

Θεωρούμε το πρόβλημα δυαδικής ταξινόμησης κατά Bayes σε δύο κλάσεις ω_1 και ω_2 .

Υποθέτουμε ότι οι εκ των προτέρων πιθανότητες των δύο κλάσεων είναι $P(\omega_1)$ και $P(\omega_2)$, και οι πιθανότητες παρατήρησης ενός χαρακτηριστικού $x \in \mathbb{R}^d$ στις δύο κλάσεις είναι $P(x|\omega_1)$ και $P(x|\omega_2)$, αντίστοιχα. Ποιοι από τους παρακάτω ισχυρισμούς για την ταξινόμηση κατά Bayes είναι ορθοί;

1. Ο κανόνας απόφασης για την ταξινόμηση του x στην ω_1 ή ω_2 είναι να επιλέγει η κλάση με τη μεγαλύτερη εκ των υστέρων πιθανότητα.
2. Η εκ των υστέρων πιθανότητα $P(\omega_i|x)$ είναι ανάλογη με το γινόμενο της εκ των προτέρων πιθανότητας $P(\omega_i)$ και της πιθανότητας $P(x|\omega_i)$.
3. Η υπερεπιφάνεια απόφασης είναι το σύνολο των σημείων x που οι εκ των υστέρων πιθανότητες $P(\omega_1|x)$ και $P(\omega_2|x)$ είναι ίσες.
4. Ο λόγος πιθανότητας $P(x|\omega_1)/P(x|\omega_2)$ μπορεί να χρησιμοποιηθεί από τον ταξινομητή Bayes για την ταξινόμηση του x σε μια από τις δύο κλάσεις.

①

Bayes

→ ορισμός

→ κανόνας Bayes

→ ορισμός

→ πρέπει να γνωρ
 $P(\omega_1), P(\omega_2)$

Ταξινόμητης Bayes

► Βέλτιστος κανόνας απόφασης που ελαχιστοποιεί ποσοστό σφάλματος

Ταξινόμ δεδομένο παρατηρησιακό x , στην κλάση με την μεγαλύτερη εκ των υστέρων πιθανότητα $P(\omega_i|x)$

Κανόνας απόφασης: ταξιν το x στην ω_i : $\arg\max_{\omega_i} P(\omega_i|x) = \arg\max_{\omega_i} \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

Υπερεπιφάνεια απόφασης: σύνολο σημείων x : $P(\omega_1|x) = P(\omega_2|x)$

Λόγος πιθανοφάνειας: $\frac{P(x|\omega_1)}{P(x|\omega_2)}$ χρησιμ όταν χωρίζουμε $P(\omega_1), P(\omega_2)$

⊕ είναι βέλτιστος

⊖ απαιτεί εκ των προτέρων γνώση $P(\omega_i)$ και πιθανοφάνειας $P(x|\omega_i)$

εκ των υστέρων πιθαν: $P(\omega_i|x)$
 εκ των προτέρων πιθαν: $P(x|\omega_i)$ } Bayes: $P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

Δίνονται τα σημεία: (1, 2), (2, 3), (3, 5), (4, 7) και έστω η ευθεία με εξίσωση $y = \theta_0 + \theta_1 x$ που περιγράφει με βέλτιστο τρόπο τα δεδομένα αυτά με βάση το κριτήριο ελαχίστων τετραγώνων.

Η τιμή της παραμέτρου κλίσης θ_1 θα είναι:

(A) $\theta_1 = 1.4$

(B) $\theta_1 = 1.7$

(Γ) $\theta_1 = 1.5$

(Δ) $\theta_1 = 1.8$

②

Least Squares

Least Squares

► χρησιμοποιείται για την εκτίμηση παραμ. γραμμικού μοντέλου

γραμ. μοντέλο: $y = \theta_0 + \theta_1 x \sim \bar{y} = \bar{X} \cdot \bar{\theta}$

στόχος: ελαχιστ. $J(\theta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \|y - X\theta\|^2$ ($\hat{y} = \bar{X}\bar{\theta}$, προβλ. απλή)

λύση: $\theta = (X^T X)^{-1} X^T y$

design matrix: $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$ διαν. y εξόδου: $y = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 7 \end{bmatrix}$ $\leadsto \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = (X^T X)^{-1} X^T y$
 (ανάγν. παράγεις)

$$\bar{x} = \frac{\sum_{i=1}^4 x_i}{4} = \frac{1+2+3+4}{4} = 2.5$$

$$\bar{y} = \frac{\sum_{i=1}^4 y_i}{4} = \frac{2+3+5+7}{4} = 4.25$$

σημείο	απόκλιση από (2.5, 4.25)
(1, 2)	(-1.5, -2.25)
(2, 3)	(-0.5, -1.25)
(3, 5)	(0.5, 0.75)
(4, 7)	(1.5, 2.75)

$$\theta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\theta_1 = \frac{(-1.5)(-2.25) + (-0.5)(-1.25) + 0.5 \cdot 0.75 + 1.5 \cdot 2.75}{(-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2} = 1.7$$

$$\bar{y} = \theta_1 \bar{x} + \theta_0 \Rightarrow \theta_0 = 4.25 - 1.7 \cdot 2.5 \Rightarrow \theta_0 = 0$$

Θεωρούμε $x_1, x_2, \dots, x_N \in \mathbb{R}^+$ ανεξάρτητες παρατηρήσεις που προέρχονται από τυχαίες μεταβλητές που ακολουθούν την κατανομή Weibull:

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0$$

με παραμέτρους k, λ . Με δεδομένο ότι η παράμετρος k είναι γνωστή, η εκτίμηση μέγιστης πιθανότητας της παραμέτρου λ θα είναι:

(a) $\hat{\lambda}_{MLE} = \left(\frac{1}{N} \sum_{i=1}^N x_i^k\right)^{1/k}$

(b) $\hat{\lambda}_{MLE} = \left(\frac{1}{N} \sum_{i=1}^N x_i^{1/k}\right)$

(c) $\hat{\lambda}_{MLE} = \frac{1}{N} \left(\sum_{i=1}^N x_i\right)^k$

(d) $\hat{\lambda}_{MLE} = \left(\frac{1}{N} \sum_{i=1}^N x_i^k\right)^k$

③

εκτίμηση μέγιστης πιθανότητας

συνάρτηση μέγιστης πιθανοφ.

$$L(k, \lambda) = \prod_{i=1}^N f(x_i | k, \lambda) = \prod_{i=1}^N \left\{ \frac{k}{\lambda} \left(\frac{x_i}{\lambda}\right)^{k-1} \cdot e^{-(x_i/\lambda)^k} \right\} = \left(\frac{k}{\lambda}\right)^N \left(\frac{\prod x_i}{\lambda^N}\right)^{k-1} e^{-\sum x_i^k / \lambda^k}$$

λογαρ συνάρτηση πιθανοφ:

$$\ell(k, \lambda) = \ln L(k, \lambda) = N \ln k - N \ln \lambda + (k-1) \sum \ln x_i - (k-1) N \cdot \ln \lambda - \frac{\sum x_i^k}{\lambda^k}$$

μηδενισμ παραγώγου:

$$\frac{\partial \ell(k, \lambda)}{\partial \lambda} = 0 \Rightarrow \frac{N}{\lambda} - \frac{(k-1)N}{\lambda} - \sum x_i^k \cdot (-k) \frac{1}{\lambda^{k+1}} = 0 \Rightarrow -N + \frac{\sum x_i^k}{\lambda^k} = 0 \Rightarrow \lambda = \left(\frac{\sum x_i^k}{N}\right)^{1/k}$$

Ερώτηση:

Θεωρούμε το πρόβλημα δυαδικής ταξινόμησης σε δύο κλάσεις ω_1 και ω_2 , όπου οι εκ των προτέρων πιθανότητες των δύο κλάσεων είναι $P(\omega_1) = 0.4$ και $P(\omega_2) = 0.6$. Δίνεται επιπλέον ότι οι πιθανότητες παρατήρησης ενός χαρακτηριστικού $x \in \mathbb{R}$ στις δύο κλάσεις είναι αντίστοιχα:

$$P(x|\omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad P(x|\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}.$$

Αν έχουμε μια νέα παρατήρηση $x = 0.5$, ποια είναι η εκ των υστέρων πιθανότητα $P(\omega_1|x = 0.5)$;

Απαντήσεις:

- (A) $\frac{1}{3}$
- (B) $\frac{1}{2}$
- (Γ) $\frac{2}{5}$
- (Δ) $\frac{4}{7}$

④

Κανόνας Bayes

Βλέπουμε ότι:

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

$$P(\omega_1|x=0.5) = \frac{P(\omega_1) \cdot P(x=0.5|\omega_1)}{P(\omega_1) \cdot P(x=0.5|\omega_1) + P(\omega_2) \cdot P(x=0.5|\omega_2)} = \frac{P(\omega_1)}{P(\omega_1) + P(\omega_2)} = 0.4$$

5

Decision Tree

5. Δίνεται ένα σύνολο δεδομένων με στιγμιότυπα που περιγράφονται με 3 χαρακτηριστικά που αποτιμώνται με 2, 3 και 4 τιμές, αντίστοιχα, και ανήκουν σε μία από δύο κατηγορίες εξόδου.

Έστω ότι χρησιμοποιείται ο αλγόριθμος ID3 για την εύρεση ενός δέντρου απόφασης. Ισχύει ότι:

(Α) Το μέγιστο ύψος του δέντρου απόφασης που θα υπολογιστεί είναι 4 και το ελάχιστο είναι 2.

(Β) Το μέγιστο ύψος του δέντρου απόφασης που θα υπολογιστεί είναι 4 και το ελάχιστο είναι 4.

(Γ) Το μέγιστο ύψος του δέντρου απόφασης που θα υπολογιστεί είναι 9 και το ελάχιστο είναι 2.

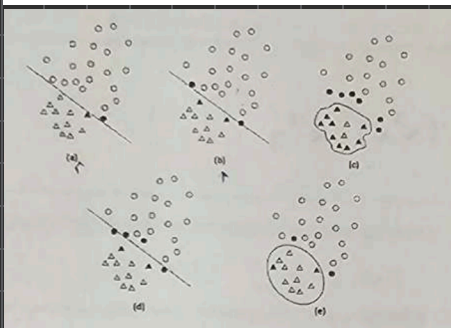
(Δ) Το μέγιστο ύψος του δέντρου απόφασης που θα υπολογιστεί είναι 9 και το ελάχιστο είναι 4.

μέγιστο: καρ 1
 αθροισι καρ 2
 όλα τα καρ 3
 χαρακι απόφαση

ελάχιστο: καρ i
 αθροισι μόνο απόφαση
 ένα χαρακ

ύψος 4
 ύψος 2

Στο Σχήμα 1 φαίνονται τα όρια απόφασης πέντε SVM που προκύπτουν από διαφορετικούς πυρήνες (kernels) ή/και διαφορετικές τιμές ποινής χαλαρότητας C . Οι δύο κλάσεις δεδομένων αναπαρίστανται με τρίγωνα και κύκλους αντίστοιχα, ενώ τα συμπαγή (σκούρα) τρίγωνα και κύκλοι αναπαριστούν διανύσματα υποστήριξης.



C : κόστος λανθ ταξινόμ

$C=0$: δεν μας ενδιαφ οι λανθ ταξινόμησεις

$C \rightarrow \infty$: έμφαση σε σωστή ταξινόμ ηρωτώνων

6. Ποιο από τα παρακάτω διαγράμματα του Σχήματος 1 αντιστοιχεί σε γραμμικό SVM με $C = 0.1$;

- (Α) Το διάγραμμα (a).
- (Β) Το διάγραμμα (b).
- (Γ) Το διάγραμμα (c).
- (Δ) Το διάγραμμα (d).

μικρό C : επερ σφαλματα $\rightarrow d$

7. Ποιο από τα παρακάτω διαγράμματα του Σχήματος 1 αντιστοιχεί σε γραμμικό SVM με $C = 5$;

- (Α) Το διάγραμμα (b).
- (Β) Το διάγραμμα (c).
- (Γ) Το διάγραμμα (d).
- (Δ) Το διάγραμμα (e).

μεγάλο C : λίγα σφάλματα $\rightarrow a, b$

8. Ποιο από τα παρακάτω διαγράμματα του Σχήματος 1 αντιστοιχεί σε SVM με πυρήνα $K(u, v) = (u^T v)^2$;

- (Α) Το διάγραμμα (b).
- (Β) Το διάγραμμα (c).
- (Γ) Το διάγραμμα (d).
- (Δ) Το διάγραμμα (e).

kernel Φ_m : μη γραμ διαχωρ

Δίνονται οι παρακάτω προτάσεις:

(Α) Σε ένα κάθε ελάχιστο δέντρο απόφασης για ένα σύνολο δεδομένων, όλα τα φύλλα βρίσκονται στο ίδιο ύψος.

(Β) Υπάρχουν εκθετικά πολλά διαφορετικά δέντρα απόφασης, σε σχέση με το πλήθος των χαρακτηριστικών των στιγμιτύπων.

(Γ) Κάθε δέντρο απόφασης έχει εκθετικά πολλούς διαφορετικούς κόμβους, σε σχέση με το πλήθος των χαρακτηριστικών των στιγμιτύπων.

Πόσες από τις παραπάνω προτάσεις ισχύουν;

- (Α) Καμία
- (Β) 1
- (Γ) 2
- (Δ) 3

9

Decision Tree

Δίνονται ένα σύνολο δεδομένων για το οποίο υπολογίζουμε ένα τυχαίο δάσος (random forest) με τη μέθοδο της τμηματοποίησης του συνόλου δεδομένων.

Α. Τα δέντρα απόφασης του τυχαίου δάσους είναι γενικά μικρότερου ύψους από αντίστοιχα δέντρα απόφασης που θα υπολογίζαμε αν δεν τμηματοποιούσαμε το σύνολο δεδομένων.

Β. Τα δέντρα απόφασης του τυχαίου δάσους είναι γενικά μεγαλύτερου ύψους από αντίστοιχα δέντρα απόφασης που θα υπολογίζαμε αν δεν τμηματοποιούσαμε το σύνολο δεδομένων.

Γ. Τα δέντρα απόφασης του τυχαίου δάσους είναι γενικά ίδιου ύψους με αντίστοιχα δέντρα απόφασης που θα υπολογίζαμε αν δεν τμηματοποιούσαμε το σύνολο δεδομένων.

Δ. Καμία από τις άλλες προτάσεις δεν είναι σωστή.

10

Random Forest

Δίνεται το σύνολο δεδομένων D του Πίνακα 1. Κατά τη διαδικασία εκτέλεσης του αλγορίθμου CART, ποιο χαρακτηριστικό επιλέγεται στη ρίζα του δέντρου απόφασης; [Δίνεται ότι: $\text{gini}(D) = 1 - \sum_{i \in \text{labels}(D)} p_i^2$]

(Α) Το X_1 .

(Β) Το X_2 .

(Γ) Επειδή περισσότερα από ένα χαρακτηριστικά έχουν το ίδιο κέρδος πληροφορίας, ο αλγόριθμος δεν θα τερματίσει.

(Δ) Επειδή περισσότερα από ένα χαρακτηριστικά έχουν το ίδιο κέρδος πληροφορίας, θα επιλεγεί τυχαία ένα από αυτά.

Πίνακας 1: Σύνολο δεδομένων D

X_1	X_2	Y
TRUE	TRUE	No
TRUE	FALSE	Yes
FALSE	TRUE	Yes
FALSE	FALSE	No

11

Decision Tree

X_1, X_2 :

True $\rightarrow (Y, N) = (1, 1)$

False $\rightarrow (Y, N) = (1, 1)$

} $\Rightarrow \text{Gini}(X_1) = \text{Gini}(X_2) \rightarrow \text{τυχαία επιλογή}$

Ποιός από τους παρακάτω ισχυρισμούς είναι λανθασμένος σχετικά με τους αλγορίθμους ομαδοποίησης k-means και DBSCAN;

(Α) Ο DBSCAN μπορεί να ομαδοποιήσει αποτελεσματικά δεδομένα με διαφορετικές πυκνότητες, ενώ ο k-means όχι.

(Β) Ο k-means είναι πιο αποτελεσματικός από τον DBSCAN για δεδομένα με θόρυβο και ακραία σημεία (outliers).

(Γ) Ο k-means είναι ευαίσθητος στην αρχικοποίηση των κέντρων των κλάσεων.

(Δ) Ο k-means υποθέτει ότι οι κλάσεις είναι σφαιρικές, ενώ ο DBSCAN όχι.

12

k-means, DBSCAN

Ιδιότητα	k-means	DBSCAN
Υποθέσεις για τις κλάσεις	Οι κλάσεις είναι σφαιρικές και ισομεγέθεις.	Δεν υπάρχουν υποθέσεις για το σχήμα των κλάσεων.
Ευαισθησία σε outliers	Πολύ ευαίσθητος. Τα outliers μπορούν να αλλοιώσουν σημαντικά τα κέντρα των κλάσεων.	Ανθεκτικός σε outliers, καθώς τα αγνοεί όταν δεν ανήκουν σε περιοχές υψηλής πυκνότητας.
Μέθοδος λειτουργίας	Ομαδοποιεί τα δεδομένα γύρω από κεντρικά σημεία (centroids) με βάση την ελαχιστοποίηση της απόστασης.	Βασίζεται στην πυκνότητα των σημείων σε μια περιοχή για να εντοπίσει κλάσεις.
Πλήθος κλάσεων (clusters)	Χρειάζεται να δοθεί εκ των προτέρων ο αριθμός των κλάσεων.	Δεν απαιτεί προκαθορισμένο αριθμό κλάσεων, καθώς τις προσδιορίζει δυναμικά.
Ευαισθησία στην αρχικοποίηση	Πολύ ευαίσθητος στην αρχικοποίηση των κέντρων.	Όχι ευαίσθητος σε αρχικοποίηση.
Ευαισθησία σε θόρυβο	Ευαίσθητος σε θορυβώδη δεδομένα.	Μπορεί να αγνοήσει σημεία θορύβου (outliers).
Πυκνότητα δεδομένων	Δεν λαμβάνει υπόψη την πυκνότητα δεδομένων.	Εστιάζει στην πυκνότητα δεδομένων και ανιχνεύει αραιές και πυκνές περιοχές.
Καταλληλότητα για μεγάλα δεδομένα	Γρήγορος και κατάλληλος για μεγάλα σύνολα δεδομένων. ↓	Μπορεί να γίνει αργός για μεγάλα σύνολα δεδομένων με υψηλές διαστάσεις.
Περίπτωση μη γραμμικών ορίων κλάσεων	Δεν μπορεί να διαχειριστεί μη γραμμικά όρια.	Μπορεί να διαχειριστεί μη γραμμικά όρια μεταξύ κλάσεων.
Υπολογιστική πολυπλοκότητα	$O(n \cdot k \cdot \ell)$, όπου n τα δεδομένα, k οι κλάσεις, και ℓ οι επαναλήψεις.	$O(n^2)$ (μπορεί να μειωθεί με κατάλληλες δομές δεδομένων όπως R-trees).
Χρήση σε πραγματικές εφαρμογές	Πολύ δημοφιλής λόγω της απλότητας και της ταχύτητας.	Χρήσιμος για δεδομένα με θόρυβο ή μη σφαιρικές κλάσεις, αλλά πιο σύνθετος στη ρύθμιση.

13. Σας δίνονται τα εξής $N = 4$ σημεία x_i :

$\{(2, 3), (3, 3), (4, 5), (5, 7)\}$.

Το ποσοστό της διασποράς των δεδομένων αυτών που "ερμηνεύεται" από την κύρια συνιστώσα της ανάλυσης PCA προκύπτει ότι είναι περίπου ίσο με:

- (Α) 97.5%
- (Β) 93.5%
- (Γ) 89.5%
- (Δ) 85.5%

Σημείωση:

Δίνεται ότι ο αμερόληπτος εκτιμητής του πίνακα συνδιακύμανσης είναι:

$$C_{unbiased} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T$$

13 PCA

$$\bar{x} = \frac{\sum x_i}{N} = (3.5, 4.5)$$

$$(2, 3) \rightarrow (-1.5, -1.5)$$

$$(3, 3) \rightarrow (-0.5, -1.5)$$

$$(4, 5) \rightarrow (0.5, 0.5)$$

$$(5, 7) \rightarrow (1.5, 2.5)$$

$$C_{unbiased} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N-1} \sum_{i=1}^N \begin{bmatrix} (x_i - \bar{x})^2 & (x_i - \bar{x})(y_i - \bar{y}) \\ (y_i - \bar{y})(x_i - \bar{x}) & (y_i - \bar{y})^2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 5 & 7 \\ 7 & 11 \end{bmatrix}$$

$$\sum (x_i - \bar{x})^2 = (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 = 5$$

$$\sum (y_i - \bar{y})^2 = (-1.5)^2 + (-1.5)^2 + 0.5^2 + 2.5^2 = 11$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (-1.5)(-1.5) + 0.5(-1.5) + 0.5(0.5) + 1.5(2.5) = 7$$

$$C_{unbiased} = \begin{bmatrix} 5/3 & 7/3 \\ 7/3 & 11/3 \end{bmatrix}$$

$$\det(C - \lambda I) = 0 \Rightarrow \begin{vmatrix} 5/3 - \lambda & 7/3 \\ 7/3 & 11/3 - \lambda \end{vmatrix} = 0 \Rightarrow (5 - 3\lambda)(11 - 3\lambda) - 49 = 0 \Rightarrow \begin{cases} \lambda_1 = (8 + \sqrt{36})/3 \\ \lambda_2 = (8 - \sqrt{36})/3 \end{cases}$$

$$\text{ποσοστό διασποράς} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{8 + \sqrt{36}}{16} = 96.7\% \quad + +$$

14. Ποιοι από τους παρακάτω ισχυρισμούς είναι λανθασμένοι σχετικά με τον k-NN αλγόριθμο;

(I) Ο k-NN δεν χρειάζεται φάση εκπαίδευσης.

(II) Ο k-NN είναι ένας μη-επιβλεπόμενος αλγόριθμος μάθησης.

(III) Στον k-NN το σύνολο απόφασης είναι μη-γραμμικό.

(IV) Το πλεονέκτημα του k-NN είναι η χαμηλή αριθμητική του πολυπλοκότητα στη φάση πρόβλεψης.

Επιλογές:

- (Α) I και III
- (Β) I και IV
- (Γ) II και III
- (Δ) II και IV
- (Ε) I και II
- (Ζ) III και IV

14

k-NN

▷ Δεν απαιτεί εκπαίδευση

▷ Supervised $\rightarrow$ απαιτεί labels

▷ Ίσχυο αποφαση δεν είναι απαραίτητα γραμμικό

Θ μεγάλη υπολογιστική πολυπλοκότητα

Θ ευαισθησία σε θόρυβο

Θ κρισιμο k: μετ k $\sim$ underfitting, μικρο k $\sim$ overfitting

15. Θεωρούμε μια παραλλαγή του αλγορίθμου k-means, όπου στο πρώτο βήμα του αλγορίθμου, αντί της Ευκλείδειας απόστασης, χρησιμοποιείται η ακόλουθη μετρική στο $\mathbb{R}^m$:

$$d(x_i, \theta_j) = |x_{im} - \theta_{jm}| + \frac{1}{1/2 + 1} \sum_{t=1, t \neq m}^l |x_{it} - \theta_{jt}|$$

όπου:

$$|x_{im} - \theta_{jm}| = \max_{m=1,2,\dots,l} |x_{im} - \theta_{jm}|$$

Τα δεδομένα είναι τα εξής:

$$x_1 = (0, 1), x_2 = (1, 2), x_3 = (2, 1), x_4 = (0, -1), x_5 = (0, -2)$$

Οι αρχικές θέσεις των κέντρων δίνονται ως:

$$\theta_1^{(0)} = (-1, 0) \quad \text{και} \quad \theta_2^{(0)} = (2, 0)$$

Εφαρμόζουμε τον παραπάνω αλγόριθμο για τον διαχωρισμό των σημείων σε δύο ομάδες. Αν τα κέντρα των ομάδων μετά την ολοκλήρωση της δεύτερης επανάληψης είναι:

$$(A) \theta_1^{(2)} = (0, -2/3) \quad \text{και} \quad \theta_2^{(2)} = (3/2, 3/2)$$

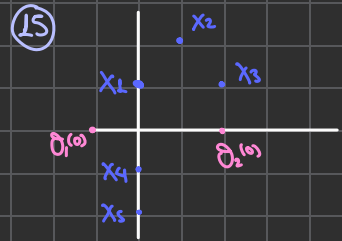
$$(B) \theta_1^{(2)} = (0, -3/2) \quad \text{και} \quad \theta_2^{(2)} = (1, 4/3)$$

$$(C) \theta_1^{(2)} = (2/3, 1/3) \quad \text{και} \quad \theta_2^{(2)} = (1/2, 0)$$

$$(D) \theta_1^{(2)} = (0, -1/2) \quad \text{και} \quad \theta_2^{(2)} = (1/2, 3/2)$$



Ποια θα είναι οι θέσεις των δύο κέντρων μετά την ολοκλήρωση της δεύτερης επανάληψης:



1 επανάληψη:

κλάση 1: $\{x_1, x_4, x_5\}$

κλάση 2: $\{x_2, x_3\}$

$$\rightarrow \theta_1^{(1)} = (0, -2/3)$$

$$\rightarrow \theta_2^{(1)} = (3/2, 3/2)$$

→ ομοίως και στην 2^η

16. Γιατί είναι σημαντικό να κανονικοποιούμε τα δεδομένα πριν εφαρμόσουμε τη μέθοδο Principal Component Analysis (PCA);

- (A) Για να εξασφαλίσουμε ότι τα δεδομένα θα ακολουθούν την κανονική κατανομή.
- (B) Για να εξασφαλίσουμε ότι όλα τα χαρακτηριστικά θα έχουν την ίδια μέση τιμή.
- (Γ) Για να εξασφαλίσουμε ότι όλα τα χαρακτηριστικά θα έχουν την ίδια συνεισφορά στην ανάλυση.
- (Δ) Για να εξασφαλίσουμε ότι τα χαρακτηριστικά θα είναι γραμμικά ανεξάρτητα.

(16)

PCA

Στόχοι: ανάδειξη κατευθύνσεων με μέγιστη διακύμανση
μείωση διασποράς δεδομ., κρατώντας πληροφ.ορία

→ απαιτείται κανονικοποίηση ⇒ όλα τα χαρακτηρ συνεισφέρουν ισότιμα

Δίνεται το σύνολο δεδομένων

$$X = \{(1, 2), (2, 3), (4, 4), (8, 7), (9, 6), (7, 8)\}$$

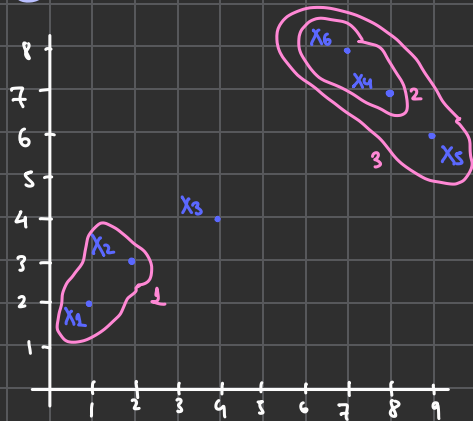
στο $\mathbb{R}^2$. Χρησιμοποιούμε τον ιεραρχικό αλγόριθμο πλήρους δεσμού για την ομαδοποίηση των δεδομένων αυτών. Ποια ομαδοποίηση έχει προκύψει μετά την ολοκλήρωση του τρίτου βήματος του αλγορίθμου;

Απαντήσεις:

- (Α) $\{(1, 2), (2, 3), (4, 4)\}, \{(8, 7)\}, \{(9, 6)\}, \{(7, 8)\}$
- (Β) $\{(1, 2)\}, \{(2, 3)\}, \{(4, 4)\}, \{(7, 8)\}, \{(8, 7), (9, 6)\}$
- (Γ) $\{(1, 2), (2, 3), (4, 4)\}, \{(7, 8), (8, 7)\}, \{(9, 6)\}$
- (Δ) $\{(1, 2), (2, 3)\}, \{(4, 4)\}, \{(7, 8), (8, 7), (9, 6)\}$



(17)



18. Ποια από τις παρακάτω προτάσεις για τη μέθοδο ελαχίστων τετραγώνων είναι εσφαλμένη;

- (Α) Το διάνυσμα των βαρών προσδιορίζεται επακριβώς (λύση κλειστής μορφής).
- (Β) Έχει μεγάλη ευαισθησία σε έκτοπες τιμές (outliers).
- (Γ) Χρησιμοποιείται για τον προσδιορισμό των χαρακτηριστικών γραμμικών συναρτήσεων απόφασης για προβλήματα δυαδικής ταξινόμησης.

(Δ) Βασίζεται στην υπόθεση ότι τα χαρακτηριστικά των δεδομένων υπόκεινται στην κανονική κατανομή.

(18)

Least Squares

→ τα σφάλματα,
όχι τα χαρακτηριστικά

Στόχος: ▸ εύρεση βαρών w που ελαχιστοποιούν τετραγ σφάλμα

▸ λύση κλειστής μορφής $w = (X^T X)^{-1} X^T y$

Ιδιότητες: ευαισθησία σε outliers

χρησιμ για γραμμικά δεδομένα

θεωρεί ότι τα σφάλμα ακολουθούν καν κατανομή

δεν λειτουργεί καλά σε υψηλές διασπασεις

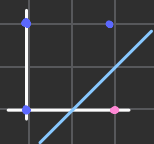
19. Έστω η boolean συνάρτηση $y = (\neg x_1) \vee x_2$, με $x_1, x_2 \in \{0, 1\}$. Ποια από τις παρακάτω προτάσεις είναι σωστή;

- (Α) Η συνάρτηση δεν μπορεί να αναπαρασταθεί από απλό perceptron, αλλά μπορεί να αναπαρασταθεί από πολυεπίπεδο perceptron (MLP) με ένα κρυμμένο στρώμα δύο νευρώνων και στρώμα εξόδου ενός νευρώνα.
- (Β) Η συνάρτηση δεν μπορεί να αναπαρασταθεί από απλό perceptron, αλλά μπορεί να αναπαρασταθεί από πολυεπίπεδο perceptron (MLP) με δύο κρυμμένα στρώματα δύο νευρώνων και στρώμα εξόδου ενός νευρώνα.
- (Γ) Το συγκεκριμένο πρόβλημα δεν είναι γραμμικά διαχωρίσιμο, όμως ένα νευρωνικό δίκτυο ADALINE μπορεί να συγκλίνει.
- (Δ) Η συνάρτηση μπορεί να αναπαρασταθεί από απλό perceptron.

19

Perceptron

Class 1: (0,0) (0,1) (1,1) } γραμ διαχωρισιμα $y = \text{sign}(\omega_1 x_1 + \omega_2 x_2 + b)$
 Class 0: (1,0)



$$x=(0,0), y=1: b > 0$$

$$x=(0,1), y=1: \omega_2 + b > 0$$

$$x=(1,0), y=0: \omega_1 + b < 0$$

$$x=(1,1), y=1: \omega_1 + \omega_2 + b > 0$$

$$\left. \begin{array}{l} x=(0,0), y=1: b > 0 \\ x=(0,1), y=1: \omega_2 + b > 0 \\ x=(1,0), y=0: \omega_1 + b < 0 \\ x=(1,1), y=1: \omega_1 + \omega_2 + b > 0 \end{array} \right\} y = \text{sign}(-x_1 + x_2 + 0.5)$$

20

SVM

20. Ποια από τις παρακάτω προτάσεις για τα Support Vector Machines (SVMs) είναι εσφαλμένη;

- (Α) Αν απομακρύνουμε από το σύνολο δεδομένων ένα σημείο που ταξινομείται σωστά και βρίσκεται μακριά από το όριο απόφασης, τότε το όριο απόφασης (βέλτιστο υπερεπίπεδο διαχωρισμού) δεν θα επηρεαστεί.
- (Β) Εισάγοντας συναρτήσεις πυρήνα (kernel functions) γίνεται έμμεσα απεικόνιση των δεδομένων σε μη γραμμικό χώρο χωρίς να εμπλέκεται πουθενά στις πράξεις η συνάρτηση μετασχηματισμού $\Phi()$ μόνη της.
- (Γ) Η προσθήκη των μεταβλητών χαλαρότητας (slack variables) ξ μπορεί να οδηγήσει στην εύρεση λύσης σε περιπτώσεις μη γραμμικά διαχωρίσιμων κλάσεων επιτρέποντας και εσφαλμένες ταξινομήσεις.
- (Δ) Καμία από τις παραπάνω.