Support Vector Machines (SVM)

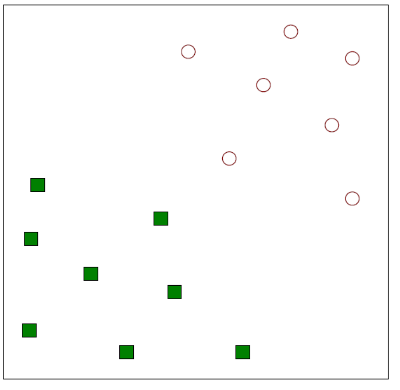
Support Vector Machines

- Οι **μηχανές διανυσμάτων υποστήριξης** είναι μία μέθοδος λήψης απόφασης που συνδυάζει τη θεωρία υπολογιστικής μάθησης με τη θεωρία βελτιστοποίησης και με μεθόδους λήψης απόφασης που βασίζονται σε γραμμικές συναρτήσεις διάκρισης.
- Οι SVM ονομάζονται επίσης και **ταξινομητές μέγιστου περιθωρίου** (*maximum margin classifiers*).
- Αν και η μέθοδος αυτή μπορεί να γενικευθεί και για την περίπτωση μη-διαχωρίσιμων δεδομένων, θα υποτεθεί στη συνέχεια ότι τα δεδομένα του εκπαιδευτικού συνόλου ανήκουν σε δύο κλάσεις και είναι γραμμικά διαχωρίσιμα.

Support Vector Machines

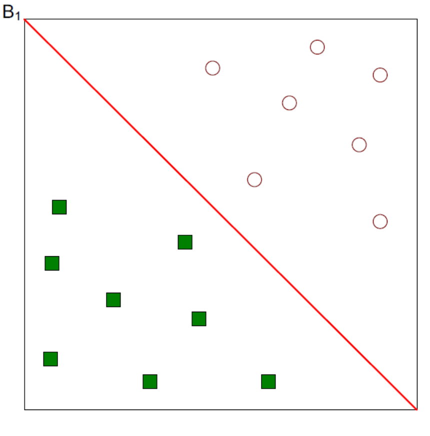
Έστω η γραμμική συνάρτηση διάκρισης $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$. Το όριο απόφασης για δεδομένα δύο κατηγοριών είναι το υπερεπίπεδο $\mathbf{w}^T\mathbf{x} + w_0 = 0$.

Έστω, επίσης, ότι τα δεδομένα είναι 2Δ όπως στο παρακάτω scatter plot.



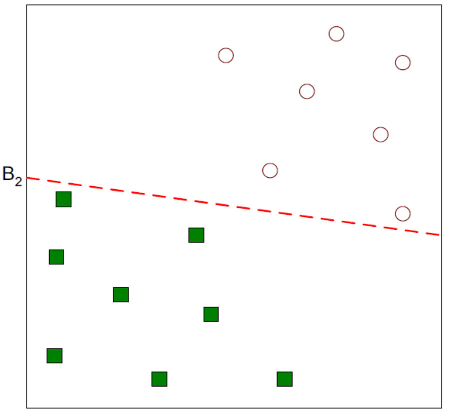
Support Vector Machines

Μιά πιθανή λύση για το γραμμικό όριο είναι η ευθεία B_1 όπως φαίνεται στο παρακάτω σχήμα.



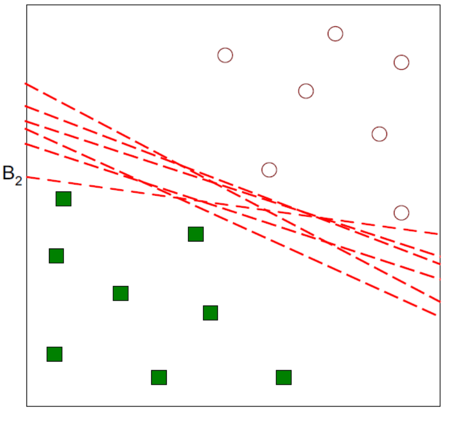
Support Vector Machines

Μιά άλλη πιθανή λύση είναι η ευθεία B_2.



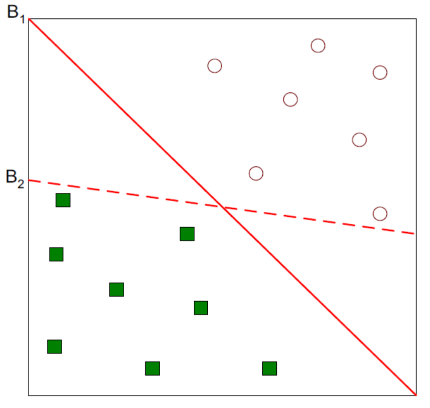
Support Vector Machines

Τελικά, υπάρχουν άπειρες πιθανές λύσεις ...



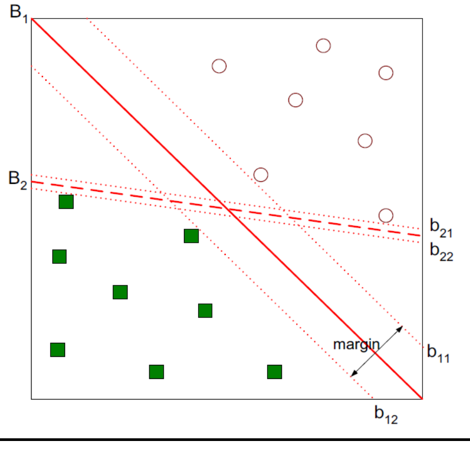
Support Vector Machines

- Ποιά λύση είναι η καλύτερη? Η B_1 ή η B_2?
- Πώς ορίζουμε ποιά λύση είναι η καλύτερη?



Support Vector Machines

- <u>Κριτήριο βελτιστοποίησης</u>: βρες το υπερεπίπεδο που μεγιστοποιεί το περιθώριο (margin) ⇒ Η B_1 είναι καλύτερη λύση από την B_2.



Εισαγωγή

- Οι μηχανές διανυσμάτων υποστήριξης ανακαλύφθηκαν από τον Vladimir Vapnik την δεκαετία του 1970 στη Ρωσία. Στη Δύση έγιναν γνωστές μόλις από τη δεκαετία του 1990.
- Οι SVMs είναι γραμμικοί ταξινομητές που βρίσκουν ένα υπερεπίπεδο για τον διαχωρισμό δεδομένων δύο κλάσεων με διπολική (± 1) κωδικοποίηση των εξόδων (κατηγοριών).
- Στην περίπτωση που οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες η μέθοδος SVM χρησιμοποιεί συναρτήσεις πυρήνα.
- Οι SVM στηρίζονται σε αυστηρές θεωρητικές βάσεις και παρουσιάζουν μεγαλύτερη ακρίβεια ταξινόμησης από πληθώρα άλλων μεθόδων ιδιαίτερα σε εφαρμογές με δεδομένα υψηλών διαστάσεων.
- Αποτελούν, μαζί με μεθόδους βαθιάς μάθησης, έναν από τους καλύτερους ταξινομητές σε προβλήματα ταξινόμησης κειμένου.

Βασικές έννοιες

- Έστω ότι D είναι το σύνολο των παραδειγμάτων εκπαίδευσης:

  $$D = \{ (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_p, y_p)\}$$

  όπου $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in})^T \in \mathbb{R}^n$ είναι ένα πραγματικό διάνυσμα εισόδου και y_i είναι η διπολική έξοδος (η επισήμανση της κατηγορίας στην οποία ανήκει η είσοδος), $y_i \in \{+1, -1\}$.
- Η SVM μέθοδος βρίσκει τα $\mathbf{w}$ και w_0 μιας γραμμικής συνάρτησης της μορφής

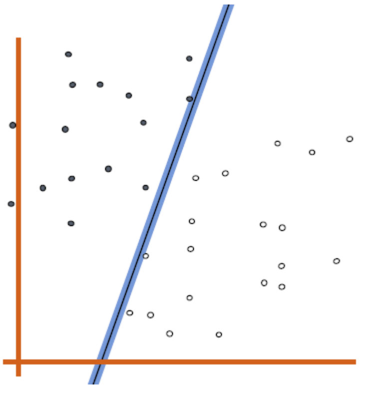
  $$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

  και υπολογίζει την έξοδο σύμφωνα με τον κανόνα απόφασης

  $$y_i = \begin{cases} +1 & \text{αν} \quad \mathbf{w}^T\mathbf{x} + w_0 > 0 \\ -1 & \text{αν} \quad \mathbf{w}^T\mathbf{x} + w_0 < 0 \end{cases}$$

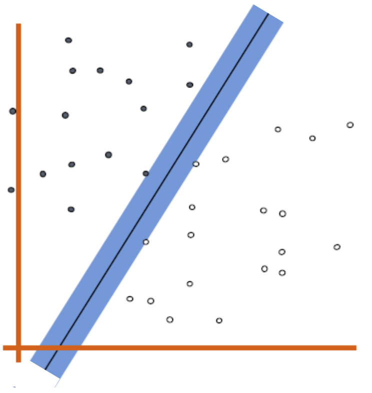
Περιθώριο ταξινομητή

- Ονομάζουμε **περιθώριο** (*margin*) ενός γραμμικού ταξινομητή το πλάτος στο οποίο θα μπορούσε να αυξηθεί το όριο των κλάσεων πριν συναντήσει κάποιο σημείο του εκπαιδευτικού συνόλου.



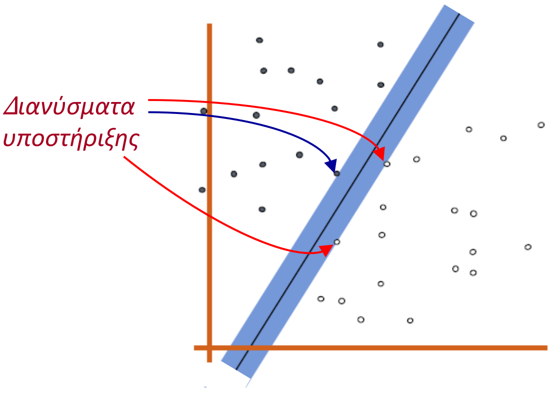
Γραμμικός ταξινομητής μέγιστου περιθωρίου

- Ο **γραμμικός ταξινομητής μέγιστου περιθωρίου** είναι αυτός με το μέγιστο περιθώριο. Ο ταξινομητής αυτός ονομάζεται **LSVM** (linear SVM).



Διανύσματα υποστήριξης

- Τα **διανύσματα υποστήριξης** είναι εκείνα τα σημεία του εκπαιδευτικού συνόλου που περιορίζουν το πλάτος του περιθωρίου γύρω από το όριο.



Γεωμετρική ερμηνεία

- Εξίσωση διαχωριστικού υπερεπιπέδου H: $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$
- Απόσταση διανύσματος $\mathbf{x}$ από το H: $r = |g(\mathbf{x})|/\|\mathbf{w}\| = |\mathbf{w}^T\mathbf{x} + w_0|/\|\mathbf{w}\|$
- Απόσταση θετικού διαν. υποστήριξης ($\mathbf{x}_s^+$):

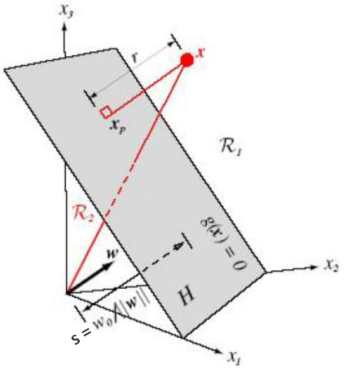
  $r_s^+ = (\mathbf{w}^T\mathbf{x}_s^+ + w_0)/\|\mathbf{w}\|$
- Απόσταση αρνητικού διαν. υποστήριξης ($\mathbf{x}_s^-$):

  $r_s^- = -(\mathbf{w}^T\mathbf{x}_s^- + w_0)/\|\mathbf{w}\|$
- Επειδή δε τα $\mathbf{x}_s^+$ και $\mathbf{x}_s^-$ ισαπέχουν από το H:

  $r_s^+ = r_s^-$

  και επειδή αν πολλαπλασιάσουμε όλα τα βάρη με έναν συντελεστή ρ δεν μεταβάλλεται το H, για κάποιο ρ θα έχουμε: $|\mathbf{w}^T\mathbf{x}_s + w_0| = 1$ και συνεπώς:

  $r_s^+ = r_s^- = 1/\|\mathbf{w}\| \quad \Rightarrow \quad \text{Περιθώριο} = r_s^+ + r_s^- = 2/\|\mathbf{w}\|$



Πρόβλημα βελτιστοποίησης υπό περιορισμούς

- Επιπλέον, ισχύει ότι: $\begin{cases} \mathbf{w}^T\mathbf{x}_i + w_0 \geq 1 & \forall \mathbf{x}_i \in \mathcal{R}_1 \\ \mathbf{w}^T\mathbf{x}_i + w_0 \leq -1 & \forall \mathbf{x}_i \in \mathcal{R}_2 \end{cases}$
- Ισοδύναμα: $y_i\,(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall \mathbf{x}_i \in D$
- Συνεπώς καταλήγουμε στο εξής πρόβλημα βελτιστοποίησης:
  - Ελαχιστοποίηση του: $\|\mathbf{w}\|^2/2$
  - υπό τους περιορισμούς: $y_i\,(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \forall \mathbf{x}_i \in D$
- Η βελτιστοποίηση γίνεται με ορισμό της συνάρτησης Lagrange

  $$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = ½\,\|\mathbf{w}\|^2 - \sum_{i=1}^{p} \lambda_i\,[y_i\,(\mathbf{w}^T\mathbf{x}_i + w_0) - 1]$$

  όπου $\lambda_i \geq 0$ είναι οι πολλαπλασιαστές Lagrange και $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_p]^T$.
- Θέλουμε **μεγιστοποίηση** της L **ως προς $\boldsymbol{\lambda}$** και **ελαχιστοποίηση ως προς $\mathbf{w}$.**

Πρόβλημα βελτιστοποίησης υπό περιορισμούς

Έστω

$$L^*(\mathbf{w}, w_0) = \max_\lambda L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \max_\lambda \left(½\,\|\mathbf{w}\|^2 - \sum_{i=1}^{p} \lambda_i\,[y_i\,(\mathbf{w}^T\mathbf{x}_i + w_0) - 1]\right)$$

Εφόσον $\lambda_i \geq 0$ και $[y_i\,(\mathbf{w}^T\mathbf{x}_i + w_0) - 1] \geq 0$, ο δεύτερος όρος είναι αρνητικός ή μηδέν και άρα η μέγιστη τιμή του ως προς τα λ_i θα είναι το μηδέν. Άρα,

$$L^*(\mathbf{w}, w_0) = ½\,\|\mathbf{w}\|^2$$

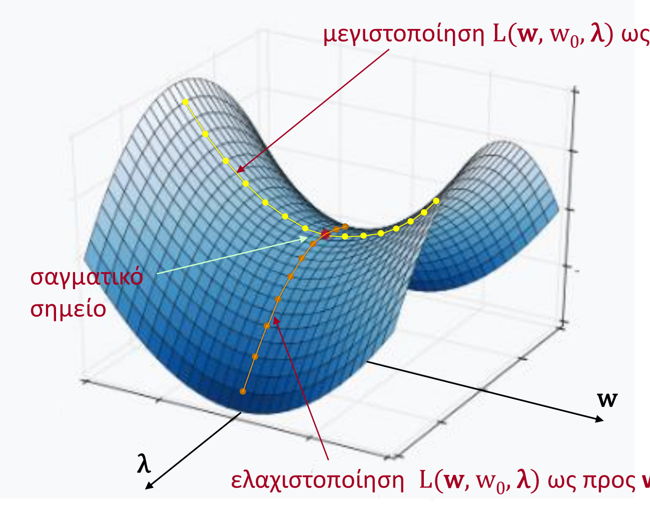
και συνεπώς:

$$\min_w (½\,\|\mathbf{w}\|^2) = \min_w L^*(\mathbf{w}, w_0) = \min_w \max_{\lambda \geq 0} L(\mathbf{w}, w_0, \boldsymbol{\lambda})$$

Μπορεί να δειχθεί ότι για την τετραγωνική συνάρτηση $½\,\|\mathbf{w}\|^2$ και τους γραμμικούς ανισοτικούς περιορισμούς, η θέση του ελαχίστου $\mathbf{w}^*$ για τους αντίστοιχους πολλαπλασιαστές Lagrange $\boldsymbol{\lambda}^*$, είναι σαγματικό σημείο (saddle point) της συνάρτησης Lagrange, για το οποίο ισχύει:

$$L(\mathbf{w}^*, w_0^*, \boldsymbol{\lambda}^*) = \min_w \max_{\lambda \geq 0} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \max_{\lambda \geq 0} \min_w L(\mathbf{w}, w_0, \boldsymbol{\lambda})$$

Πρόβλημα βελτιστοποίησης υπό περιορισμούς



Πρόβλημα βελτιστοποίησης υπό περιορισμούς

Για την επίλυση του προβλήματος **τετραγωνικής βελτιστοποίησης** (quadratic optimization), εφαρμόζουμε τις **συνθήκες Karush-Kuhn-Tucker** για τη θέση του σαγματικού σημείου (ελάχιστο ως προς **w**, w_0, μέγιστο ως προς **λ**):

1) $\nabla_{\mathbf{w}} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0$
2) $\frac{\partial}{\partial w_0} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0$
3) $\lambda_i \geq 0 \quad \forall i = 1, 2, \ldots, p$
4) $\lambda_i \,[y_i\,(\mathbf{w}^T \mathbf{x}_i + w_0) - 1] = 0 \quad \forall i = 1, 2, \ldots, p$

όπου L( ) είναι η συνάρτηση Lagrange:

$$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = ½\, \mathbf{w}^T\mathbf{w} - \sum_{i=1}^{p} \lambda_i\, y_i\, \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^{p} \lambda_i\, y_i\, w_0 + \sum_{i=1}^{p} \lambda_i$$

Τετραγωνική βελτιστοποίηση υπό περιορισμούς

- Από την (1):

$$\mathbf{w} - \sum_{i=1}^{p} \lambda_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{p} \lambda_i y_i \mathbf{x}_i \qquad (5)$$

- Από την (2):

$$\sum_{i=1}^{p} \lambda_i y_i = 0 \qquad (6)$$

Συνεπώς, με αντικατάσταση των (5) και (6) στη συνάρτηση Lagrange θα έχουμε:

$$L(\boldsymbol{\lambda}) = ½ \left(\sum_{i=1}^{p} \lambda_i y_i \mathbf{x}_i^T\right)\left(\sum_{j=1}^{p} \lambda_j y_j \mathbf{x}_j\right) - \sum_{i=1}^{p} \lambda_i y_i \left(\sum_{j=1}^{p} \lambda_j y_j \mathbf{x}_j\right)^T \mathbf{x}_i - \cancelto{0}{\sum_{i=1}^{p} \lambda_i y_i} w_0 + \sum_{i=1}^{p} \lambda_i =$$

$$= -½ \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{p} \lambda_i$$

Δυϊκό πρόβλημα βελτιστοποίησης για εύρεση των λ_i

και οι πολλαπλασιαστές Lagrange βρίσκονται από το δυϊκό πρόβλημα βελτιστοποίησης:

Μεγιστοποίησε την $L(\boldsymbol{\lambda}) = \sum_{i=1}^{p} \lambda_i - ½ \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$

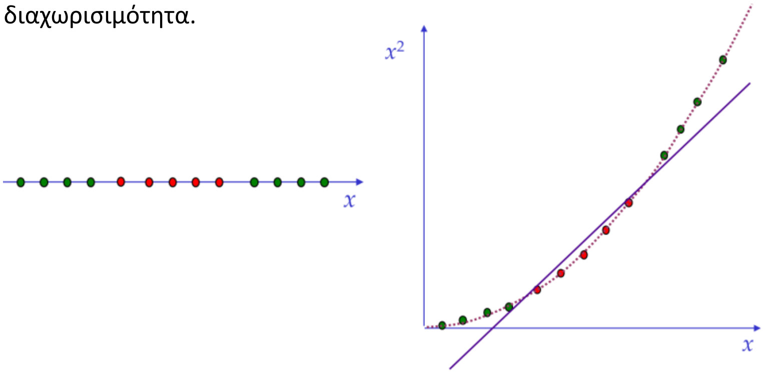
Έτσι ώστε $\sum_{i=1}^{p} \lambda_i y_i = 0 \,, \quad \lambda_i \geq 0 \;\; \forall i = 1, 2, \ldots, p$

- Αφού βρεθούν βέλτιστες τιμές για τα λ_i, το διάνυσμα βαρών **w** υπολογίζεται ως $\mathbf{w} = \sum_{x_i \in D} \lambda_i y_i \mathbf{x}_i$ και επειδή $\lambda_i = 0$ για όλα τα διανύσματα εισόδου εκτός από τα διανύσματα υποστήριξης, θα έχουμε $\mathbf{w} = \sum_{x_i \in S} \lambda_i y_i \mathbf{x}_i$ όπου S είναι το σύνολο των διανυσμάτων υποστήριξης.
- Επίσης, έχουμε ότι $w_0 = 1/y_i - \mathbf{w}^T \mathbf{x}_i$ ή $w_0 = \frac{1}{|S|} \sum_{x_i \in S} (1/y_i - \mathbf{w}^T \mathbf{x}_i)$
- Συνεπώς, η συνάρτηση διάκρισης θα είναι:

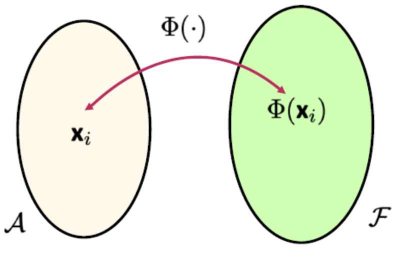
$$g(\mathbf{x}) = \left(\sum_{x_i \in S} \lambda_i y_i \mathbf{x}_i\right)^T \mathbf{x} + w_0 = \sum_{x_i \in S} \lambda_i y_i \mathbf{x}_i^T \mathbf{x} + \frac{1}{|S|} \sum_{x_i \in S} \left(1/y_i - \sum_{x_j \in S} \lambda_j y_j \mathbf{x}_j^T \mathbf{x}_i\right)$$

SVM και μη γραμμικά διαχωρίσιμες κλάσεις: η μη γραμμική περίπτωση

- Στην περίπτωση που οι κλάσεις δεν διαχωρίζονται γραμμικά μετασχηματίζουμε τον χώρο εισόδων σε χώρο χαρακτηριστικών περισσοτέρων διαστάσεων στον οποίο έχουμε γραμμική διαχωρισιμότητα.



Απεικόνιση σε γραμμικά διαχωρίσιμες κλάσεις



$\mathcal{A}$: χώρος εισόδου

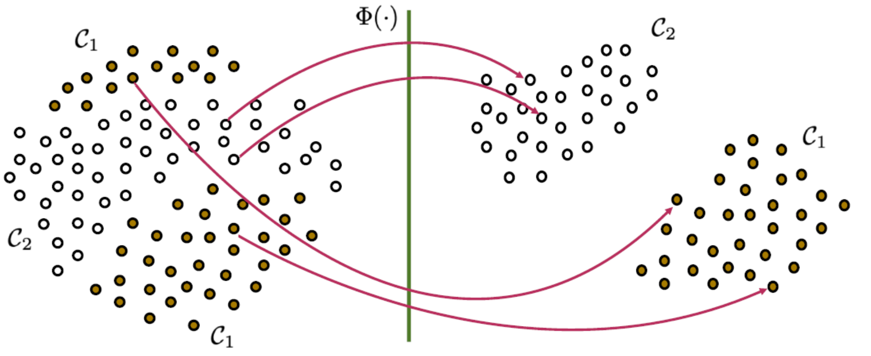
$\mathcal{F}$: χώρος χαρακτηριστικών

$\Phi(\cdot)$: μη-γραμμική συνάρτηση απεικόνισης

Θεώρημα Cover

Κάθε πολυδιάστατος χώρος με μη γραμμικά διαχωρίσιμα πρότυπα, μπορεί να μετασχηματιστεί σε ένα νέο χώρο στον οποίο τα πρότυπα είναι γραμμικά διαχωρίσιμα με *υψηλή πιθανότητα*, αρκεί ο μετασχηματισμός να είναι μη γραμμικός και ο νέος αυτός χώρος να έχει την απαραίτητη διάσταση

Απεικόνιση σε γραμμικά διαχωρίσιμες κλάσεις



Απεικόνιση σε γραμμικά διαχωρίσιμες κλάσεις

- $\mathbf{w} = \sum_{\Phi(\mathbf{x}_i)\in S} \lambda_i y_i \Phi(\mathbf{x}_i)$
- $w_0 = \frac{1}{|S|} \sum_{\Phi(\mathbf{x}_i)\in S} \left(\frac{1}{y_i} - \mathbf{w}^T \Phi(\mathbf{x}_i) \right)$
- Συνάρτηση διάκρισης:

$$g^*(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + w_0 = \sum_{\Phi(\mathbf{x}_i)\in S} \lambda_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + w_0$$

- Παρατηρούμε ότι δεν χρειάζεται να υπολογίσουμε αναλυτικά τον μετασχηματισμό $\Phi(\cdot)$ των δεδομένων $\mathbf{x}$ προς ταξινόμηση. Αρκεί να υπολογίσουμε τα πολύ "φθηνότερα" εσωτερικά γινόμενα $\Phi(\mathbf{x})^T \Phi(\mathbf{y})$.

Χρήση συναρτήσεων πυρήνα

Ορισμός

Ορίζουμε τη συνάρτηση $k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})^\top\Phi(\mathbf{y})$, την οποία θα ονομάζουμε συνάρτηση πυρήνα.

Χρησιμοποιώντας τη συνάρτηση πυρήνα κάνουμε οικονομία πράξεων ειδικά όταν η διάσταση του $\Phi(\mathbf{x})$ είναι μεγαλύτερη από τη διάσταση του $\mathbf{x}$ (όπως συνήθως συμβαίνει)

Παράδειγμα.

Έστω $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ $\quad \Phi(\mathbf{x}) = \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix}^\top$

Για $\mathbf{x} = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top$ $\quad \Phi([1\ 2]^\top) = \begin{bmatrix} 1 & 2\sqrt{2} & 4 \end{bmatrix}^\top$

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= \Phi(\mathbf{x})^\top\Phi(\mathbf{y}) = (x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2) \\ &= (x_1y_1 + x_2y_2)^2 = (\mathbf{x}^\top\mathbf{y})^2 \end{aligned}$$

Χρήση συναρτήσεων πυρήνα

- Συνεπώς, αντί να υπολογίσουμε αναλυτικά το $\mathbf{w}$, υπολογίζουμε κατευθείαν την τιμή της συνάρτησης διάκρισης $g^*(\mathbf{x})$ για κάθε $\mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbf{w}^\mathrm{T}\Phi(\mathbf{x}) + w_0 = \sum_{x_i \in D} \lambda_i y_i \Phi(\mathbf{x}_i)^\mathrm{T}\Phi(\mathbf{x}) + w_0 = \sum_{x_i \in D} \lambda_i y_i k(\mathbf{x}_i, \mathbf{x}) + w_0$$

  όπου οι πολλαπλασιαστές Lagrange βρίσκονται από την επίλυση αντίστοιχου (δυϊκού) προβλήματος βελτιστοποίησης υπό περιορισμούς.

- Κανόνας απόφασης για ταξινόμηση νέων δεδομένων $\mathbf{x}$:

  Ταξινόμησε το $\mathbf{x}$ στην ω_1 αν $g^*(\mathbf{x}) > 0$, ειδάλλως στην ω_2.