ε.δε.μ^2

ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Δυαδική Ταξινόμηση - Αλγόριθμοι Πυρήνα (Kernel Methods)

1. Διαχωρισιμότητα Προτύπων, Θεώρημα του Cover
2. Radial-Basis Function (RBF) Networks
3. RBF Hybrid Learning
4. Support Vector Machines (SVM)

καθ. Βασίλης Μάγκλαρης

maglaris@netmode.ntua.gr

www.netmode.ntua.gr

Αίθουσα 002, Νέα Κτίρια ΣΗΜΜΥ

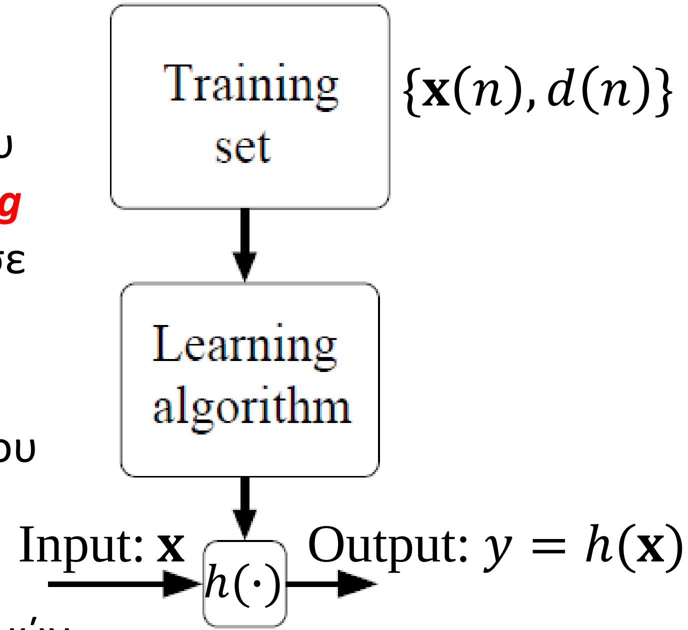
Τρίτη 23/4/2024

Γενικό Μοντέλο Επιβλεπόμενης Μάθησης - Supervised Learning (*επανάληψη*)

Βασισμένο στο Andrew Ng, “CS229 Lecture Notes”, Stanford University, Fall 2018

- Στόχος του συστήματος είναι η αντιστοίχηση ενός δειγματικού στοιχείου εισόδου (***input sample point***, ***example***, ***instance***) $\mathbf{x} = [x_1\ x_2 \dots x_m]^{\mathrm{T}}$ σε τιμές εξόδου y που εκτιμούν επιθυμητές τιμές d (***labels***, ***targets***) π.χ. πρόβλεψη ή ταξινόμηση. Τα στοιχεία x_i είναι αριθμητικές τιμές που κωδικοποιούν m ειδοποιά χαρακτηριστικά (***features***) του δειγματικού στοιχείου $\mathbf{x}$

  Ζητείται ο προσδιορισμός της συνάρτησης εισόδου - εξόδου $y = h(\mathbf{x}) \cong d$ που προκύπτει από δείγμα μάθησης (***Training Set***) N ***labeled*** ζευγών $\{\mathbf{x}(n), d(n)\}$, $n = 1,2, \dots, N$ γνωστών σε εξωτερικό εκπαιδευτή (***supervisor***)



- Η μορφή και οι παράμετροι της $h(\cdot)$ προσδιορίζονται με αλγόριθμο μάθησης που συγκλίνει σε προσέγγιση του στόχου της υπόθεσης για τα N στοιχεία του δείγματος μάθησης

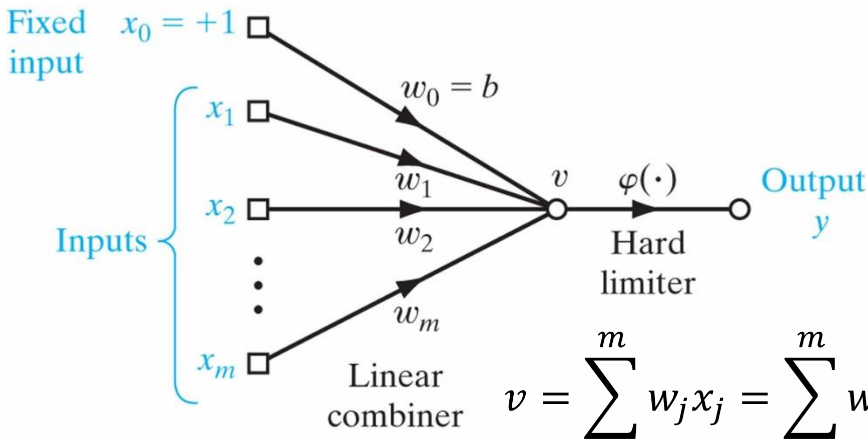
$$d(n) \cong y(n) = h(\mathbf{x}(n))$$

- Αν ο στόχος ικανοποιείται με μικρό αριθμό διακριτών επιλογών (κλάσεων) της y πρόκειται για πρόβλημα Ταξινόμησης, ***Classification*** (για δύο κλάσεις έχουμε δυαδική ταξινόμηση)

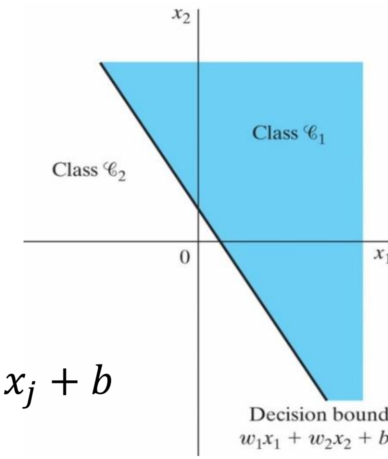
- Αν η έξοδος y λαμβάνει συνεχείς τιμές, το πρόβλημα αναφέρεται σαν Παλινδρόμηση, ***Regression***

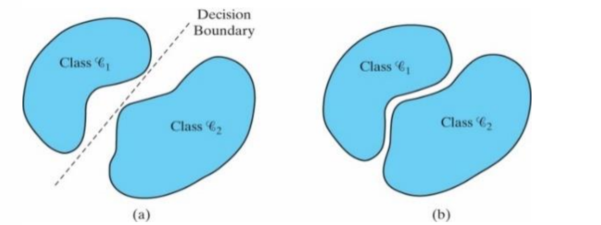
ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Rosenblatt's Perceptron (*επανάληψη*)



$$v = \sum_{j=0}^{m} w_j x_j = \sum_{j=1}^{m} w_j x_j + b$$





Σύνοψη:

1. Ένας νευρώνας με γραμμικό ***induced local field*** v και συνάρτηση ενεργοποίησης $\varphi(v)$ κατωφλίου (***Threshold Function***, ***Hard Limiter***) ή πρόσημου (***Signum Function***) για δυαδική ταξινόμηση στοιχείων $\mathbf{x} = [x_0\ x_1 \dots x_m]^{\mathrm{T}}$ σε δύο ***γραμμικά διαχωριζόμενες*** κλάσεις:

   C_1 αν $y = \varphi(v) = 1$, $\quad C_2$ αν $y = \varphi(v) = 0$ ή αν $y = \varphi(v) = -1$

1. Τα βάρη $\mathbf{w} = [w_0\ w_1 \dots w_m]^{\mathrm{T}}$ ρυθμίζονται on-line (stochastic iterative method) με την εφαρμογή ***Error-correction Algorithm*** σε δειγματικά στοιχεία μάθησης $\{\mathbf{x}(n), d(n)\}$, $n = 1,2,\dots,N$ σε περιβάλλον ***supervised learning*** προς ελαχιστοποίηση σφαλμάτων $[d(n) - y(n)]$

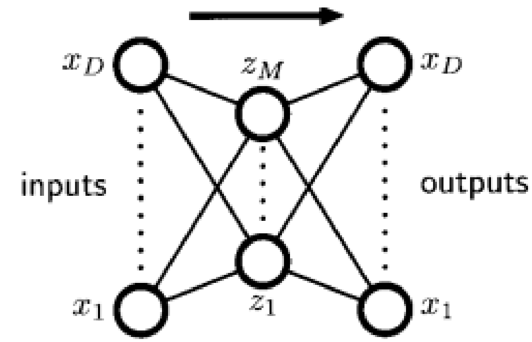
   $$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta[d(n) - y(n)]\mathbf{x}(n)$$

   Η *hyperparameter* η, $0 < \eta \leq 1$ (*learning-rate parameter*) αν είναι *μικρή* οδηγεί την επαναληπτική διαδικασία μάθησης σε σύγκλιση. Αν είναι *μεγάλη* μπορεί να επιταχύνει τη σύγκλιση π.χ. σε περιβάλλοντα με μεγάλες αποκλίσεις των δεδομένων $\mathbf{x}(n)$, αλλά μπορεί να οδηγήσει σε αστάθειες λόγω ταλαντώσεων περί την βέλτιστη τιμή

Σε περιβάλλον δειγματικών στοιχείων $\mathbf{x}$ κατανομής Gauss, η ταξινόμησή τους σε δύο κλάσεις C_1 , C_2 μέσω **Bayes Classifier** (ελαχιστοποίηση ρίσκου σφάλματος με βάση a-priori πιθανότητες p_1, p_2) ταυτίζεται με το **Rosenblatt Perceptron**

Αντιστοίχιση Προτύπων - Pattern Association (S. Haykin: *Introduction, Section 9*)

Διαδικασία *συσχετιστικής μάθησης* (*associative learning*) για αντιστοίχιση *παραδειγμάτων-κλειδιών* $\mathbf{x}_k$ (*key patterns*) σε q *αποθηκευμένα πρότυπα* $\mathbf{y}_k$ (*memorized patterns*)



Μέθοδοι Συσχετιστικής Μάθησης :

➢ **Αυτοαντιστοίχιση** (*Autoassociation*): $\mathbf{x}_k = \mathbf{y}_k$

Τα διανύσματα $\mathbf{x}_k$, $\mathbf{y}_k$ έχουν D διαστάσεις. Με ***Multilayer Perceptron*** (***MLP***) κωδικοποιούμε τα $\mathbf{x}_k$ σε κρυφά (***latent***) διανύσματα $\mathbf{z}_k$ διαστάσεων $M < D$ και σε επόμενο στρώμα (***layer***) αποκωδικοποιούμε κατά προσέγγιση (ελαχίστων τετραγώνων) τα διανύσματα εισόδου (όπως σε ***autoencoders***). Οι παράμετροι του ***MLP*** ρυθμίζονται στη φάση μάθησης με ***Unsupervised Learning*** με δείγμα μάθησης τα q παραδείγματα-κλειδιά $\mathbf{x}_k = \mathbf{y}_k, \quad k = 1,2,\ldots,q$

https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

➢ **Ετεροαντιστοίχηση** (*Heteroassociation*): $\mathbf{x}_k \neq \mathbf{y}_k$ ***Supervised Learning***

Φάσεις Αντιστοίχησης Προτύπων:

- Αποθήκευση (*Storage*): Αποθήκευση στη μνήμη των *key patterns* με κλειδί ανάκτησης τα $\mathbf{x}_k$
- Ανάκληση (*Recall*): Αντιστοίχιση νέου στοιχείου $\mathbf{x}_k$ (*stimulus*, *input vector*, π.χ. χειρόγραφα δεκαδικών αριθμητικά ψηφία ή παραμορφωμένες εικόνες) σε αποθηκευμένο πρότυπο $\mathbf{y}_k$

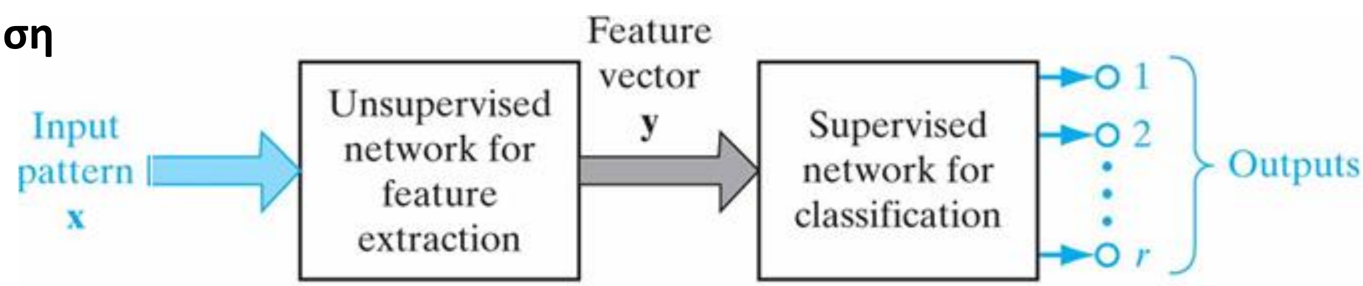
Αναγνώριση Προτύπων - Pattern Recognition (S. Haykin: *Introduction, Section 9*)

Αναγνώριση εισόδου, εξαγωγή κύριων χαρακτηριστικών και ανάθεση – ταξινόμηση νέων προτύπων (*patterns*) σε ορισμένη κατηγορία (*class*) με κριτήριο τη στατιστική συνάφεια με αποθηκευμένα πρότυπα στο σύστημα κατά τη διαδικασία μάθησης

Η διαδικασία συνήθως περιλαμβάνει 2 στάδια:

- **Στάδιο Εξαγωγής Χαρακτηριστικών** (**Feature Extraction**): Μετασχηματισμός εισόδου $\mathbf{x}$ (διάνυσμα διαστάσεως m) σε ενδιάμεσο διάνυσμα $\mathbf{y}$ διαστάσεως $q \leq m$ με *unsupervised learning*. Αν $q < m$ έχουμε συμπίεση δεδομένων ή επιλογή σημαντικών χαρακτηριστικών (*important features*) για απλοποίηση της διαδικασίας ταξινόμησης

- **Στάδιο Ταξινόμησης** (**Classification**): Αντιστοίχηση του ενδιάμεσου πρότυπου $\mathbf{y}$ σε r διακριτές κλάσεις (*supervised learning* σε κρυφά στρώματα). Αν $r = 2$ έχουμε **δυαδική ταξινόμηση**



Παράδειγμα Labeled Δείγματος Μάθησης: *MNIST Database* για ταξινόμηση χειρόγραφων αριθμών σε $r = 10$ κλάσεις $(0 , \ldots , 9)$ https://en.wikipedia.org/wiki/MNIST_database

Διαχωρισιμότητα Προτύπων (Separability of Patterns)

Ταξινόμηση Παραδειγμάτων μέσω Διαχωρίσιμων Προτύπων

Αντιστοίχηση παραδειγμάτων εισόδου $\mathbf{x}$ (examples, instances) σε N πρότυπα (patterns) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ ($\mathbf{x}_i$ διανύσματα με m_0 συνιστώσες) για δυαδική ταξινόμηση σε δύο διαχωρίσιμες κλάσεις C_1 και C_2

Θεώρημα του Cover (1965)

- Περίπλοκο πρόβλημα ταξινόμησης προτύπων μη γραμμικά ορισμένο σε χώρο πολλών διαστάσεων, είναι πιθανότερο να είναι γραμμικά διαχωρίσιμο (linearly separable) από ότι σε χώρο λίγων διαστάσεων, αν δεν υπάρχουν πυκνά σημεία (πρότυπα)
- Για εύκολη ταξινόμηση προτύπων προτιμάται η *γραμμική διαχωρισιμότητα* μέσω *μη γραμμικού μετασχηματισμού συντεταγμένων* και ας απαιτούνται *περισσότερες διαστάσεις*

Κρυφές Συναρτήσεις (Hidden Functions)

Τα $\mathbf{x}$ μετασχηματίζονται (μη γραμμικά) σε $\boldsymbol{\varphi}(\mathbf{x})$ με $m_1 \geq m_0$ διαστάσεις

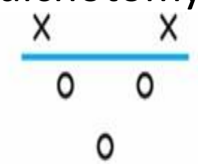
$$\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x}) = \left[\varphi_1(\mathbf{x}) \;\; \varphi_2(\mathbf{x}) \dots \varphi_{m_1}(\mathbf{x})\right]^{\mathrm{T}}$$

Οι $\varphi_j(\mathbf{x}) \in \mathbb{R}, j = 1,2,\dots,m_1$ *είίναι κρυφές συναρτήσεις*

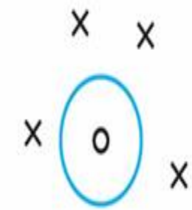
Ο χώρος των προτύπων είναι *γραμμικά διαχωρίσιμος* κατά $\boldsymbol{\varphi}$ ($\boldsymbol{\varphi}$-separable dichotomy) όταν υπάρχει διάνυσμα $\mathbf{w}$ με $m_1 \geq m_0$ συνιστώσες που να ορίζει δύο γραμμικά διακριτές περιοχές αντίστοιχες με τις C_1 και C_2 των $\mathbf{x}$:

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}) > 0 \Rightarrow \mathbf{x} \in C_1 \text{ και } \mathbf{w}^{T}\boldsymbol{\varphi}(\mathbf{x}) < 0 \Rightarrow \mathbf{x} \in C_2$$

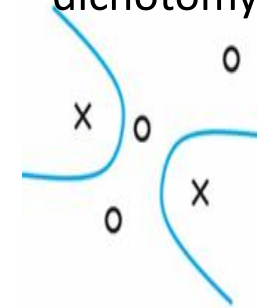
(a) Linearly separable dichotomy



(b) Spherically separable dichotomy



(c) Quadrically separable dichotomy



ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

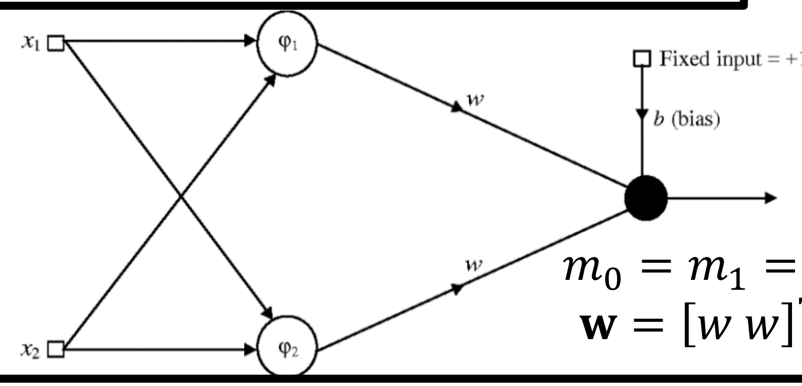
Διαχωρισιμότητα Προτύπων – Το πρόβλημα XOR

Συνήθης επιλογή για $\varphi_j(\mathbf{x})$: ***Gaussian Radial-Basis Function*** (***RBF***)

$\varphi_j(\mathbf{x}) = \exp\left(-\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|^2\right)$ όπου $\boldsymbol{\mu}_j$ διάνυσμα διαστάσεως m_0 των *μέσων τιμών* (κέντρων) της $\varphi_j(\mathbf{x})$ και $\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|$ η *Ευκλείδεια απόσταση* του σημείου $\mathbf{x}$ από το $\boldsymbol{\mu}_j$

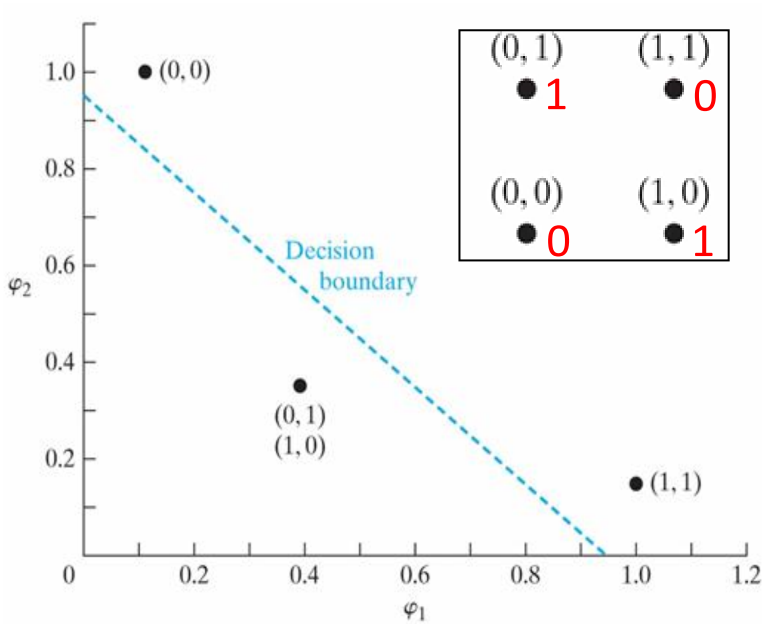
TABLE 5.1 Specification of the Hidden Functions for the XOR Problem of Example 1

Input Pattern $\mathbf{x}$	First Hidden Function $\varphi_1(\mathbf{x})$	Second Hidden Function $\varphi_2(\mathbf{x})$
(1,1)	1	0.1353
(0,1)	0.3678	0.3678
(0,0)	0.1353	1
(1,0)	0.3678	0.3678



$m_0 = m_1 = 2$
$\mathbf{w} = [w\ w]^{\mathrm{T}}$

$$\mathbf{x} = [x_1\ x_2]^{\mathrm{T}} \rightarrow \boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x})\ \varphi_2(\mathbf{x})]^{\mathrm{T}}$$
$$\varphi_1(\mathbf{x}) = \exp(-\|\mathbf{x} - \boldsymbol{\mu}_1\|^2),\ \boldsymbol{\mu}_1 = [1{,}1]^{\mathrm{T}}$$
$$\varphi_2(\mathbf{x}) = \exp(-\|\mathbf{x} - \boldsymbol{\mu}_2\|^2),\ \boldsymbol{\mu}_2 = [0{,}0]^{\mathrm{T}}$$



Παράδειγμα Υπολογισμού $\boldsymbol{\varphi}(\mathbf{x})$, $\mathbf{x} = [1\ 1]^{\mathrm{T}}$

$$\varphi_1(1{,}1) = \exp\left(-\left\|[1\ 1]^{\mathrm{T}} - [1\ 1]^{\mathrm{T}}\right\|^2\right) = 1$$

$$\varphi_2(1{,}1) = \exp\left(-\left\|[1\ 1]^{\mathrm{T}} - [0\ 0]^{\mathrm{T}}\right\|^2\right) = 0.1353$$

Έξοδος: $y = w\varphi_1(\mathbf{x}) + w\varphi_2(\mathbf{x}) + b$

$(1{,}1)$: $w + w \times 0.1353 + b = 0$

$(0{,}1)$: $w \times 0.3678 + w \times 0.3678 + b = 1$

$(0{,}0)$: $w \times 0.1353 + w + b = 0$

$(1{,}0)$: $w \times 0.3678 + w \times 0.3678 + b = 1$

Λύση
$w = -2.502$
$b = 2.841$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Ορισμοί Radial-Basis Function (RBF), Kernels, Hybrid Learning

(βασισμένο στο ***C. M. Bishop***, **Ch.6: Kernel Methods** https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf)

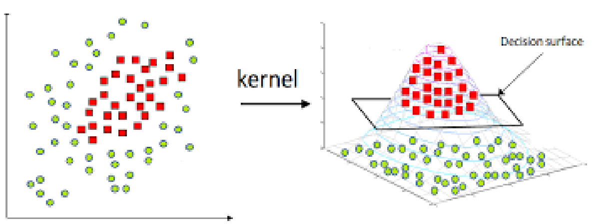
Radial-BasIs Function (RBF): $\mathbf{x} \in \mathbb{R}^{m_0} \to \varphi_j(\mathbf{x}) = \varphi(\|\mathbf{x}-\mathbf{x}_j\|) = \varphi(r) \in \mathbb{R}$

$r = \|\mathbf{x}-\mathbf{x}_j\|$ μέτρο ακτινικής απόστασης διανυσμάτων $\mathbf{x}$ και $\mathbf{x}_j$ (συνήθως *Ευκλείδεια*)

- **Μετασχηματισμός**: $\mathbf{x} \to \boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x})\ \varphi_2(\mathbf{x}) \ldots \varphi_{m_1}(\mathbf{x})]^{\mathrm{T}}, m_1 \geq m_0$
- **Παράδειγμα**: *Gaussian RBF* $\varphi_j(\mathbf{x}) = \exp\left(-\|\mathbf{x}-\mathbf{x}_j\|^2\right)$ που αφορά στην Ευκλείδεια απόσταση δειγματικού σημείου $\mathbf{x}$ με m_0 χαρακτηριστικά (*features*) από τα N δειγματικά σημεία μάθησης (πρότυπα, *patterns*) $\mathbf{x}_j$
- **Ταξινόμηση Προτύπων**: Τα $\varphi_j(\mathbf{x})$ απεικονίζουν N κρυφά χαρακτηριστικά (*hidden features*) του $\mathbf{x}$ σαν αποστάσεις από τα κέντρα $\mathbf{x}_j$ τα οποία προκύπτουν από το δείγμα μάθησης (*patterns*) σε *1η Φάση Υβριδικής Μάθησης* (***Hybrid Learning***) για ταξινόμηση προτύπων γύρω από κέντρα βάρους μέσω μη επιβλεπόμενης μάθησης (π.χ. με αλγόριθμο *K-Means*)

Kernel: $k(\mathbf{x},\mathbf{x}_j) \in \mathbb{R}$ μέτρο ομοιότητας (~εσωτερικό γινόμενο) του διανύσματος $\mathbf{x} \in \mathbb{R}^{m_0}$ με διάνυσμα $\mathbf{x}_j \in \mathbb{R}^{m_0}$

- **Σχέση με RBF**: $k(\mathbf{x},\mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_j), j = 1,2,\ldots,m_1$ (εσωτερικό γινόμενο). Λόγω του *Θεωρήματος του Cover* συνήθως $m_1 > m_0$ για μετασχηματισμό $\mathbf{x} \to \boldsymbol{\varphi}(\mathbf{x})$ σε *linearly separable* περιοχές ταξινόμησης του προτύπου $\mathbf{x}$



Kernel Trick

- **Ταξινόμηση Προτύπων**: Τελική επιλογή κλάσης για δειγματικό σημείο (*pattern*) $\mathbf{x} \in \mathbb{R}^{m_0}$ σε *2η Φάση Υβριδικής Μάθησης* (***Hybrid Learning***) στο πλησιέστερο σημείο από τα κέντρα βάρους της 1ης Φάσης, μέσω επιβλεπόμενης μάθησης και με δίκτυο ***feed-forward***

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Δυαδική Ταξινόμηση - Kernel Perceptron

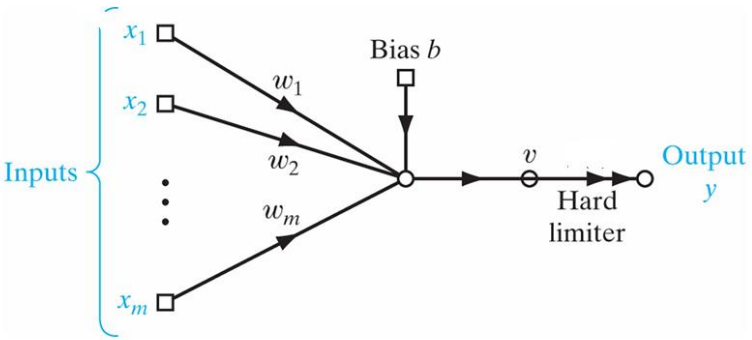
https://en.wikipedia.org/wiki/Kernel_perceptron

Γραμμική Δυαδική Ταξινόμηση, Αλγόριθμος Perceptron

$$y = \mathrm{sgn}(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \in \{-1,1\}$$

Αλγόριθμος Μάθησης: Επαναληπτικός προσδιορισμός συναπτικών βαρών με αρχικοποίηση $\mathbf{w} = [0\ 0 \dots 0]^{\mathrm{T}}$, διαδοχικές εισόδους *labeled* στοιχείων μάθησης $\{\mathbf{x}_i, d_i\}$, $d_{\mathrm{i}} \in \{-1,1\}$, $i = 1,2,\dots,N$ και εξόδους $y = \mathrm{sgn}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)$:

$$\mathbf{w} \leftarrow \begin{cases} \mathbf{w} & \text{αν } y = d_i \quad \text{(ορθή επιλογή)} \\ \mathbf{w} + d_i\mathbf{x}_i & \text{αν } y \neq d_i \quad \text{(λάθος επιλογή)} \end{cases}$$



Μη Γραμμική Δυαδική Ταξινόμηση, Kernel Perceptron

Η *Kernel Machine* αποθηκεύει N σημεία $\mathbf{x}_i$ διαστάσεως m_0 του δείγματος μάθησης $\{\mathbf{x}_i, d_i\}$, ορίζει μετρητή α_i για ταξινομήσεις $\mathbf{x}_i \to y \in \{-1,1\}$ και επιλέγει κλάση y με βάση τον κανόνα

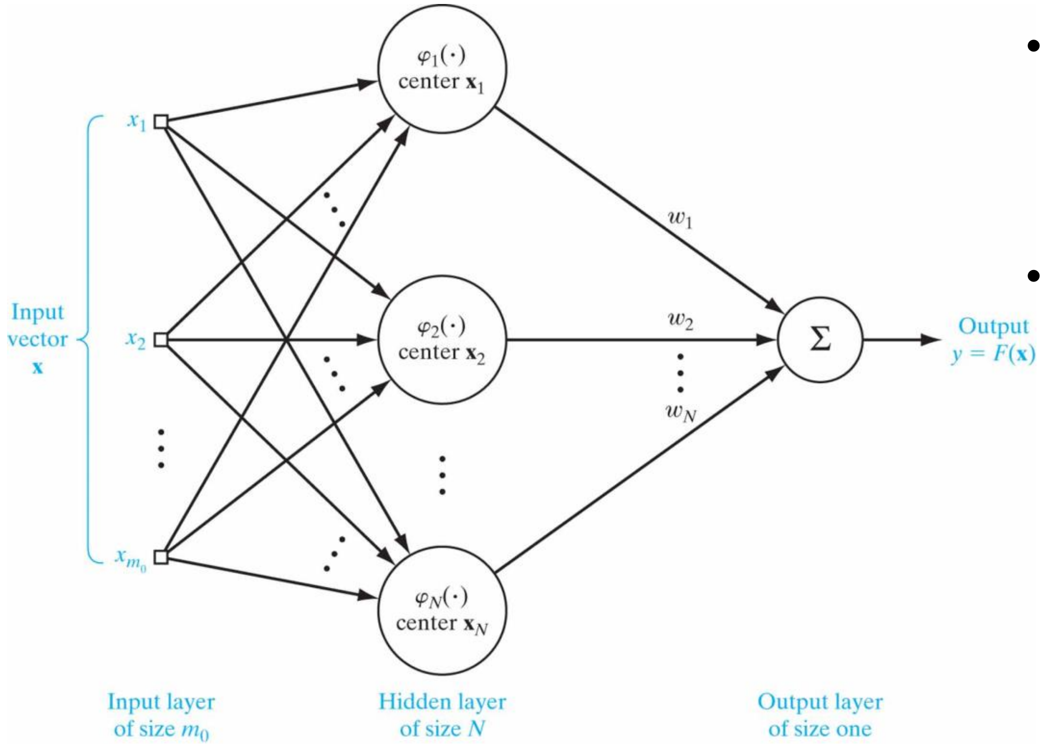
$$y = \mathrm{sgn}\sum_{i=1}^{N} \alpha_i d_i k(\mathbf{x}, \mathbf{x}_i) \in \{-1,1\}$$

Ο πυρήνας (*Kernel*) $k(\mathbf{x}, \mathbf{x}_i) \in \mathbb{R}$ ορίζεται σαν εσωτερικό γινόμενο διανυσμάτων διαστάσεως $m_1 \geq m_0$ με στοιχεία μη γραμμικές *hidden functions*: $k(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}_i)$

Αλγόριθμος Μάθησης: Ανάλογη του γραμμικού Αλγορίθμου Perceptron με $\mathbf{w} = \sum_{i=1}^{N} \alpha_i\, d_i, \mathbf{x}_i$ όπου α_i μετρητής λανθασμένων επιλογών $\mathbf{x}_i \to y \neq d_i$

Για κάθε δειγματικό σημείο μάθησης - *labeled pattern* $\{\mathbf{x}_i, d_i\}, i = 1,2,\dots,N$ υπολογίζω $y = \mathrm{sgn}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) = \mathrm{sgn}\sum_{j=1}^{N} \alpha_j d_j k(\mathbf{x}_i, \mathbf{x}_j)$. Αν $y \neq d_i$ αυξάνεται ο μετρητής $\alpha_i \leftarrow \alpha_i + 1$

Radial-Basis Function (RBF) Networks



- **Input Layer**: Είσοδος διανυσμάτων $\mathbf{x}$ με m_0 χαρακτηριστικά (*features*) που τροφοδοτούν χωρίς τροποποίηση ενδιάμεσο κρυφό επίπεδο
- **Hidden Layer**: Για *Δείγμα Μάθησης* με $N \geq m_0$ στοιχεία (*πρότυπα*), ορίζονται N κόμβοι επεξεργασίας με *Gaussian Ακτινικές Συναρτήσεις Βάσης* - *RBF*:

$$\varphi_j(\mathbf{x}) = \varphi(\mathbf{x}, \mathbf{x}_j) = \varphi(\|\mathbf{x} - \mathbf{x}_j\|)$$
$$= \exp(-\|\mathbf{x} - \mathbf{x}_j\|^2)$$

(Gaussian συνάρτηση της απόστασης $\|\mathbf{x} - \mathbf{x}_j\|^2$ του $\mathbf{x}$ από τα παραδείγματα μάθησης $\mathbf{x}_j$)

- **Output Layer**: Γραμμικός συνδυασμός συναρτήσεων βάσης $\boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x})\ \varphi_2(\mathbf{x}) \ldots \varphi_N(\mathbf{x})]^{\mathrm{T}}$

$$y = F(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=1}^{N} w_j \varphi(\|\mathbf{x} - \mathbf{x}_j\|), y \in \{-1,1\}$$

- **Training**:
  - ✓ Επίλυση γραμμικού συστήματος N εξισώσεων $F(\mathbf{x}_i) = \sum_j w_j\, \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|) = d_i$ από τα N *labeled* στοιχεία $\{\mathbf{x}_i, d_i\}$ του δείγματος μάθησης με N αγνώστους w_j
  - ✓ Οι $F(\mathbf{x}_i) = d_i$ ορίζουν στο δείγμα μάθησης *υπερεπιφάνεια δυαδικού διαχωρισμού* κλάσεων
  - ✓ Το σύστημα έχει πάντα λύση για διακριτά σημεία $\mathbf{x}_i$ (*Θεώρημα Michellis*)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Radial-Basis Function (RBF) Network για Ταξινομητή XOR

(βασισμένο στη παρουσίαση «**Υβριδική Μάθηση – RBF**», ***Α. Σταφυλοπάτη***, ΣΗΜΜΥ Ε.Μ.Π.
http://mycourses.ntua.gr/courses/ECE1080/document/%C4%E9%E1%EB%DD%EE%E5%E9%F2_2019-2020/rbf.pdf)

Δείγμα Μάθησης: $N = 4$ *labeled* στοιχεία (*πρότυπα*) $\mathbf{x}_i = [x_i(1)\ x_i(2)]^{\mathrm{T}} \rightarrow y = F(\mathbf{x}_i)$
όπου $\{x_i(1), x_i(2), y\}$ δυαδικές μεταβλητές $\in \{0,1\}$ και $i \leq N = 4$

Ακτινικές Συναρτήσεις Βάσης: $\varphi_j(\mathbf{x}) = \exp(-\|\mathbf{x} - \boldsymbol{\mu}_j\|^2)$, $j = 1,2,3,4$
γύρω από 4 κέντρα βάρους: $\boldsymbol{\mu}_1 = [1\ 1]$, $\boldsymbol{\mu}_2 = [0\ 0]$, $\boldsymbol{\mu}_3 = [0\ 1]$, $\boldsymbol{\mu}_4 = [1\ 0]$

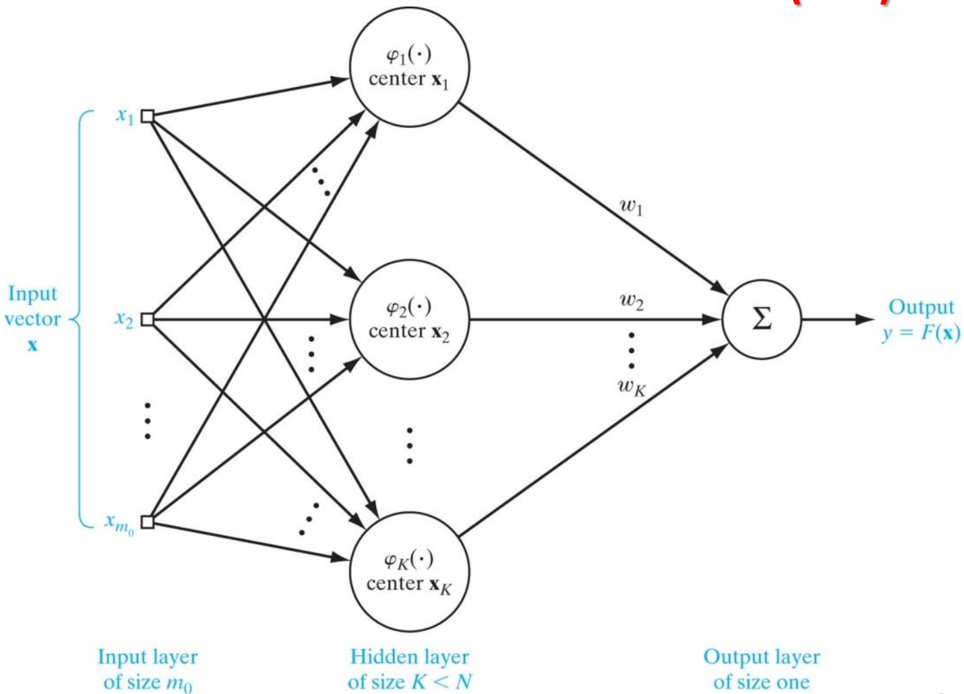
$$y = F(\mathbf{x}) = w_1\varphi_1(\mathbf{x}) + w_2\varphi_2(\mathbf{x}) + w_3\varphi_3(\mathbf{x}) + w_4\varphi_4(\mathbf{x})$$

$\mathbf{x}$	$\varphi_1(\mathbf{x})$	$\varphi_2(\mathbf{x})$	$\varphi_3(\mathbf{x})$	$\varphi_4(\mathbf{x})$	y
(1,1)	1	0.1353	0.3678	0.3678	0
(0,0)	0.1353	1	0.3678	0.3678	0
(0,1)	0.3678	0.3678	1	0.1353	1
(1,0)	0.3678	0.3678	0.1353	1	1

Ρύθμιση Παραμέτρων Δικτύου

$$w_1 = w_2 = -0.9843$$
$$w_3 = w_4 = 1.5188$$

Radial-Basis Function (RBF) Networks – Πρακτική Υλοποίηση



Τα δίκτυα RBF χαρακτηρίζονται από γρήγορη φάση μάθησης **αλλά** απαιτούν αποθήκευση μεγάλου αριθμού κρυφών κόμβων N (ίσων με τον αριθμό των στοιχείων μάθησης), ακριβείς μετρήσεις των $\{\mathbf{x}_i, d_i\}$ και παρουσιάζουν μεγάλο υπολογιστικό φόρτο στη φάση test

Προσεγγιστική Υλοποίηση

Μικρότερος αριθμός κρυφών κόμβων $K < N$ που ορίζει χώρο K διαστάσεων:

$$y = F(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=1}^{K} w_j \varphi\left(\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|\right)$$

Υβριδική Μάθηση

Προσδιορισμός των $K < N$ *Κέντρων* $\mathbf{x}_j$ και των *Συναπτικών Βαρών* w_j, $j = 1,2, \dots, K$

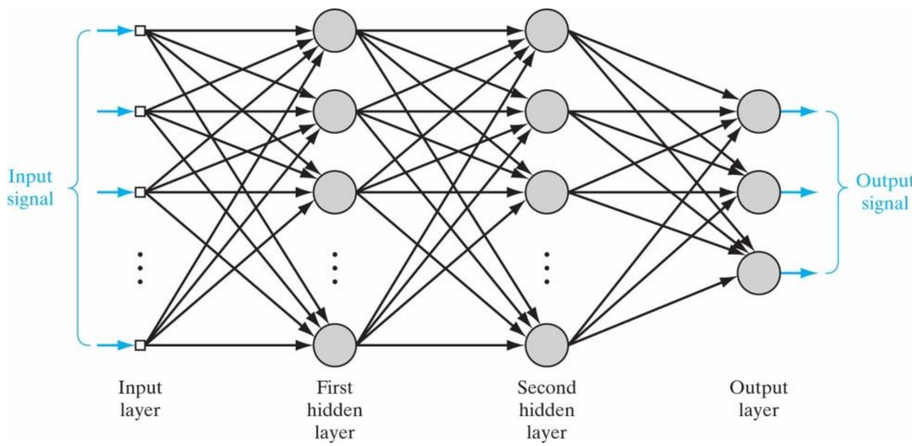
- **Input Layer**: Διάνυσμα $\mathbf{x}$ διαστάσεως m_0 (αριθμός *features*)
- **Hidden Layer**: K κρυφοί κόμβοι $\varphi\left(\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|\right)$ με κέντρα $\boldsymbol{\mu}_j$ που προκύπτουν σαν *cluster heads* των $\mathbf{x}$ από αλγόριθμο *μη επιβλεπόμενης μάθησης K-Means Clustering* με Ευκλείδειο μέτρο απόστασης $\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|^2$ (K *υπερπαράμετρος*, $m_0 \leq K < N$)
- **Output Layer**: Γραμμικός συνδυασμός των K συναρτήσεων βάσης $\varphi_j(\mathbf{x})$:

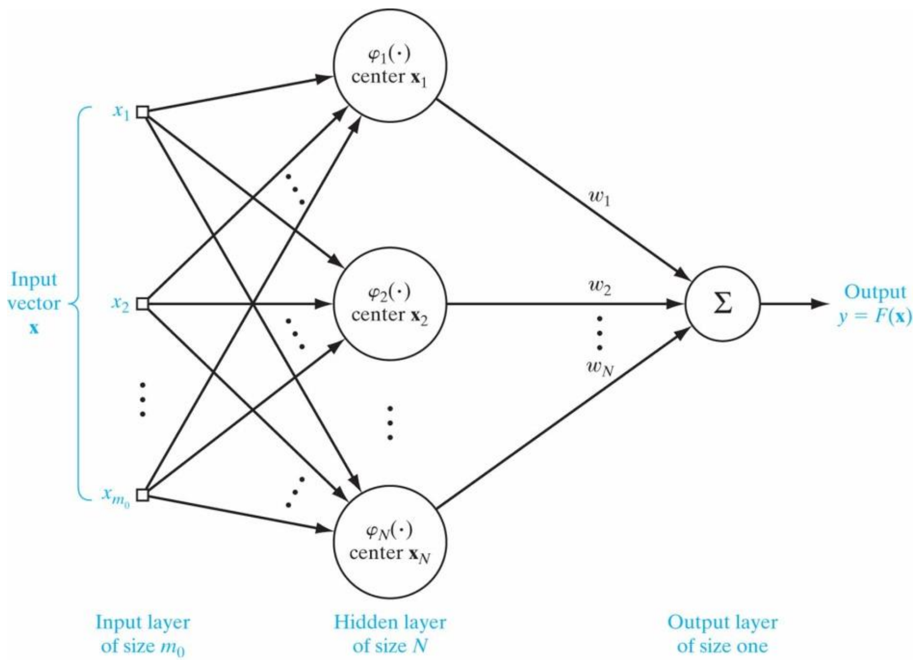
$$y = F(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}) = \textstyle\sum_{j=1}^{K} w_j \varphi\left(\left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|\right)$$

Εκτίμηση των w_j από στοιχεία του δείγματος μάθησης $\{\boldsymbol{\varphi}(\mathbf{x}_i), d_i\}$ με *επιβλεπόμενη μάθηση* κατά προσέγγιση *ελαχίστων τετραγώνων*: N εξισώσεις $d_i \cong F(\mathbf{x}_i)$, K άγνωστοι w_j $(K < N)$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Multi-layer Perceptron (MLP) vs. RBF





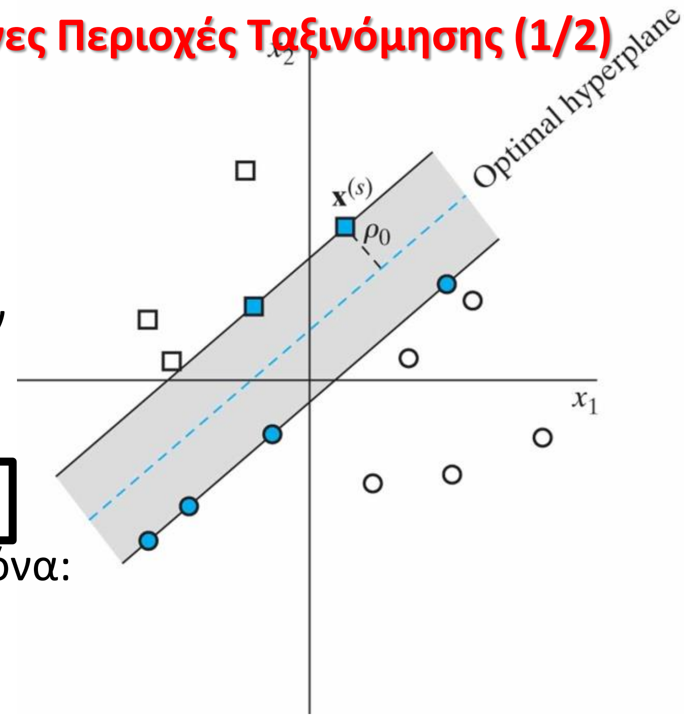
MLP

- Πολλά επίπεδα κρυφών νευρώνων
- Επιβλεπόμενη Μάθηση
- Batch ή On-line (Stochastic) Learning
- Back-propagation Algorithm
- Μη γραμμική συνάρτηση ενεργοποίησης
- Βραδεία εκπαίδευση
- Ανοχή σε ανακρίβειες μετρήσεων εισόδου

RBF

- Ένα κρυφό επίπεδο νευρώνων
- Υβριδική Μάθηση (Hybrid Learning)
- Μη γραμμικός μετασχηματισμός διανυσματικών σημείων μέσω Radial-Basis Functions (Gaussian)
- Ευελιξία στη διαχωρισιμότητα περιοχών κατάταξης διανυσματικών σημείων (pattern vectors)
- Γρήγορη εκπαίδευση
- Ευαισθησία σε ανακρίβειες μετρήσεων δειγματικών σημείων
- Η υπερ-επιφάνεια δυαδικού διαχωρισμού γενικεύεται για ανακριβή ή νέα δειγματικά σημεία με interpolation

Support Vector Machines (SVM) – Γραμμικά Διαχωριζόμενες Περιοχές Ταξινόμησης (1/2)



- Για *labeled* δείγμα μάθησης $\{\mathbf{x}_i, d_i\}$, $d_i \in \{-1, +1\}$, $i = 1,2,\dots,N$ η **SVM** ορίζει βέλτιστες περιοχές δυαδικής ταξινόμησης με τη μέγιστη διαχωριστική ζώνη (περιθώριο διαχωρισμού - *margin of separation*) μεταξύ τους
- Για περιπτώσεις γραμμικά διαχωριζόμενων διανυσματικών στοιχείων $\mathbf{x}$ (*patterns*) με m διαστάσεις (*features*) το υπερ-επίπεδο διαχωρισμού ορίζεται από την εξίσωση

$$\boxed{\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0 \text{ όπου } \mathbf{w} \text{ συναπτικά βάρη και } b \text{ bias}}$$

- Η ταξινόμηση του σημείου μάθησης $\mathbf{x}_i$ ακολουθεί τον κανόνα:

$$\boxed{\begin{array}{l} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \geq 0 \text{ αν } d_i = +1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b < 0 \text{ αν } d_i = -1 \end{array}}$$

- Η απόσταση του πλησιέστερου σημείου από το υπερεπίπεδο διαχωρισμού ορίζει το περιθώριο ρ που πρέπει να μεγιστοποιηθεί για *βέλτιστο διαχωρισμό*: $\mathbf{w}_o^{\mathrm{T}}\mathbf{x} + b_o = 0$
- Γεωμετρικά προκύπτει πως $\rho = \frac{2}{\|\mathbf{w}_o\|}$ όπου $\|\mathbf{w}_o\|$ το Ευκλείδειο μέτρο του διανύσματος $\mathbf{w}_o$
- Για τα στοιχεία του δείγματος μάθησης $\{\mathbf{x}_i, d_i\}$ για κανονικό υπερεπίπεδο διαχωρισμού:

$$\mathbf{w}_o^{\mathrm{T}}\mathbf{x}_i + b_o \geq 1 \text{ αν } d_i = +1$$
$$\mathbf{w}_o^{\mathrm{T}}\mathbf{x}_i + b_o \leq 1 \text{ αν } d_i = -1$$

- Τα διανύσματα $\mathbf{x}_i$ για τα οποία ισχύει η ισότητα σε μια από τις δύο ανισότητες είναι τα ***Support Vectors*** (*Διανύσματα Υποστήριξης*) $\mathbf{x}_i^S$ στα όρια της διαχωριστικής ζώνης
- Οι ανισότητες ενοποιούνται σαν περιορισμοί (*constraints*) για το δείγμα μάθησης:

$$d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, i = 1,2,\dots,N$$

Support Vector Machines (SVM) – Γραμμικά Διαχωριζόμενες Περιοχές Ταξινόμησης (2/2)

(βασισμένο στη παρουσίαση «**Μηχανές Διανυσμάτων Υποστήριξης**», ***Γ. Στάμου***, ΣΗΜΜΥ Ε.Μ.Π.
http://mycourses.ntua.gr/courses/ECE1078/document/%D5%EB%E9%EA%FC_%C4%E9%E1%EB%DD%EE%E5%F9%ED_2019-2020/NN-SVM-handouts.pdf)

Διατύπωση σαν Πρόβλημα Μη Γραμμικού Προγραμματισμού

Μεγιστοποίηση περιθωρίου διαχωρισμού $\rho = \frac{2}{\|\mathbf{w}_o\|} \Leftrightarrow$ Ελαχιστοποίηση $\|\mathbf{w}_o\|^2 = \mathbf{w}_o^{\mathrm{T}}\mathbf{w}_o$

Πρόβλημα βελτιστοποίησης με περιορισμούς για προσδιορισμό των παραμέτρων της SVM (synaptic weights $\mathbf{w}$ και bias b):

$$\min_{\mathbf{w}} \Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \quad \text{όταν} \quad d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, i = 1,2, \dots, N$$

Η συνάρτηση κόστους είναι άθροισμα τετραγώνων και οι περιορισμοί γραμμικοί. Το βέλτιστο $\mathbf{w}$ μπορεί να προσδιορισθεί με κλασσική μέθοδο μη γραμμικού προγραμματισμού, π.χ. με χρήση ***Lagrange Multipliers*** λ_i για τους περιορισμούς $d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1$

Ορίζω συνάρτηση κόστους $J(\mathbf{w}, b, \lambda_1, \lambda_2, \dots, \lambda_N) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{N} \lambda_i [\, d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)]$

Στο βέλτιστο σημείο και για τα N στοιχεία μάθησης $\mathbf{x}_i$ ισχύουν οι συνθήκες ***Kuhn-Tucker***:

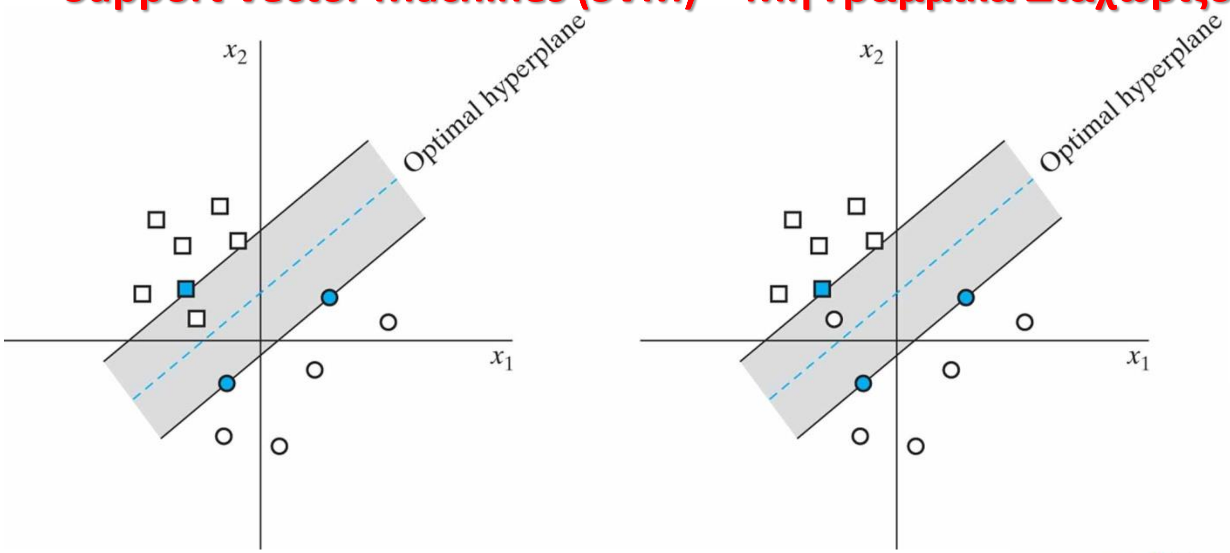
$$\frac{\partial J}{\partial \mathbf{w}} = 0 \;\rightarrow\; \mathbf{w} = \sum_{i=1}^{N} \lambda_i d_i \mathbf{x}_i$$

$$\frac{\partial J}{\partial b} = 0 \;\rightarrow\; \sum_{i=1}^{N} \lambda_i d_i = 0$$

- Τα $\mathbf{w}, b$ προσδιορίζουν το βέλτιστο υπερεπίπεδο διαχωρισμού $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$
- Τα ***Support Vectors*** $\mathbf{x}_i^S$ αντιστοιχούν σε $\lambda_i > 0$. Τα υπόλοιπα $\mathbf{x}_i$ σε $\lambda_i = 0$

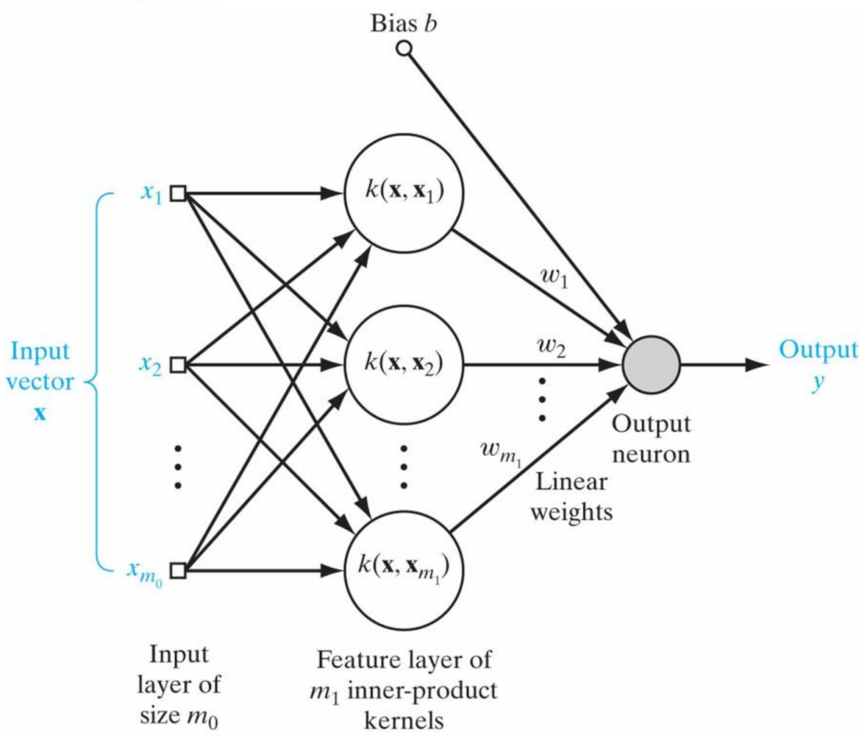
ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Support Vector Machines (SVM) – Μη Γραμμικά Διαχωριζόμενες Περιοχές Ταξινόμησης



Παραβάσεις Γραμμικής Διαχωρισιμότητας:

- $\{\mathbf{x}_i, d_i\}$ εντός της διαχωριστικής ζώνης από την σωστή πλευρά του βέλτιστου υπερεπιπέδου
- $\{\mathbf{x}_i, d_i\}$ εντός της διαχωριστικής ζώνης από την λάθος πλευρά του βέλτιστου υπερεπιπέδου



Αρχιτεκτονική SVM με χρήση Δικτύου RBF

Χρήση μεγάλου αριθμού *hidden nodes* m_1 (μικρότερο ή ίσο από τον αριθμό στοιχείων του δείγματος μάθησης N) που μετασχηματίζουν μη γραμμικά διαχωρίσιμες περιοχές των διανυσμάτων εισόδου $\mathbf{x}$ διαστάσεως $m_0 \ll m_1$ σε γραμμικά διαχωρίσιμες περιοχές