

Ερώτημα 1

Δίνονται οι ακόλουθοι ισχυρισμοί σχετικά με την Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA):

PCA

- 1) Είναι μέθοδος χωρίς επίβλεψη.
- 2) Αναζητά αυτές τις κατευθύνσεις έτσι ώστε τα δεδομένα να έχουν τη μεγαλύτερη διασπορά.
- 3) Ο μέγιστος αριθμός των κύριων συνιστωσών $\leq$ του αρχικού αριθμού χαρακτηριστικών.
- 4) Όλες οι συνιστώσες είναι ορθογώνιες μεταξύ τους.

Ποιο από τα παρακάτω ισχύει;

- a) Όλα.
- b) 1 και 2.
- c) 1 και 3.
- d) Μόνο 3.

PCA (Principal Components Analysis)

Στόχοι: ▷ μείωση διαστάσεων πολυδιάστατων δεδομένων

▷ εύρεση νέων χαρακτηριστικών που ετήχουν μεγαλύτερη διακύμανση δεδομένων ²

Χαρακτηριστικά: ▷ δεν λαμβάνει υπόψη labels → unsupervised ¹

▷ αρχικό $\xrightarrow[\text{μικρότερο}]{\text{μεγαλύτερο}}$ νέο σύνολο } principal components: κάθετες / ορθογωνί ⁴ / ασυσχετιστές

▷ 1^η κύρια συνιστώσα: μεγαλύτερη διακύμανση στα δεδομένα

Βήματα: ① Κανονικοποίηση δεδομένων

② Πίνακας συμμεταβατότητας: $S = \frac{1}{n-1} \cdot X^T X$

③ Ιδιοτιμές / Ιδιοδιανύσματα του S : κάθε ιδιοσ καθορίζει σημαντικότητα αντισφ ιδιοδιαν

④ Ταξινόμηση λ_i με φθίνουσα σειρά λ_i

⑤ Προβοθή στον νέο χώρο των κύριων συνιστωσών

Μαθηματική ανάλυση: ▷ πίνακας δεδομένων X : $n \times m$

▷ ορθοκανονικός πίνακας ιδιοδιαν συμμεταβα W : $n \times k$

▷ πίνακας προβοθών Z : $m \times k$

n : # παρατηρη (δείγματα)

m : # χαρακτηρ (διαστάσεις)

k : # κύριων συνιστωσών

} $k \leq m$ ³

Εξτρά: ▷ ΔΕΝ εφασφατίζει θεάτωση προβοθών για ταξινόμηση

▷ εφαιρετική για unsupervised εφαρμογές

▷ δεν λαμβάνει υπόψη labels → λιχότερο αποτελεσματική για ταξινόμηση

▷ για διαχωρισμό κλάσεων προτιμούνται supervised methods (πχ LDA)

Ερώτημα 2

Ποιο από τα παρακάτω ισχύουν για τη μέθοδο Linear Discriminant Analysis (LDA);

- a) Μεγιστοποιεί τόσο την απόσταση μεταξύ διαφορετικών κατηγοριών όσο και την απόσταση εντός κατηγοριών.
- b) Ελαχιστοποιεί τόσο την απόσταση μεταξύ διαφορετικών κατηγοριών όσο και την απόσταση εντός κατηγοριών.
- c) Μεγιστοποιεί την απόσταση μεταξύ διαφορετικών κατηγοριών και ελαχιστοποιεί την απόσταση εντός κατηγοριών.
- d) Ελαχιστοποιεί την απόσταση μεταξύ διαφορετικών κατηγοριών και μεγιστοποιεί την απόσταση εντός κατηγοριών.

LDA

LDA (Linear Discriminant Analysis)

Στόχοι: ▷ μείωση διάστασης διατηρώντας διακριτική ικανότητα κλάσεων

▷ εύρεση ιδιοδίων κατά τις διευθύνσεις οι οποίες ταξίδια διαχωρίζονται καλύτερα

Χαρακτηριστικά: ▷ λαμβάνει υπόψη labels → supervised

▷ εύρεση υπόχωρου: between-class scatter / διαταξική διασπορά $\sim \max S_b$
within class scatter / ενδοταξική διασπορά $\sim \min S_w$

▷ μείωση διαστάσεων πολυδιάστατων δεδομένων

Μαθηματική ανάλυση: ▷ within-class Scatter matrix: $S_w = \sum_{i=1}^C \sum_{j=1}^{n_i} (x_j - m_i)(x_j - m_i)^T$
($x_j - m_i$): απόκλιση x_j από μέσο κατηγορίας

▷ between-class Scatter matrix: $S_b = \sum_{i=1}^C n_i (m_i - m)(m_i - m)^T$
($m_i - m$): απόκλιση μέσου m_i κατηγορίας i από το γενικό μέσο m

C : # classes x_j : observation vector in class i

n_i : # observations συν class i m_i : mean obs vector in class i

m : γενικό μέσο όλων των δειγμάτων

$$\triangleright \arg \max_w J = \frac{|S_b|}{|S_w|} = \frac{|W^T S_b W|}{|W^T S_w W|}$$

! Η μήτρα S_w δεν είναι αντιστρέψιμη όταν τα δείγματα D είναι λιγότερα από αρχικές διαστάσεις N

→ εφαρμογή PCA στον πίνακα δεδομ X ($N \times D$) → προβάλλεται σε Y χαμηλ διαστάσεις με διαστάσεις $\leq N-1$

→ διαχωρ κατεύθυνσεις $\leq C-1$

Ερώτημα 3

Έστω ότι έχουμε τα εξής δεδομένα $[(x^{(1)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})]$ και τη linear regression λύση για αυτά τα δεδομένα $y = w_1 \cdot x + b_1$. Θεωρήστε, επίσης και το εξής σύνολο δεδομένων $[(x^{(1)} + \alpha, y^{(2)} + \beta), \dots, (x^{(n)} + \alpha, y^{(n)} + \beta)]$, με $\alpha, \beta > 0$ και $w_1 \cdot \alpha \neq \beta$. Η linear regression λύση για αυτό το σύνολο δεδομένων είναι $y = w_2 \cdot x + b_2$. Επιλέξτε ποιος από τους παρακάτω ισχυρισμούς για τα w_1, b_1 (για οποιαδήποτε επιλογή των τιμών α, β εντός των δοθέντων περιορισμών):

- a) $w_1 = w_2, b_1 = b_2$
- b) $w_1 \neq w_2, b_1 = b_2$
- c) $w_1 = w_2, b_1 \neq b_2$
- d) $w_1 \neq w_2, b_1 \neq b_2$

Linear Regression

	Σύνολο Δεδομένων		Linear regression
1	$\{(x_1^{(1)}, y_1^{(1)}), \dots, (x_1^{(n)}, y_1^{(n)})\}$	$\{(x_1^{(i)}, y_1^{(i)})\}_{i=1}^n$	$y_1 = w_1 x_1 + b_1$
2	$\{(x_2^{(1)}, y_2^{(1)}), \dots, (x_2^{(n)}, y_2^{(n)})\}$	$\{(x_2^{(i)}, y_2^{(i)})\}_{i=1}^n$	$y_2 = w_2 x_2 + b_2$
	$\{(x_1^{(1)} + \alpha, y_1^{(1)} + \beta), \dots, (x_1^{(n)} + \alpha, y_1^{(n)} + \beta)\}$	$\{(x_1^{(i)} + \alpha, y_1^{(i)} + \beta)\}_{i=1}^n$	

$$y_1 = w_1 \cdot x_1 + b_1 \Rightarrow$$

$$y_1 + \beta - \beta = w_1(x_1 + \alpha) - w_1 \alpha + b_1 \Rightarrow$$

$$y_2 - \beta = w_1 \cdot x_2 - w_1 \alpha + b_1 \Rightarrow$$

$$y_2 = \underbrace{w_1}_{w_2} \cdot x_2 + \underbrace{b_1 + \beta - w_1 \alpha}_{b_2}$$

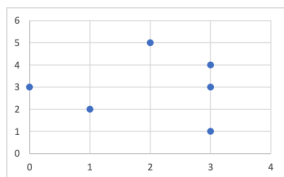
$$\triangleright w_2 = w_1$$

$$\triangleright b_2 = b_1 + \beta - w_1 \alpha \quad \left. \vphantom{\begin{matrix} \triangleright w_2 = w_1 \\ \triangleright b_2 = b_1 + \beta - w_1 \alpha \end{matrix}} \right\} b_2 \neq b_1$$

$$\triangleright \beta \neq \alpha w_1$$

Ερώτημα 4

Σας δίνονται τα εξής σημεία $\mathbf{x}_i = (x_i, y_i)$, $i = 1, \dots, N = 6$, στο επίπεδο $(0,3), (1,2), (2,5), (3,1), (3,3), (3,4)$ του παρακάτω σχήματος.



Το ιδιοδιάνυσμα που αντιστοιχεί στη κύρια συνιστώσα της ανάλυσης PCA προκύπτει ότι είναι ίσο με:

- a) $\begin{pmatrix} -0.93 \\ 1 \end{pmatrix}$
- b) $\begin{pmatrix} 0.55 \\ 1 \end{pmatrix}$
- c) $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- d) $\begin{pmatrix} 1.29 \\ 1 \end{pmatrix}$

Σημείωση: Τα παραπάνω ιδιοδιανύσματα έχουν κανονικοποιηθεί ως προς τη 2^η συντεταγμένη τους. Επίσης, σας δίνεται ότι ο unbiased εκτιμητής του πίνακα συνδιακύμνωσης:

$$C_{\text{unbiased}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}) \cdot (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Μέσος όρος: $\bar{\mathbf{x}} = 1/N \sum_{i=1}^N \mathbf{x}_i = 1/6 (12, 18) = (2, 3)$

Unbiased εκτιμητής covariance matrix:

$$C_{\text{unbiased}} = 1/(N-1) \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N-1} \sum_{i=1}^N \begin{bmatrix} (x_i - \bar{x})^2 & (x_i - \bar{x})(y_i - \bar{y}) \\ (y_i - \bar{y})(x_i - \bar{x}) & (y_i - \bar{y})^2 \end{bmatrix}$$

$$\sum (x_i - \bar{x})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 1^2 + 1^2 = 8$$

$$\sum (y_i - \bar{y})^2 = 0^2 + (-1)^2 + 2^2 + (-2)^2 + 0^2 + 1^2 = 10$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (-2) \cdot 0 + (-1) \cdot (-1) + 0 \cdot 2 + 1 \cdot (-2) + 1 \cdot 0 + 1 \cdot 1 = 0$$

$$\left. \begin{matrix} \sum (x_i - \bar{x})^2 = 8 \\ \sum (y_i - \bar{y})^2 = 10 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) = 0 \end{matrix} \right\} C_{\text{unbiased}} = \frac{1}{5} \begin{bmatrix} 8 & 0 \\ 0 & 10 \end{bmatrix} = \begin{bmatrix} 1.6 & 0 \\ 0 & 2 \end{bmatrix}$$

Ιδιοδιανύσματα: $\det(C - \lambda I_2) = 0 \Rightarrow \begin{vmatrix} 1.6 - \lambda & 0 \\ 0 & 2 - \lambda \end{vmatrix} = 0 \Rightarrow (1.6 - \lambda)(2 - \lambda) = 0 \Rightarrow \begin{cases} \lambda_1 = 2 \\ \lambda_2 = 1.6 \end{cases}$

Η κύρια συνιστώσα αντιστοιχεί στη μεγαλύτερη ιδιοτιμή $\lambda_1 = 2$

Έστω $\mathbf{u}_1$ το αντιστοίχο ιδιοδιάνυσμα: $\mathbf{u}_1 = [a \ 1]^T$

$$C \cdot \mathbf{u}_1 = \lambda_1 \cdot \mathbf{u}_1 \Rightarrow \begin{bmatrix} 1.6 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} a \\ 1 \end{bmatrix} = 2 \cdot \begin{bmatrix} a \\ 1 \end{bmatrix} \Rightarrow \begin{cases} 1.6a = 2a \\ 2 = 2 \end{cases} \Rightarrow a = 0 \Rightarrow \mathbf{u}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Ερώτημα 5

Έστω ένα βαθύ νευρωνικό δίκτυο πολλαπλών επιπέδων-στρωμάτων με δεδομένα εισόδου γκριζες εικόνες διαστάσεων 300×300 pixels. Με εξαίρεση λίγα τελικά επίπεδα, σε κάθε επίπεδο υπολογίζονται 10 χάρτες βαθμωτών χαρακτηριστικών (feature maps) διαστάσεων 300×300 , είτε με αρχιτεκτονική ενός Fully Connected Neural Network (FCNN) είτε με αρχιτεκτονική ενός συνελκτικού νευρωνικού δικτύου (Convolutional Neural Network - CNN) που χρησιμοποιεί πυρήνες 3×3 βαρών για κάθε χαρακτηριστικό. Ποιο από τα παρακάτω ισχύει;

- a) Οι αρχιτεκτονικές FCNN και CNN έχουν περίπου την ίδια τάξη αριθμού παραμέτρων, αλλά το CNN προτιμάται επειδή προσφέρει την εξαγωγή χαρακτηριστικών.
- b) Το CNN επιτυγχάνει μείωση του αριθμού παραμέτρων ανά επίπεδο δικτύου έναντι του FCNN κατά $10^5:1$.
- c) Το CNN επιτυγχάνει μείωση του αριθμού παραμέτρων ανά επίπεδο δικτύου έναντι του FCNN κατά $10^4:1$.
- d) Καμία από τις παραπάνω.

Input: 300×300 grayscale images
⊕ 10 feature maps on each level

① FCNN

Κάθε-είσοδος συνδέεται με κάθε-έξοδο = μεγάλος αριθμός παραμέτρων

10 feature maps,

{ κάθε feature map υπολογ από όλα τα 90.000 pixels } = $\forall$ επίπεδο: παραμ. = $10 \cdot (90.000 + 1) = \underline{900.010}$
κάθε feature map έχει μια σταθερά (bias)

② CNN

3×3 kernels που είναι shared σε όλα τα pixels του map vector

10 feature maps,

{ κάθε feature map έχει πυρήνια με 9 βάρη } = $\forall$ επίπεδο: παραμ. = $10 \cdot (9 + 1) = \underline{100}$
κάθε feature map έχει μια σταθερά (bias)

$$\text{Αναλογία παραμ.: } \frac{900,010}{100} \approx \frac{10^6}{10^2} = \frac{10^4}{1}$$

FCNN (Fully Connected Neural Network)

Περιγραφή: ▶ αποτελείται από επίπεδα νευρώνων, όπου κάθε νευρώνας ενός επιπέδου συνδέεται με κάθε νευρώνα του επόμενου επιπέδου
Όλοι συνδέονται αντιστοιχούν σε βάση που προσαρμόζεται από εκπαίδευση για ↓ σφ

Pros η Cons: ① δυνατότητα εκμάθησης πολύπλοκων σχέσεων
② υψηλός αριθμός παραμέτρων, κίνδυνος overfitting

Χρήση: - πλήρες σύνολο χαρακτηριστικών διαθέσιμο
- όχι περιορισμοί σε χωρητικότητα υπολογισμών

CNN (Convolutional Neural Network)

Χαρακτηριστικά: ▶ ελάττει τοπικές ιδιότητες όπως άκρα ή υφές

Βασικές Διαστάσεις: ① Convolution Step: δεδομ $\xrightarrow[\text{για ελάττωση}]{\text{νύσση (kernel)}}$ feature maps
εισόδου

② Nonlinearity Step: χρησιμοποίηση μη γραμμικής συνάρτησης (πχ ReLU) για να ελαττώσει μη γραμμικότητα

③ Pooling Step: συμπίεση πληροφορίας { μείωση διαστάσεων
διατήρηση πληροφορίας

1. νύσση $H, k \times k$, συνείληθη pixel (n, m) : $O(n, m) = \sum_{i=1}^k \sum_{j=1}^k h_{ij} \cdot I(n+i-1, m+j-1)$
~ τοπικές ιδιότητες, υφές, άκρες (τιμή σε δάν παρ) (input)

3. Pooling: μείωση διαστάσεων, διατηρούνται σημαντικά χαρακτηριστικά
{ max pooling: μέγιστη τιμή σε κάθε παράθυρο } ~ ενισχύει ανθεκτικότητα σε μικρές παραμ. εικόνες
average pooling: μέσος όρος τιμών

Pros η Cons: ① μείωση παραμ. δ' χω κοινών βάσεων

② ικανότητα μάθησης τοπικών χαρακτηριστ

③ απαιτεί μεγάλη υπολ. πολυπλοκότητα και μέγ. ποσότητες δεδομ

Ερώτημα 6

Αν $x_1, \dots, x_N \in R^+$ είναι ανεξάρτητες παρατηρήσεις που προέρχονται από τυχαίες μεταβλητές που ακολουθούν την εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας $p(x) = \lambda \cdot e^{-\lambda \cdot x}, x \geq 0$, η εκτίμηση μέγιστης πιθανοφάνειας της παραμέτρου λ της κατανομής θα δίνεται από τη σχέση:

$$a) \lambda_{ML} = N \frac{1}{\sum_{i=1}^N x_i}$$

$$b) \lambda_{ML} = \frac{1}{N \sum_{i=1}^N x_i}$$

$$c) \lambda_{ML} = \frac{1}{N^2 \sum_{i=1}^N x_i}$$

$$d) \lambda_{ML} = N \frac{1}{\sum_{i=1}^N x_i^2}$$

$$\text{σηη εκθετικής: } p(x) = \lambda \cdot e^{-\lambda x}, x \geq 0$$

$$\text{συνάρτηση πιθανοφάνειας: } L(\lambda) = \prod_{i=1}^N p(x_i) = \prod_{i=1}^N \lambda \cdot e^{-\lambda x_i} = \lambda^N \cdot e^{-\lambda \sum_{i=1}^N x_i}$$

$$\text{συνάρτηση λογ πιθανοφάνειας: } \ell(\lambda) = \log L(\lambda) = N \cdot \log \lambda - \lambda \sum_{i=1}^N x_i$$

$$\text{μηδενισμός παραγώγου: } \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \frac{N}{\lambda} - \sum_{i=1}^N x_i = 0 \Rightarrow \lambda = \frac{N}{\sum_{i=1}^N x_i}$$

Ερώτημα 7

Έστω ένα πρόβλημα ταξινόμησης σε δύο κλάσεις ω_1 και ω_2 στο R με τη χρήση του ταξινομητή Bayes. Οι 2 κλάσεις θεωρούνται ισοπίθανες, δηλαδή $P(\omega_1) = P(\omega_2) = 0.5$. Έστω, επίσης, ότι τα δεδομένα στην κλάση ω_1 ακολουθούν την κατανομή $p(x|\omega_1) = 6 \cdot (1-x), 0 \leq x \leq 1$ και στην ω_2 την κατανομή $p(x|\omega_2) = 6 \cdot (x-0.5) \cdot (1.5-x), 0.5 \leq x \leq 1.5$. Αν τα λάθη που σχετίζονται με τις 2 κλάσεις δεν έχουν την ίδια βαρύτητα, αλλά τα αντίστοιχα βάρη είναι $\lambda_{12} = 0.75$ και $\lambda_{21} = 1$, η τιμή του κατωφλίου x_r , που διαχωρίζει τις 2 κλάσεις και ελαχιστοποιεί το μέσο ρίσκο είναι:

$$a) 0.7$$

$$b) 0.6$$

$$c) 0.75$$

$$d) 0.65$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$p(x|\omega_1) = 6(1-x)(1.5-x), 0.5 \leq x \leq 1.5$$

$$\text{βάρη σφαλμάτων: } \lambda_{12} = 0.75, \lambda_{21} = 1$$

$$p(x|\omega_2) = 6(1-x), 0 \leq x \leq 1$$

κανόνας ελαχιστοποίησης κινδύνου:

$$\lambda_{12} P(\omega_1) \cdot p(x|\omega_1) = \lambda_{21} P(\omega_2) \cdot p(x|\omega_2) \Rightarrow$$

$$\lambda_{12} \cdot P(\omega_1 | x) = \lambda_{21} P(\omega_2 | x) \Rightarrow$$

$$0.75 \cdot 6 \cdot (1-x_r) = 6 \cdot (x_r - 0.5)(1.5 - x_r) \Rightarrow$$

$$x_r^2 - 2.75x_r + 1.5 = 0 \Rightarrow$$

$$x_r = 0.75 \text{ ή } x_r = 2 \quad \left. \vphantom{x_r = 0.75 \text{ ή } x_r = 2} \right\} x_r = 0.75$$

$$0.5 \leq x_r \leq 1$$

Ερώτημα 8

Έστω το πρόβλημα ταξινόμησης κατά Bayes σε 2 ισοπίθανες κλάσεις στο $\mathbb{R}^2$, όπου τα $p(x|\omega_1)$, $p(x|\omega_2)$ είναι Gaussian κατανομές με μέσα διανύσματα $\mu_1 = [1, 3]^T$, $\mu_2 = [2, -1]^T$ και κοινό πίνακα συμμεταβλητότητας:

$$\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

Η εξίσωση της ευθείας που διαχωρίζει τις 2 κλάσεις κατά βέλτιστο τρόπο, ελαχιστοποιώντας το σφάλμα, είναι:

a) $y = \frac{1}{4}x - \frac{3}{4}$

b) $y = \frac{1}{3}x + \frac{5}{8}$

c) $y = \frac{1}{3}x - \frac{3}{4}$

d) $y = \frac{1}{4}x + \frac{5}{8}$

$$P_2(\omega_1) = P(\omega_2) = 0.5$$

$P_{x|\omega_1}, P_{x|\omega_2} \sim \text{Gaussian}$

$$\mu_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \Sigma^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

συνάρτηση $g(x) = w^T x + w_0$

$$\text{διαν. συντελεστής: } w = \Sigma^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} -2 & 8 \end{bmatrix}$$

$$\text{σταθερά bias } w_0: w_0 = -\frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2) = -\frac{1}{2} (20 - 10) = -5$$

$$\mu_1^T \Sigma^{-1} \mu_1 = \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = 2(1)^2 + 2(3)^2 = 20$$

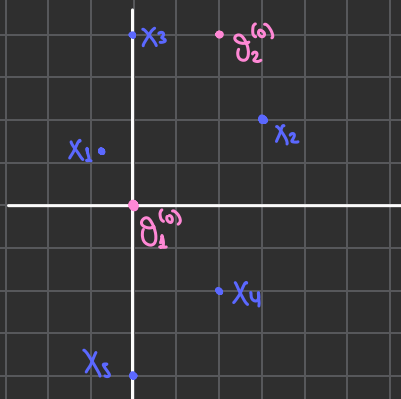
$$\mu_2^T \Sigma^{-1} \mu_2 = 2(2)^2 + 2(-1)^2 = 10$$

$$\text{Συνολικά: } g(x) = \begin{bmatrix} -2 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 5 = -2x_1 + 8x_2 - 5 \Rightarrow y = \frac{1}{4}x + \frac{5}{8}$$

Ερώτημα 9

Χρησιμοποιούμε τον αλγόριθμο k-means για τον διαχωρισμό των σημείων $x_1 = [-0.4, 0.6]$, $x_2 = [1.5, 1]$, $x_3 = [0, 2]$, $x_4 = [1, -1]$, $x_5 = [0, -2]$ σε 2 ομάδες. Αν τα κέντρα των ομάδων αρχικοποιούνται ως $\theta_1^{(0)} = [0, 0]$ και $\theta_2^{(0)} = [1, 2]$, οι θέσεις των κέντρων μετά την ολοκλήρωση της 1^{ης} επανάληψης του αλγορίθμου θα είναι:

- a) $\theta_1^{(1)} = [0.4, 1.25]$, $\theta_2^{(1)} = [0.4, -0.6]$
- b) $\theta_1^{(1)} = [0.4, -0.6]$, $\theta_2^{(1)} = [0.5, 1.25]$
- c) $\theta_1^{(1)} = [0.75, 1.5]$, $\theta_2^{(1)} = [0.2, -0.8]$
- d) $\theta_1^{(1)} = [0.2, -0.8]$, $\theta_2^{(1)} = [0.75, 1.5]$



1^η επανάληψη:

κλάση 1: x_3, x_4, x_5

κλάση 2: x_2, x_1

$$\theta_1^{(1)} = \frac{x_3 + x_4 + x_5}{3} = [0.2, -0.8]$$

$$\theta_2^{(1)} = \frac{x_2 + x_1}{2} = [0.75, 1.5]$$

Ερώτημα 10

Δίνεται ο παρακάτω πίνακας εγγύτητας (ανομοιότητας) 5 σημείων $x_i \in \mathbb{R}^2$, $i = 1, \dots, 5$:

$$P(X) = \begin{pmatrix} 0 & 3.42 & 2.24 & 3.61 & 5.83 \\ 3.42 & 0 & 2.83 & 3.16 & 3.61 \\ 2.24 & 2.83 & 0 & 1.41 & 4.12 \\ 3.16 & 3.16 & 1.41 & 0 & 2.15 \\ 5.83 & 3.61 & 4.12 & 2.15 & 0 \end{pmatrix}$$

Αν με βάση τον $P(X)$ εφαρμόσουμε τον ιεραρχικό συσσωρευτικό αλγόριθμο πλήρους δεσμού (complete link), η ομαδοποίηση που προκύπτει μετά το 2^ο βήμα του αλγορίθμου γίνεται:

- a) $\{x_1\}, \{x_2, x_3, x_4\}, \{x_5\}$
- b) $\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5\}$
- c) $\{x_1, x_2\}, \{x_3\}, \{x_4, x_5\}$
- d) $\{x_1, x_2\}, \{x_3, x_4\}, \{x_5\}$

Complete Link
απόσταση 2 clusters $\rightarrow \max$

Single Link
απόσταση 2 clusters $\rightarrow \min$

Αρχικά: $\{x_1\} \{x_2\} \{x_3\} \{x_4\} \{x_5\}$

Βήμα 1: min απόσταση $d_{34} = d_{43} = 1.41 \leadsto \{x_1\} \{x_2\} \{x_3, x_4\} \{x_5\}$

Βήμα 2: $d(\{x_3, x_4\}, x_5) = \max \{d(x_3, x_5), d(x_4, x_5)\} = \max \{2.24, 3.16\} = 3.16$

$d(\{x_3, x_4\}, x_2) = \max \{d(x_3, x_2), d(x_4, x_2)\} = \max \{2.83, 3.16\} = 3.16$

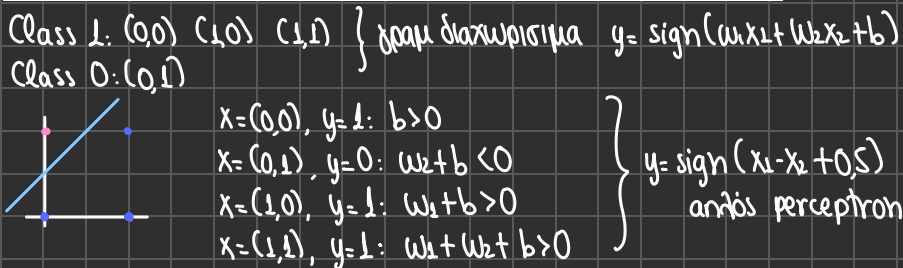
$d(\{x_3, x_4\}, x_1) = \max \{d(x_3, x_1), d(x_4, x_1)\} = \max \{4.12, 2.15\} = 4.12$

$$P(x) = \begin{bmatrix} 0 & 3.42 & 3.16 & 5.83 \\ 3.42 & 0 & 3.16 & 3.61 \\ 3.16 & 3.16 & 0 & 4.12 \\ 5.83 & 3.61 & 4.12 & 0 \end{bmatrix} \quad \begin{array}{l} \text{Βήμα 3: min απόσταση } d_{13} = d_{31} = 3.16 \\ \text{ήτανά clusters:} \\ \left\{ \begin{array}{l} \{x_1, x_3, x_4\} \{x_2\} \{x_5\} \\ \{x_1\} \{x_2, x_3, x_4\} \{x_5\} \end{array} \right\} \end{array}$$

Ερώτημα 11

Έστω η Boolean συνάρτηση $y = x_1 \cup \neg(x_2)$, με $x_1, x_2 \in \{0,1\}$. Ποια από τις παρακάτω προτάσεις είναι σωστή;

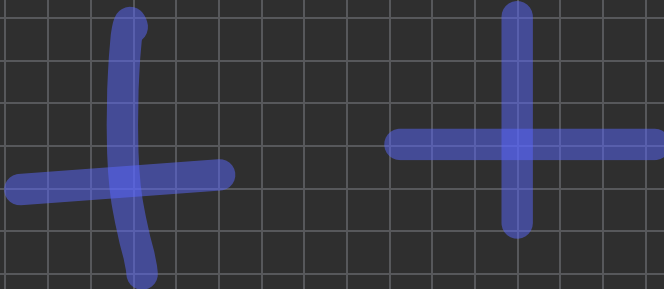
- a) Η συνάρτηση μπορεί να αναπαρασταθεί από απλό perceptron.
- b) Η συνάρτηση δεν μπορεί να αναπαρασταθεί από απλό perceptron, αλλά μπορεί να αναπαρασταθεί από πολυστρωματικό perceptron (MLP) με 1 κρυμμένο στρώμα 2 νευρώνων και στρώμα εξόδου ενός νευρώνα.
- c) Η συνάρτηση δεν μπορεί να αναπαρασταθεί από απλό perceptron, αλλά μπορεί να αναπαρασταθεί από πολυστρωματικό perceptron (MLP) με 1 κρυμμένο στρώμα 3 νευρώνων και στρώμα εξόδου ενός νευρώνα.
- d) Το συγκεκριμένο πρόβλημα δεν είναι γραμμικά διαχωρίσιμο, όμως ένα νευρωνικό δίκτυο ADALINE μπορεί να συγκλίνει επιτρέποντας ακριβώς μία λανθασμένη ταξινόμηση.



Ερώτημα 12

Έστω πολυστρωματικό perceptron (MLP) που περιέχει συνολικά 100 συναντικά βάρη (δεν περιέχει πολώσεις – bias terms). Το επίπεδο εξόδου αποτελείται από 4 εξόδους. Ως συνάρτηση κόστους χρησιμοποιείται μέσο τετραγωνικό σφάλμα (MSE). Κατά τη διάρκεια της εκπαίδευσης, το MLP επεξεργάζεται συνολικά 300 minibatches, το καθένα μεγέθους 20 δειγμάτων. Ποιο είναι το συνολικό πλήθος των ανανεώσεων βαρών που πραγματοποιούνται κατά την οπισθοδιάδοση (backpropagation);

- a) 600.000
- b) 30.000
- c) 2.000
- d) Καμία από τις παραπάνω.



Ερώτημα 13

Ποια από τις παρακάτω προτάσεις για τα Support Vector Machines (SVMs) είναι **λανθασμένη**;

SVM

- a) Η εισαγωγή των μεταβλητών χαλαρότητας (slack variables) ξ μπορεί να οδηγήσει στην εύρεση λύσης σε περιπτώσεις μη γραμμικά διαχωρίσιμων κλάσεων επιτρέποντας λανθασμένες ταξινομήσεις.
- b) Αν αφαιρέσουμε από το σύνολο δεδομένων ένα σημείο που ταξινομείται ορθά και βρίσκεται μακριά από το όριο απόφασης, τότε το όριο απόφασης (βέλτιστο υπερεπίπεδο διαχωρισμού) δεν θα επηρεαστεί.
- c) Ένα σημαντικό μειονέκτημα των SVMs είναι ότι συχνά παγιδεύονται σε τοπικά ελάχιστα, σε αντίθεση με τα MLPs.
- d) Με τη χρήση συναρτήσεων πυρήνα (kernel functions) γίνεται έμμεσα απεικόνιση των δεδομένων σε μη γραμμικό χώρο χωρίς να εμφανίζεται πουθενά στις πράξεις η συνάρτηση $M/\Sigma \Phi()$ μόνη της.

SVM (Support Vector Machines)

Χαρακτηριστικά: ▸ μεθοδολογία ML για ταξινόμηση και παλινδρόμηση

▸ εσκότζει υπερ-επιπέδο που διαχωρίζει δύο κλάσεις δεδομένων με τον καλύτερο δυνατό τρόπο, μεγιστοποιώντας το περιθώριο (margin) μεταξύ δεδομένων και υπερ-επιπέδου

Βασικά: ▸ Υπερ-επιπέδο (hyperplane): γραμ απόφαση $f(x) = w^T x + b \rightarrow y_i \cdot f(x_i) \geq 1$
σύνολο δεδομένων: (x_i, y_i) , $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

▸ Περιθώριο (margin): μέτρο-μεγιστοποίηση

▸ Σημεία στήριξης (Support Vectors):^b σημεία κοντά στο hyperplane και καθορίζουν την κατεύθυνση και το μέγεθος του. Τα υπόλοιπα σημεία δεν το επηρεάζουν.

▸ Γραμ διαχ δεδομένα: ελαχιστοποίηση $\min_{w,b} \frac{1}{2} \|w\|^2$ με περιορισμούς $y_i(w^T x_i + b) \geq 1 \quad \forall i$

▸ Μη γραμ διαχ δεδομ: ① Stack Variables (γ): προστίθενται για να επιτρέψουν λάθη ταξιν^a
$$\min_{w,b,\gamma} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \gamma_i \quad C: \text{trade-off μεταξύ μεγιστ margin και σφαλμάτων}$$

② Kernel Functions: απεικόνιση δεδομένων σε υψηλότερο χώρο διαστάσεων, όπου πιθανότερο να είναι γραμ διαχωρίσιμα.

· γραμ πυρήνα: $K(x_i, x_j) = x_i^T x_j$

· πολυωνυμ πυρήνα: $K(x_i, x_j) = (x_i^T x_j + c)^d$

· Gaussian RBF πυρήνας: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$

Pros η Cons: ④ αποτ^a σε υψηλές διαστάσεις

④ χεικ^aευση, ανθεκτικό σε overfitting

με παρκ λίσση

④ δεν παγιδεύεται σε τοπικά min^c γιατί λύνουν κυρτό πρόβλημα βελ^a

④ ευελιξία με πυρήνες: χειρίζονται εύκολα μη γραμ δεδομένα με kernel fun

⊖ υποδοχ κόστος

⊖ δύσκολη επιλογή πυρήνα και παραμέτρων

! MLP ~~αποτ^a~~ χρησιμοποιεί οπισθοπαράδοση (backpropagation, που βασίζεται σε μη κυρτά συναρ^a, μπορεί να παγιδευτεί σε τοπικά min

Ερώτημα 14

Έστω πρόβλημα ταξινόμησης σε 3 κλάσεις, το οποίο θέλουμε να επιλύσουμε με χρήση Support Vector Machines (SVMs). Έστω M το πλήθος των δυαδικών SVMs που θα πρέπει να εκπαιδευσουμε αν ακολουθήσουμε τη μέθοδο one-against-one και N το πλήθος των δυαδικών SVMs που θα πρέπει να εκπαιδευσουμε αν ακολουθήσουμε τη μέθοδο one-against-all. Για τα M, N ισχύει:

- a) $M > N$
- b) $M < N$
- c) $M = N$
- d) Δεν γνωρίζουμε.

κλάσεις $k=3$

One-against-one: $M = k(k-1)/2 = 3 \sim [1,2] [2,3] [3,1]$

One-against-all: $N = K = 3 \sim [1, (2,3)] [2, (1,3)] [3, (1,2)]$

Ερώτημα 15

Δίνονται οι παρακάτω προτάσεις σχετικά με τον ε-άπληστο αλγόριθμο στην ενισχυτική μάθηση. Επιλέξτε την πρόταση που είναι **λανθασμένη**.

- a) Στη διάρκεια της εκτέλεσής του, σε κάποιες περιπτώσεις επιλέγεται η ενέργεια με τη μέγιστη εκτιμώμενη ανταμοιβή και σε άλλες κάποια άλλη ενέργεια με τυχαίο τρόπο.
- b) Προτείνει έναν τρόπο αντιμετώπισης του μειονεκτήματος της ελλιπούς εκμετάλλευσης που διακρίνει τον απλό άπληστο αλγόριθμο.
- c) Για κάποια οριακή τιμή του ϵ εκφυλίζεται στον απλό άπληστο αλγόριθμο.
- d) Είναι δυνατό η αύξηση της εξερεύνησης σε κάποιες περιπτώσεις να μην βελτιώνει την επίδοση του αλγορίθμου.

??

ϵ -greedy: $\begin{cases} \epsilon \text{ ποσοστό πόντου} \rightarrow \text{τυχαία επιλογή (εξερεύνηση)} \\ 1-\epsilon \text{ ποσοστό πόντου} \rightarrow \text{επιλογή max ανταμοιβής (εκμετάλλ.)} \end{cases}$

Θαύμα πρόβλημα ελαφινούς εκμετάλλευσης greedy αλγ

! $\epsilon \rightarrow 0$: εκφυλίζει απλό greedy γιατί δεν κάνει εξερεύνη

Θ $\exists$ περιπτώσεις που η αύξηση της ϵ δεν βελτιώνει απόδοση αλγορίθμου
κυρίως αν οι εκτιμώμενες ανταμ είναι κοντά στη βέλτεια

Ερώτημα 16

Δίνεται ένα σύνολο δεδομένων με στοιχεία που περιγράφονται από χαρακτηριστικά που αποτιμώνται ως Αληθή ή Ψευδή και ανήκουν σε μία από τις 2 κατηγορίες εξόδου. Σχετικά με την εύρεση ενός δέντρου απόφασης που ταξινομεί σωστά τα στοιχεία του συνόλου δεδομένων ισχύει ότι:

- a) Υπάρχουν πιθανά πολλά δέντρα που ταξινομούν σωστά τα στοιχεία, αλλά δεν μπορούμε να τα βρούμε με αποδοτικό αλγόριθμο.
- b) Υπάρχει ένα μοναδικό δέντρο που ταξινομεί σωστά τα στοιχεία και μπορούμε να το βρούμε με αποδοτικό αλγόριθμο.
- c) Υπάρχουν πιθανά πολλά δέντρα που ταξινομούν σωστά τα στοιχεία και μπορούμε να τα βρούμε με αποδοτικό αλγόριθμο.
- d) Υπάρχει ένα μοναδικό δέντρο που ταξινομεί σωστά τα στοιχεία, αλλά δεν μπορούμε να το βρούμε με αποδοτικό αλγόριθμο.

??

$\exists$ πολλά σωστά δέντρα

Μπορούμε να βρούμε ένα με απόδοτικό τρόπο: CART, ID3

Η εύρεση του βέλτστου: NP-hard

Ερώτημα 17

Έστω ένας δυαδικός ταξινομητής δύο πραγματικών χαρακτηριστικών, για το οποίο γνωρίζουμε ότι μπορεί να κατακερματισθεί ένα (γενικό) σύνολο σημείων του επιπέδου, οποιασδήποτε πληθικότητας, επιλέγοντας αντίστοιχες τιμές παραμέτρων. Για τον ταξινομητή αυτό μπορούμε να αποδείξουμε ότι:

- a) Είναι PAC εκπαιδευσιμος, μόνο για το συγκεκριμένο σύνολο εισόδου.
- b) Είναι PAC εκπαιδευσιμος.
- c) Δεν είναι PAC εκπαιδευσιμος.
- d) Δεν γνωρίζουμε.

PAC-εκπαιδευσιμος

① $\forall$ κατ'ον δεδομ D , $\forall$ σφάλμα $\epsilon > 0$ και πιθαν αποστ $\delta > 0$: $\exists$ πλήθος δείγμ εκη m :
ο αλγ παράχει υπόθεση $h \in H$: $P[\text{Error}(h) \leq \epsilon] \geq 1 - \delta$ (Error: χρεικευμ σφάλμα)

② χρόνος εκη $\oplus$ αριθμ δείγμ m εταρε πολυνυμ απο:

- VC διάσταση χώρου υποθ H

- τιμές $1/\epsilon$, $1/\delta$

$\triangleright$ PAC-learnable $\Rightarrow$ ο ταξινομ μπορεί να χρεικει χαμηλό σφάλμα σε νέα δεδομ
 $\Rightarrow$ χρόνος εκη $\oplus$ πλήθος δείγμ $\Rightarrow$ πολυνυμ περιερασμ

Διάσταση VC: μετροση πολυνυμ ταξινομ

ταξινομητη διαχ τείδια η ασικεία $\Rightarrow VC \geq n$

χαμηλό VC $\left\{ \begin{array}{l} \text{μικρή εκφρ ισχύ} \\ \text{πιο ανθεκτ σε overfitting} \end{array} \right.$

SVM: PAC εκη με υψηλή χρεικευση

decision trees: χα περιωρ βαθμ δειγτρου

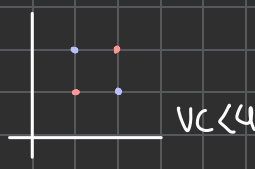
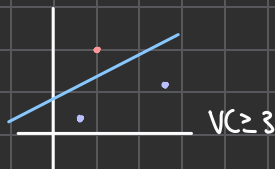
linear classifiers: (πχ perceptron) αυ τα δει είναι γραμ διαχ

Ερώτημα 20

Η διάσταση VC (VC dimension) ενός δυαδικού ταξινομητή 2 πραγματικών χαρακτηριστικών που αποτελείται από ένα απλό νευρώνα τύπου perceptron είναι:

- a) 1.
- b) 2.
- c) 3.
- d) 4.

Υπόδειξη: Για τον υπολογισμό της πληθικότητας του συνόλου σημείων που κατακερματίζονται από τον ταξινομητή, να θεωρήσετε γενική τοποθέτηση των σημείων στο επίπεδο, το οποίο σημαίνει ότι για $n = 5$, για παράδειγμα, μην θεωρήσετε ότι τα 5 σημεία είναι στην ίδια ευθεία.



$$\Rightarrow VC = 3$$

Ερώτημα 18

Σε ένα δέντρο απόφασης που έχει κατασκευαστεί από ένα σύνολο δεδομένων με τον αλγόριθμο ID3, εφαρμόζουμε τεχνική κλάδεματος επιλέγοντας μία ελάχιστη εμπιστοσύνη. Δίνονται οι παρακάτω προτάσεις:

- 1) Ενώ πριν το κλάδεμα μπορεί κάποια στοιχεία του συνόλου δεδομένων να μην ταξινομούνται σωστά, μετά το κλάδεμα θα ταξινομούνται όλα σωστά.
- 2) Ενώ πριν το κλάδεμα όλα στοιχεία του συνόλου δεδομένων ταξινομούνται σωστά, μετά το κλάδεμα μπορεί κάποια να μην ταξινομούνται σωστά.
- 3) Το δέντρο μετά το κλάδεμα έχει μικρότερο μέγεθος, αλλά χειρότερη γενίκευση.
- 4) Το δέντρο μετά το κλάδεμα έχει μικρότερο μέγεθος και καλύτερη γενίκευση.

Ποιες από αυτές είναι ορθές;

- a) 1 και 3.
- b) 2 και 3.
- c) 1 και 4.
- d) 2 και 4.

κλάδεμα $\rightarrow$ βελτίωση γενίκευσης, αποφυγή υπερπροσαρμογής
 $\rightarrow$ δημιουργεί μικρά σφάλματα στο σύνολο εκπαίδ

Ερώτημα 19

Δίνεται το παρακάτω σύνολο δεδομένων. Κατά τη διαδικασία εκτέλεσης του αλγορίθμου CART, ποιο χαρακτηριστικό επιλέγεται στη ρίζα του δέντρου απόφασης;

Υπόδειξη: Στον υπολογισμό του κέρδους πληροφορίας για ένα σύνολο D, χρησιμοποιείται το μέτρο:

$$gini(D) = \sum_{i \in \text{labels}(D)} p_i \cdot (1 - p_i) = \sum_{i \in \text{labels}(D)} p_i^2$$

Outlook	Temperature	Windy	Play Tennis
Sunny	Hot	False	No
Overcast	Hot	False	Yes
Overcast	Cool	True	Yes
Overcast	Hot	False	Yes

- a) Outlook.
- b) Temperature.
- c) Windy.
- d) Επειδή περισσότερα από 1 χαρακτηριστικά έχουν το ίδιο κέρδος πληροφορίας, θα επιλεγεί τυχαία 1 από αυτά.

$(\text{Outlook} == \text{Sunny}) \rightarrow \text{No}$ } το Outlook περιέχει την περισσότερη πληροφορία
 $(\text{Outlook} == \text{Overcast}) \rightarrow \text{Yes}$

$$\begin{aligned} \text{Outlook} \begin{cases} \text{Sunny} : (4,1) = (0,1) \\ \text{Overcast} : (4,1) = (3,0) \end{cases} & \left\{ \begin{aligned} gini(\text{Sunny}) &= 1 - (0^2 + 1^2) = 0 \\ gini(\text{Overcast}) &= 1 - (1^2 + 0^2) = 0 \end{aligned} \right. \left\{ \begin{aligned} gini(\text{Outlook}) &= \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0 = 0 \end{aligned} \right. \end{aligned}$$

$$\begin{aligned} \text{Temper} \begin{cases} \text{Hot} : (4,1) = (2,1) \\ \text{Cool} : (4,1) = (1,0) \end{cases} & \left\{ \begin{aligned} gini(\text{Hot}) &= 1 - ((2/3)^2 + (1/3)^2) = 4/9 \\ gini(\text{Cool}) &= 1 - (1^2 + 0^2) = 0 \end{aligned} \right. \left\{ \begin{aligned} gini(\text{Temp}) &= \frac{3}{4} \cdot \frac{4}{9} + \frac{1}{4} \cdot 0 = \frac{1}{3} \end{aligned} \right. \end{aligned}$$

$$\begin{aligned} \text{Windy} \begin{cases} \text{T} : (4,1) = (3,0) \\ \text{F} : (4,1) = (2,1) \end{cases} & \left\{ \begin{aligned} gini(\text{T}) &= 1 - ((2/3)^2 + (1/3)^2) = 4/9 \\ gini(\text{F}) &= 1 - (1^2 + 0^2) = 0 \end{aligned} \right. \left\{ \begin{aligned} gini(\text{Windy}) &= \frac{3}{4} \cdot \frac{4}{9} + \frac{1}{4} \cdot 0 = \frac{1}{3} \end{aligned} \right. \end{aligned}$$

$gini(\text{Outlook}) < gini(\text{Temp}) = gini(\text{Windy}) \rightarrow \downarrow gini \rightarrow \uparrow \text{κέρδος πληροφορίας} \rightarrow \text{επιλογή Outlook}$