} + w_0 \begin{matrix} < \\ > \end{matrix} 0$$

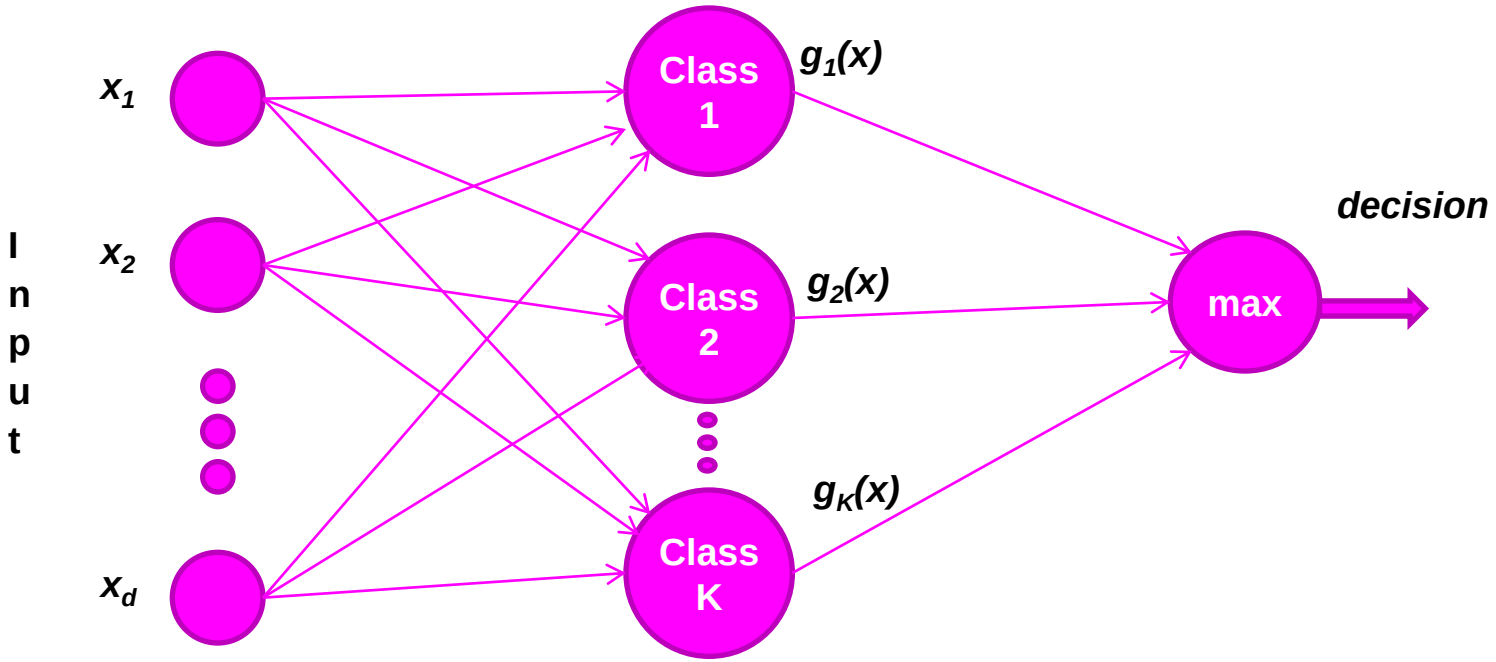
- **γραμμική επιφάνεια διάκρισης** με συντελεστές:

$$w_i = \ln\left(\frac{\theta_{1i}}{\theta_{2i}}\right) \qquad w_0 = -\ln\left(\frac{1-\rho}{\rho}\right) = \ln\left(\frac{\rho}{1-\rho}\right)$$

Διακρίνουσες συναρτήσεις – Επιφάνειες διάκρισης

- Χρήση διακρίνουσων συναρτήσεων (discriminant functions) σε προβλήματα ταξινόμησης, $g_j(x)$ $j=1,\ldots,K$. Τότε, ένα πρότυπο x θα ανήκει στην κατηγορία ω_j αν ισχύει:

$$g_j > g_k \quad \forall k \neq j$$



- Η συνάρτηση διάκρισης στους Μπεϋζιανούς ταξινομητές:

$$g_j(x) = P(\omega_j \mid x)$$

- Μπορούμε να χρησιμοποιήσουμε αντί για *g(x)* μία άλλη συνάρτηση *f(g(x))*, όπου *f()* μία συνάρτηση μονότονη αύξουσα. Το αποτέλεσμα της απόφασης ταξινόμησης δεν μεταβάλλεται:

$$g_j(x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

$$g_j(x) = P(\omega_j \mid x) \quad \Rightarrow \quad g_j(x) = p(x \mid \omega_j)P(\omega_j)$$

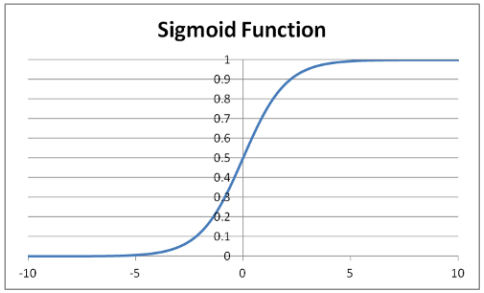
$$g_j(x) = \ln p(x \mid \omega_j) + \ln P(\omega_j)$$

Εναλλακτικές διακρίνουσες συναρτήσεις:

- sigmoid

(ή *logistic*)

$$g_j(x) = \frac{1}{1+e^{-a_j(x)}} = \sigma\left(a_j(x)\right)$$



όπου $$a_j(x) = \ln \frac{p(x \mid \omega_j)P(\omega_j)}{\sum_{k=1}^{K} p(x \mid \omega_k)P(\omega_k)}$$

- Normalized exponential (soft-max)

$$g_j(x) = \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}$$

όπου $$a_j(x) = \ln\, p(x \mid \omega_j)P(\omega_j)$$

Περίπτωση Κανονικής κατανομής

$$x/\omega_j \sim N(\mu_j, \Sigma_j)$$

$$p(x/\omega_j) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_j\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}$$

- Χρησιμοποιώντας την λογαριθμική διακρίνουσα συνάρτηση παίρνουμε :

$$g_j(x) = \ln p(x \mid \omega_j) + \ln P(\omega_j)$$

$$g_j(x) = -\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_j\right| + \ln P(\omega_j)$$

- Διακρίνουσα συνάρτηση:

$$g_j(x) = -\frac{1}{2}\left[x^T\Sigma_j^{-1}x + \mu_j^T\Sigma_j^{-1}\mu_j - 2\mu_j^T\Sigma_j^{-1}x\right] - \frac{1}{2}\ln\left|\Sigma_j\right| + \ln P(\omega_j) =$$

$$= x^T W_j x + w_j^T x + w_{j0}$$

όπου

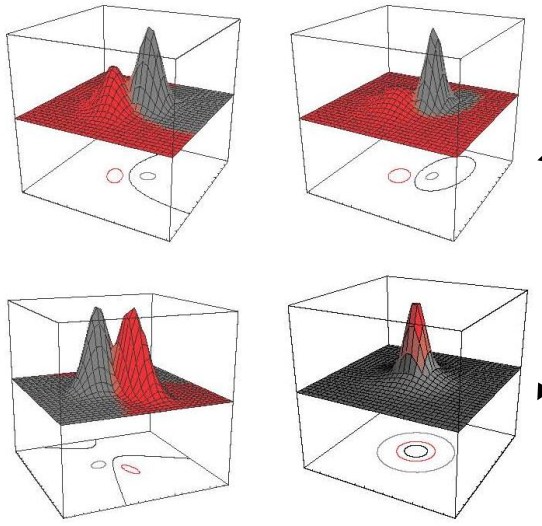
$$W_j = -\frac{1}{2}\Sigma_j^{-1} \qquad w_j = -\Sigma_j^{-1}\mu_j$$

$$w_{j0} = -\frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\ln\left|\Sigma_j\right| + \ln P(\omega_j)$$

- Τα *όρια απόφασης* προκύπτουν από την εξίσωση

$$g_1(x) = g_2(x)$$

- Η *επιφάνεια απόφασης* είναι μη-γραμμική (υπερ-παραβολικές, υπερ-ελλειπτικές, κλπ.)



non-linear decision boundaries

Περίπτωση κοινού πίνακα Σ , $\Sigma_j = \Sigma$

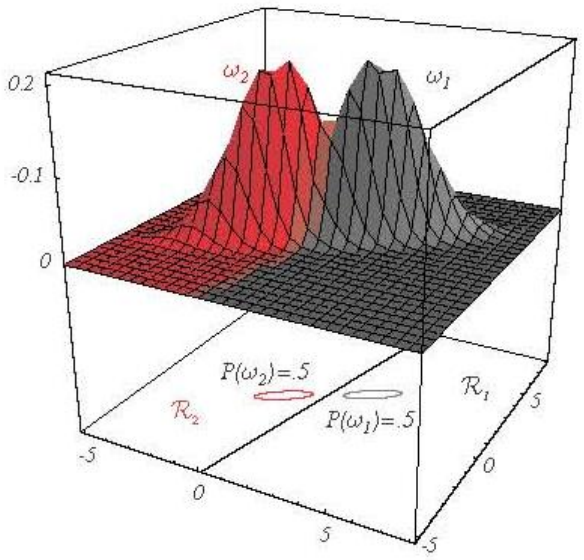
- Διακρίνουσα συνάρτηση :

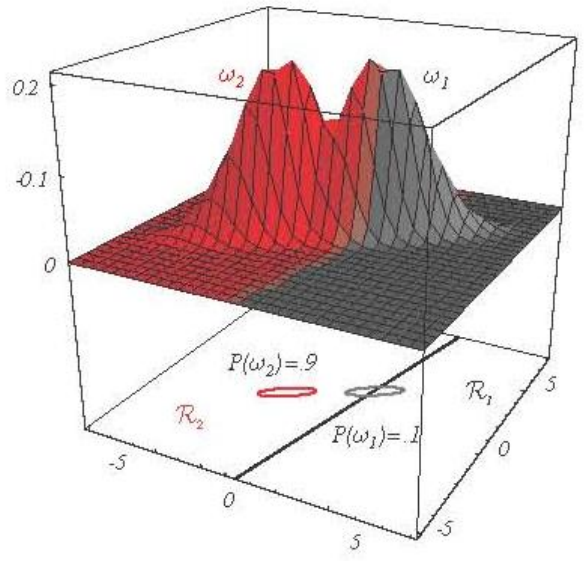
$$g_j(x) = -\frac{1}{2}\left[x^T\Sigma_j^{-1}x + \mu_j^T\Sigma_j^{-1}\mu_j - 2\mu_j^T\Sigma_j^{-1}x\right] - \frac{1}{2}\ln|\Sigma_j| + \ln P(\omega_j)$$

$$g_j(x) = w_j^T x + w_{j0} \qquad \textbf{\textit{γραμμική}}$$

όπου

$$w_j = -\Sigma^{-1}\mu_j \qquad w_{j0} = -\frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \ln P(\omega_j)$$





$$x_0 = \frac{1}{2}(\mu_1 + \mu_2) - \frac{\ln \frac{P(\omega_1)}{P(\omega_2)}}{(\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_2)} (\mu_1 - \mu_2)$$

Αν $P(\omega_i) \neq P(\omega_j)$, τότε το x_0 απομακρύνεται από την πιθανότερη κατηγορία.

- Equivalent to Mahalanobis distance classifier
  - Σε περίπτωση όπου όλες οι κατηγορίες είναι επιπλέον ισοπίθανες (όλες οι πιθανότητες $P(\omega_i)$ *ίσες)* τότε η διακρίνουσα συνάρτηση μετατρέπεται σε:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{i}) + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{i})$$

δηλ. ανάλογη με την Mahalanobis απόσταση

Περίπτωση κοινού "σφαιρικού" πίνακα $\Sigma_j=\Sigma=\sigma^2 I$

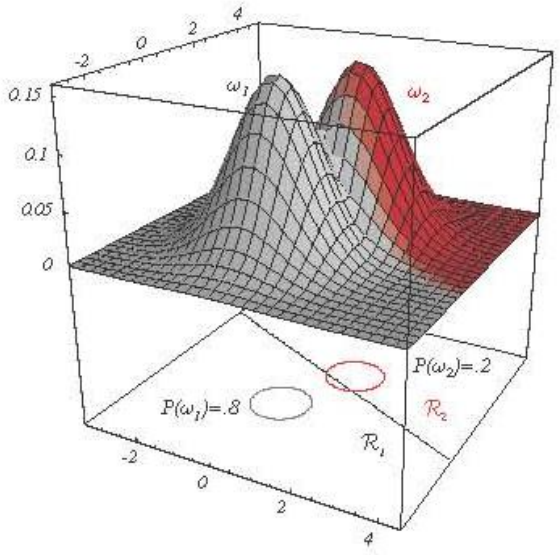
- Διακρίνουσα συνάρτηση :

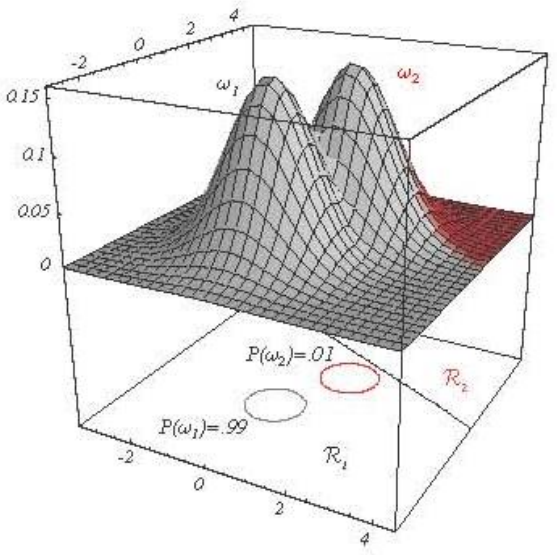
$$g_j(x)=-\frac{1}{2}\left[x^T\Sigma_j^{-1}x+\mu_j^T\Sigma_j^{-1}\mu_j-2\mu_j^T\Sigma_j^{-1}x\right]-\frac{1}{2}\ln|\Sigma_j|+\ln P(\omega_j)$$

$$g_j(x)=w_j^T x+w_{j0} \qquad \textbf{\textit{γραμμική}}$$

όπου

$$w_j=-\frac{1}{\sigma^2}\mu_j \qquad w_{j0}=-\frac{1}{2\sigma^2}\mu_j^T\mu_j+\ln P(\omega_j)$$





$$x_0 = \frac{1}{2}(\mu_1 + \mu_2) - \frac{\sigma^2}{\|\mu_1 - \mu_2\|^2} \ln \frac{P(\omega_1)}{P(\omega_2)} (\mu_1 - \mu_2)$$

Αν $P(\omega_i) \neq P(\omega_j)$, τότε το x_0 απομακρύνεται από την πιθανότερη κατηγορία.

- Equivalent to Minimum Euclidean distance classifier

  – Όταν επιπλέον όλες οι πιθανότητες $P(\omega_i)$ είναι ίσες τότε η διακρίνουσα συνάρτηση μετατρέπεται σε:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i) \quad \Rightarrow \quad g_i(\mathbf{x}) = -\|\mathbf{x}-\mu_i\|^2$$

  Δηλ. απόφαση με βάση την Ευκλείδια απόσταση από το μέσον (κέντρο) της κατηγορίας.