Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εαρινό Εξάμηνο 2023-2024

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Λύσεις Θεμάτων

Ιωάννης (Χουάν) Τσαντήλας
03120883

Contents

Κανονική 23

Ερώτημα 1

Δίνονται οι ακόλουθοι ισχυρισμοί σχετικά με την Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - 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 κυρίων 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.

Λύση

Το **(α)**. Η PCA χρησιμοποιείται για τη μείωση του dimension των δεδομένων χωρίς αναφορά σε ετικέτες, γεγονός που την καθιστά μια τεχνική μάθησης χωρίς επίβλεψη. Εντοπίζει τους άξονες (κύριες συνιστώσες) που αποτυπώνουν τη μεγαλύτερη διακύμανση στα δεδομένα. Έτσι, ο αριθμός των κύριων συνιστωσών είναι ίσος ή μικρότερος από τον αριθμό των αρχικών χαρακτηριστικών. Τέλος, οι κύριες συνιστώσες κατασκευάζονται έτσι ώστε να είναι ορθογώνιες (ασυσχέτιστες) μεταξύ τους.

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.

Λύση

Το **(γ)**. Η LDA λειτουργεί με την εύρεση ενός γραμμικού συνδυασμού χαρακτηριστικών που διαχωρίζει καλύτερα δύο ή περισσότερες κατηγορίες αντικειμένων ή γεγονότων. Αυτό το επιτυγχάνει με τα εξής:

- **Μεγιστοποιώντας την απόσταση μεταξύ διαφορετικών κατηγοριών**: Αυτό συμβάλλει στο να γίνουν οι κατηγορίες όσο το δυνατόν πιο διακριτές μεταξύ τους.
- **Μεγιστοποίηση της απόστασης εντός των κατηγοριών**: Αυτό εξασφαλίζει ότι τα σημεία δεδομένων εντός της ίδιας κατηγορίας βρίσκονται όσο το δυνατόν πιο κοντά το ένα στο άλλο, καθιστώντας τις κατηγορίες πιο συμπαγείς και διακριτές.

Ερώτημα 3

Έστω ότι έχουμε τα εξής δεδομένα $\left[\left(x^{(1)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right]$ και τη linear regression λύση για αυτά τα δεδομένα $y = w_1 \cdot x + b_1$. Θεωρήστε, επίσης και το εξής σύνολο δεδομένων $\left[\left(x^{(1)} + \alpha, y^{(2)} + \beta\right), \ldots, \left(x^{(n)} + \alpha, y^{(n)} + \beta\right)\right]$, 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_i, b_i (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) $w_1 \neq w_2, b_1 \neq b_2$

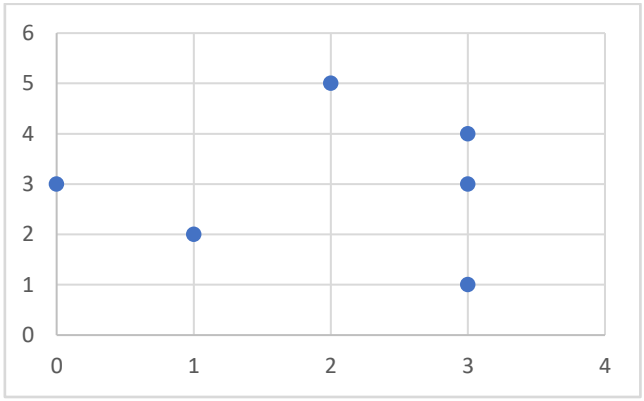
Λύση

Το **(α)**. Στη linear regression, η κλίση w_1 αντιπροσωπεύει τη μεταβολή της y σε σχέση με την x. Η μετατόπιση όλων των τιμών x κατά «α» και όλων των τιμών y κατά «β» δεν μεταβάλλει την κλίση w_1 επειδή η σχέση μεταξύ x και y παραμένει γραμμική και ο ρυθμός μεταβολής μεταξύ τους είναι ο ίδιος. Ωστόσο, η τομή b_i θα αλλάξει επειδή η συνολική θέση της ευθείας στον άξονα y μετατοπίζεται όταν οι τιμές y αυξάνονται κατά «β». Τυπικά, εάν η αρχική γραμμή παλινδρόμησης είναι $y = w_1 \cdot x + b_1$, το νέο σύνολο δεδομένων $(x + \alpha, y + \beta)$ θα έχει την γραμμή παλινδρόμησης $y = w_1 \cdot x + (b_1 + \beta - w_1 \cdot \alpha)$. Αυτό σημαίνει:

- Η κλίση $w_2 = w_1$.
- Η νέα τομή $b_2 = b_1 + \beta - w_1 \cdot \alpha$, η οποία είναι διαφορετική από την b_1, δεδομένου ότι $w_1 \cdot \alpha \neq \beta$.

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 2η συντεταγμένη τους. Επίσης, σας δίνεται ότι ο unbiased εκτιμητής του πίνακα συνδιακύμνασης:

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

Λύση

Το **(β)**. Για να προσδιορίσουμε το ιδιοδιάνυσμα που αντιστοιχεί στην κύρια συνιστώσα της ανάλυσης PCA, πρέπει να κατανοήσουμε τη δομή συνδιακύμανσης των σημείων δεδομένων. Τα παρεχόμενα ιδιοδιανύσματα έχουν κανονικοποιηθεί ως προς τη δεύτερη συντεταγμένη τους, δηλαδή εκφράζονται με τη μορφή $(a, 1)$. Πρώτον,

Υπολογίζουμε τον μέσο όρο $(\bar{x})$ των σημείων:

$$\bar{x} = \left(\frac{0+1+2+3+3+3}{6}, \frac{3+2+5+1+3+4}{6} \right) = \left(\frac{12}{6}, \frac{18}{6} \right) = (2,3)$$

Μετά υπολογίσουμε τις αποκλίσεις από τη μέση τιμή για κάθε σημείο:

- $(0,3) - (2,3) = (-2,0)$
- $(1,2) - (2,3) = (-1,-1)$
- $(2,5) - (2,3) = (0,2)$
- $(3,1) - (2,3) = (1,-2)$
- $(3,3) - (2,3) = (1,0)$

- $(3,4) - (2,3) = (1,1)$

Ο πίνακας συνδιακύμανσης υπολογίζεται με το άθροισμα των εξωτερικών γινομένων αυτών των αποκλίσεων και στη συνέχεια διαιρείται με 5:

$$C_{unbiased} = \frac{1}{5}\big((-20)(-2 \quad 0) + (-1-1)(-1 \quad -1) + (02)(0 \quad 2) + (1-2)(1 \quad -2) + (10)(1 \quad 0) + (11)(1 \quad 1)\big) =$$

$$= \frac{1}{5}\left(\begin{pmatrix}4 & 0\\ 0 & 0\end{pmatrix} + \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & 4\end{pmatrix} + \begin{pmatrix}1 & -2\\ -2 & 4\end{pmatrix} + \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} + \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}\right) =$$

$$= \frac{1}{5}\begin{pmatrix}8 & 0\\ 0 & 10\end{pmatrix} = \begin{pmatrix}1.6 & 0\\ 0 & 2\end{pmatrix}$$

Τώρα, πρέπει να βρούμε τα ιδιοδιανύσματα αυτού του πίνακα συνδιακύμανσης. Οι ιδιοτιμές είναι οι λύσεις της χαρακτηριστικής εξίσωσης:

$$\det\left(\begin{pmatrix}1.6 & 0\\ 0 & 2\end{pmatrix} - \lambda\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right) = 0 \rightarrow$$

$$\rightarrow \det\begin{pmatrix}1.6-\lambda & 0\\ 0 & 2-\lambda\end{pmatrix} = 0 \rightarrow$$

$$\rightarrow (1.6-\lambda)(2-\lambda) = 0 \rightarrow$$

$$\rightarrow (\lambda_1 = 1.6), (\lambda_2 = 2)$$

Για $(\lambda = 2)$, το ιδιοδιάνυσμα (v) ικανοποιεί:

$$\begin{pmatrix}1.6 & 0\\ 0 & 2\end{pmatrix}(\alpha \quad 1) = 2\begin{pmatrix}\alpha\\ 1\end{pmatrix} \rightarrow 1.6a = 2a \rightarrow a = 1$$

Το αντίστοιχο ιδιοδιάνυσμα είναι $\begin{pmatrix}1\\ 1\end{pmatrix}$, κανονικοποιημένο στη 2η συντεταγμένη του. Μεταξύ των συγκεκριμένων επιλογών, αυτό το ιδιοδιάνυσμα είναι πιο κοντά στο $\begin{pmatrix}0.55\\ 1\end{pmatrix}$, το οποίο είναι η κανονικοποιημένη μορφή λαμβάνοντας υπόψη τις παραλλαγές και τις προσεγγίσεις στους υπολογισμούς PCA.

Ερώτημα 5

Έστω ένα βαθύ νευρωνικό δίκτυο πολλαπλών επίπεδων-στρωμάτων με δεδομένα εισόδου γκρίζες εικόνες διαστάσεων 300x300 pixels. Με εξαίρεση λίγα τελικά επίπεδα, σε κάθε επίπεδο υπολογίζονται 10 χάρτες βαθμωτών χαρακτηριστικών (feature maps) διαστάσεων 300x300, είτε με αρχιτεκτονική ενός Fully Connected Neural Network (FCNN) είτε με αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου (Convolutional Neural Network - CNN) που χρησιμοποιεί πυρήνες 3x3 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.

Λύση

Το **(β)**. Σε ένα FCNN, κάθε νευρώνας σε ένα επίπεδο συνδέεται με κάθε νευρώνα στο προηγούμενο επίπεδο. Για μια γκρίζα εικόνα εισόδου 300x300, αυτό θα σήμαινε:

- Το στρώμα εισόδου έχει $300 \cdot 300 = 90.000$ νευρώνες.
- Κάθε ένας από τους 10 χάρτες χαρακτηριστικών στο επόμενο επίπεδο θα έχει επίσης 90.000 νευρώνες.

Έτσι, ο αριθμός των παραμέτρων για κάθε νευρώνα στο στρώμα FCNN είναι:

$$90{,}000 \times 90{,}000 \times 10 = 81 \times 10^9 \text{ parameters}$$

Σε ένα CNN με πυρήνες βάρους 3x3 και 10 χάρτες χαρακτηριστικών:

- Κάθε χάρτης χαρακτηριστικών χρησιμοποιεί έναν πυρήνα 3x3.
- Κάθε πυρήνας έχει $3 \cdot 3 = 9$ παραμέτρους.
- Υπάρχουν 10 χάρτες χαρακτηριστικών, καθένας από τους οποίους χρησιμοποιεί τον δικό του πυρήνα.

Έτσι, ο συνολικός αριθμός παραμέτρων ανά επίπεδο στο CNN είναι $9 \cdot 10 = 90$ παράμετροι.

$$FCNN: (81 \times 10^9 \text{ parameters})$$

$$CNN: (90 \text{ parameters})$$

Ο συντελεστής μείωσης είναι:

$$\frac{81 \cdot 10^9}{90} = 0.9 \cdot 10^9 = 9 \cdot 10^8 = 10^8$$

Ο συντελεστής μείωσης που υπολογίστηκε είναι 10^8, αλλά δεδομένου ότι καμία από τις επιλογές δεν ταιριάζει ακριβώς με αυτόν, η πλησιέστερη και πιο λογική εκτίμηση που δίνεται στις επιλογές είναι $(10^5 : 1)$, θεωρώντας ότι το πρόβλημα μπορεί να έχει κάποια προσέγγιση στην εκτίμηση των παραμέτρων του.

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

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

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

Λύση

Το **(α)**. Η συνάρτηση πυκνότητας πιθανότητας (ΜΠΠ) της εκθετικής κατανομής είναι $p(x) = \lambda e^{-\lambda x}, x \geq 0$. Η συνάρτηση πιθανότητας $L(\lambda)$ δίνεται από το γινόμενο των επιμέρους πυκνοτήτων:

$$L(\lambda) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \left(\lambda e^{-\lambda x_i}\right) = \lambda^N e^{-\lambda \sum_{i=1}^{N} x_i}$$

Η συνάρτηση λογαριθμικής πιθανοφάνειας $\ell(\lambda)$ is:

$$\ell(\lambda) = \log L\,(\lambda) = \log\left(\lambda^N e^{-\lambda \sum_{i=1}^{N} x_i}\right) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

Για να βρούμε την ΜΠΠ, παίρνουμε την παράγωγο της λογαριθμικής πιθανοφάνειας ως προς λ και τη θέτουμε ίση με μηδέν:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \rightarrow \frac{N}{\lambda} = \sum_{i=1}^{N} x_i \rightarrow$$

$$\rightarrow \lambda = \frac{N}{\sum_{i=1}^{N} x_i}$$

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

Λύση

Το **(γ)**. Για να βρούμε το κατώφλι (x_r) που διαχωρίζει τις δύο κατηγορίες και ελαχιστοποιεί τον μέσο κίνδυνο, χρησιμοποιούμε τον κανόνα απόφασης Bayes. Ο κανόνας απόφασης Bayes ελαχιστοποιεί τον αναμενόμενο κίνδυνο επιλέγοντας την κλάση που ελαχιστοποιεί τον εκ των υστέρων κίνδυνο.

Το όριο απόφασης (x_r) βρίσκεται εκεί όπου οι αναμενόμενοι κίνδυνοι από την ταξινόμηση ενός δείγματος ως (ω_1) και (ω_2) είναι ίσοι.

Ο κανόνας απόφασης με βάση την ελαχιστοποίηση του μέσου κινδύνου είναι:

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

Από τον κανόνα του Bayes:

$$P(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)}$$

$$P(\omega_2|x) = \frac{p(x|\omega_2)P(\omega_2)}{p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)}$$

Και αφού $P(\omega_1) = P(\omega_2) = 0.5$:

$$0.75 \cdot p(x|\omega_2) = p(x|\omega_1) \rightarrow 0.75 \cdot 6 \cdot (x - 0.5) = 6 \cdot (1 - x) \rightarrow x = \frac{8.25}{10.5} = 0.7857$$

Ερώτημα 8

Έστω το πρόβλημα ταξινόμησης κατά Bayes σε 2 ισοπίθανες κλάσεις στο 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) $y = \frac{1}{4}x + \frac{5}{8}$

Λύση

Το **(δ)**. Για τις δεδομένες γκαουσιανές κατανομές, το όριο απόφασης είναι μια γραμμική συνάρτηση διάκρισης. Η γραμμική συνάρτηση διάκρισης $g(x)$ ορίζεται ως εξής:

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

Με:

$$\Sigma^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \text{ και } \ln\frac{P(\omega_1)}{P(\omega_2)} = ln1 = 0$$

Υπολογίζουμε τα $(\mu_1 - \mu_2)$, $(\mu_1^T\Sigma^{-1}\mu_1)$ και $(\mu_2^T\Sigma^{-1}\mu_2)$:

$$\mu_1 - \mu_2 = (13) - (2-1) = -14$$

$$\mu_1^T\Sigma^{-1}\mu_1 = (1 \quad 3)\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}(13) = (1 \quad 3)(26) = 1\cdot 2 + 3\cdot 6 = 2 + 18 = 20$$

$$\mu_2^T\Sigma^{-1}\mu_2 = (2 \quad -1)\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}(2-1) = (2 \quad -1)(4-2) = 2\cdot 4 + (-1)\cdot(-2) = 8 + 2 = 10$$

Και αντικαθιστούμε:

$$g(x) = x^T\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}(-14) - \frac{1}{2}(10) = x^T(-28) - 5$$

Για ένα σημείο (x,y) έχουμε:

$$g(x,y) = -2x + 8y - 5 = 0 \rightarrow 8y = 2x + 5 \rightarrow y = \frac{1}{4}x + \frac{5}{8}$$

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 1ης 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) $\theta_1^{(1)} = [0.2,-0.8], \theta_2^{(1)} = [0.75,1.5]$

Λύση

Το **(δ)**. Αρχικά, πρέπει να αναθέσουμε κάθε σημείο στο πλησιέστερο κέντρο. Τα αρχικά κέντρα είναι: $\theta_1^{(0)} = [0,0]$, $\theta_2^{(0)} = [1,2]$. Θα υπολογίσουμε την ευκλείδεια απόσταση από κάθε σημείο και από τα δύο κέντρα.

Για $x_1 = [-0.4,0.6]$:

$$d\left(x_1, \theta_1^{(0)}\right) = \sqrt{(-0.4-0)^2 + (0.6-0)^2} = \sqrt{0.16+0.36} = \sqrt{0.52} \approx 0.72$$

$$d\left(x_1, \theta_2^{(0)}\right) = \sqrt{(-0.4-1)^2 + (0.6-2)^2} = \sqrt{1.96+1.96} = \sqrt{3.92} \approx 1.98$$

Αφού $d\left(x_1, \theta_1^{(0)}\right) < d\left(x_1, \theta_2^{(0)}\right)$, το x_1 αναθέτεται στο $\theta_1^{(0)}$.

Για $x_2 = [1.5,1]$:

$$d\left(x_2, \theta_1^{(0)}\right) = \sqrt{(1.5-0)^2 + (1-0)^2} = \sqrt{2.25+1} = \sqrt{3.25} \approx 1.8$$

$$d\left(x_2, \theta_2^{(0)}\right) = \sqrt{(1.5-1)^2 + (1-2)^2} = \sqrt{0.25+1} = \sqrt{1.25} \approx 1.12$$

Αφού $d\left(x_2, \theta_2^{(0)}\right) < d\left(x_2, \theta_1^{(0)}\right)$, x_2 αναθέτεται στο $\theta_2^{(0)}$.

For $x_3 = [0,2]$:

$$d\left(x_3, \theta_1^{(0)}\right) = \sqrt{(0-0)^2 + (2-0)^2} = \sqrt{0+4} = 2$$

$$d\left(x_3, \theta_2^{(0)}\right) = \sqrt{(0-1)^2 + (2-2)^2} = \sqrt{1+0} = 1$$

Αφού $d\left(x_3, \theta_2^{(0)}\right) < d\left(x_3, \theta_1^{(0)}\right)$, x_3 αναθέτεται στο $\theta_2^{(0)}$.

Για $x_4 = [1,-1]$:

$$d\left(x_4, \theta_1^{(0)}\right) = \sqrt{(1-0)^2 + (-1-0)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$$

$$d\left(x_4, \theta_2^{(0)}\right) = \sqrt{(1-1)^2 + (-1-2)^2} = \sqrt{0+9} = 3$$

Αφού $d\left(x_4, \theta_1^{(0)}\right) < d\left(x_4, \theta_2^{(0)}\right)$, x_4 αναθέτεται στο $\theta_1^{(0)}$.

For $x_5 = [0, -2]$:

$$d\left(x_5, \theta_1^{(0)}\right) = \sqrt{(0-0)^2 + (-2-0)^2} = \sqrt{0+4} = 2$$

$$d\left(x_5, \theta_2^{(0)}\right) = \sqrt{(0-1)^2 + (-2-2)^2} = \sqrt{1+16} = \sqrt{17} \approx 4.12$$

Αφού $d\left(x_5, \theta_1^{(0)}\right) < d\left(x_5, \theta_2^{(0)}\right)$, x_5 αναθέτεται στο $\theta_1^{(0)}$.

Στην συνέχεια υπολογίζουμε εκ νέου τα κέντρα.

Τώρα, τα σημεία που ανατίθενται στο $\theta_1^{(0)}$ είναι: $x_1 = [-0.4, 0.6], x_4 = [1, -1], x_5 = [0, -2]$. Το νέο κέντρο $\theta_1^{(1)}$ είναι ο μέσος όρος τους:

$$\theta_1^{(1)} = \frac{1}{3}([-0.4, 0.6] + [1, -1] + [0, -2]) = \frac{1}{3}([-0.4 + 1 + 0, 0.6 - 1 - 2]) = \frac{1}{3}([0.6, -2.4]) = [0.2, -0.8]$$

Τώρα, τα σημεία που ανατίθενται στο $\theta_2^{(0)}$ είναι: $x_2 = [1.5, 1], x_3 = [0, 2]$. Το νέο κέντρο $\theta_2^{(1)}$ είναι ο μέσος όρος τους:

$$\theta_2^{(1)} = \frac{1}{2}([1.5, 1] + [0, 2]) = \frac{1}{2}([1.5, 3]) = [0.75, 1.5]$$

Ερώτημα 10

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

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

Αν με βάση τον $P(X)$ εφαρμόσουμε τον ιεραρχικό συσσωρευτικό αλγόριθμο πλήρους δεσμού (complete link), η ομαδοποίηση που προκύπτει μετά το 2ο 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) $[x_1, x_2], [x_3, x_4], [x_5]$

Λύση

Το **(δ)**. Θεωρούμε πως κάθε σημείο έχει το δικό cluster $C_i = x_i$. Αναζητούμε τη μικρότερη μη μηδενική εγγραφή στον πίνακα, η οποία υποδεικνύει το πλησιέστερο ζεύγος σημείων που πρέπει να συγχωνευθεί πρώτο. Η μικρότερη μη μηδενική εγγραφή είναι 1,41 (μεταξύ x_3 και x_4). Συγχωνεύουμε τα σημεία x_3 και x_4:

$$C_{34} = \{x_3, x_4\}$$

Ενημερώνουμε τον πίνακα ανομοιότητας για το νέο cluster C_{34} χρησιμοποιώντας τη μέθοδο πλήρους σύνδεσης (μέγιστη απόσταση):

$$P_{34}(X) = \begin{pmatrix} 0 & 3.42 & 2.24 & 3.61 & 5.83 \\ 3.42 & 0 & 2.83 & 3.16 & 3.61 \\ 2.24 & 2.83 & 0 & 4.12 & - \\ 3.16 & 3.16 & 4.12 & 2.15 & - \\ 5.83 & 3.61 & 4.12 & 2.15 & 0 \end{pmatrix}$$

Στη συνέχεια, βρίσκουμε το επόμενο πλησιέστερο ζεύγος συστάδων. Η μικρότερη μη μηδενική εγγραφή στον ενημερωμένο πίνακα είναι 2.24 (μεταξύ x_3 και x_1). Συγχωνεύουμε τα σημεία x_3 και x_1:

$$C_{13} = \{x_1, x_3\}$$

Ενημερώνουμε τον πίνακα ανομοιότητας για το νέο cluster C_{13} χρησιμοποιώντας τη μέθοδο πλήρους σύνδεσης (μέγιστη απόσταση):

$$P_{13}(X) = \begin{pmatrix} 0 & 3.42 & 3.61 & 5.83 \\ 3.42 & 0 & 2.83 & 3.61 \\ 3.61 & 2.83 & 0 & 4.12 \\ 5.83 & 3.61 & 4.12 & 0 \end{pmatrix}$$

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.

Λύση

Το **(α)**. Πίνακας Αληθείας για $y = x_1 \cup \neg(x_2)$:

x_1	x_2	y	*Class*
0	0	1	1
0	1	0	0
1	0	1	1
1	1	1	1

Ένα απλό perceptron μπορεί να αναπαραστήσει μια συνάρτηση Boole αν είναι γραμμικά διαχωρίσιμη. Ας απεικονίσουμε τη συνάρτηση στο επίπεδο x_1, x_2:

- Class 1: (0,0), (1,0), (1,1).
- Class 0: (0,1).

Τα σημεία αυτά είναι γραμμικά διαχωρίσιμα, επειδή μπορούμε να σχεδιάσουμε μια ευθεία γραμμή που διαχωρίζει το (0,1) από τα άλλα σημεία. Συγκεκριμένα, ένα perceptron μπορεί να χρησιμοποιήσει βάρη και μεροληψία για να διαχωρίσει αυτά τα σημεία. Για να αναπαραστήσετε τη συνάρτηση με ένα perceptron, πρέπει να βρούμε τα βάρη w_1 και w_2 και μια προκατάληψη b έτσι ώστε:

$$y = 1: w_1 x_1 + w_2 x_2 + b > 0$$

$$y = 0: w_1 x_1 + w_2 x_2 + b \leq 0$$

Εξετάζοντας τα σημεία:

1) $(0,0): 0 \cdot w_1 + 0 \cdot w_2 + b > 0 \rightarrow b > 0$
2) $(1,0): 1 \cdot w_1 + 0 \cdot w_2 + b > 0 \rightarrow w_1 + b > 0$
3) $(1,1): 1 \cdot w_1 + 1 \cdot w_2 + b > 0 \rightarrow w_1 + w_2 + b > 0$
4) $(0,1): 0 \cdot w_1 + 1 \cdot w_2 + b \leq 0 \rightarrow w_2 + b \leq 0$

Επιλέγουμε $(b = 0.5)$, $(w_1 = 1)$, $(w_2 = -1)$ ώστε να ικανοποιούνται οι συνθήκες. Άρα, η συνάρτηση μπορεί να αναπαρασταθεί από απλό perceptron.

Ερώτημα 12

Έστω πολυστρωματικό perceptron (MLP) που περιέχει συνολικά 100 συναπτικά 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.

Λύση

Το **(β)**. Συνολικός αριθμός δειγμάτων εκπαίδευσης:

$$\text{Total samples} = \text{Number of minibatches} \times \text{Minibatch size} = 300 \times 20 = 6000$$

Κάθε minibatch οδηγεί σε μία ενημέρωση βάρους ανά βάρος. Έτσι, για κάθε μίνι παρτίδα, ενημερώνονται και τα 100 βάρη. Δεδομένου ότι κάθε 1 από τα 300 minibatch οδηγεί σε ενημέρωση και των 100 βαρών, ο συνολικός αριθμός των ενημερώσεων βαρών είναι:

$$\text{Total weight updates} = \text{Number of minibatches} \times \text{Number of weights} = 300 \times 100 = 30{,}000$$

Ερώτημα 13

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

a) Η εισαγωγή των μεταβλητών χαλαρότητας (slack variables) ξ_i 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 Μ/Σ Φ() μόνη της.

Λύση

Το **(γ)**. Οι SVM δεν παγιδεύονται σε τοπικά ελάχιστα, επειδή το πρόβλημα βελτιστοποίησής τους είναι κυρτό, πράγμα που σημαίνει ότι κάθε τοπικό ελάχιστο είναι και παγκόσμιο ελάχιστο. Αντίθετα, τα MLP (Multilayer Perceptrons) μπορούν να παγιδευτούν σε τοπικά ελάχιστα λόγω του μη κυρτού τοπίου βελτιστοποίησής τους.

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 εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-one και N το πλήθος των δυαδικών SVMs που θα πρέπει να εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-all. Για τα M, N ισχύει:

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

Λύση

Το **(γ)**. Στη μέθοδο «ένας εναντίον ενός», εκπαιδεύεται ένας SVM για κάθε ζεύγος κλάσεων. Για k κλάσεις, ο αριθμός των απαιτούμενων δυαδικών SVM δίνεται από το συνδυασμό:

$$M = \binom{k}{2} = \frac{k(k-1)}{2}$$

Για k=3 κλάσεις:

$$M = \frac{3(3-1)}{2} = \frac{3 \cdot 2}{2} = 3$$

Στη μέθοδο «ένας εναντίον όλων», ένα SVM εκπαιδεύεται για κάθε κλάση έναντι όλων των άλλων κλάσεων. Για k κλάσεις, ο αριθμός των δυαδικών SVM που απαιτούνται είναι $N = k$, δηλαδή $N = 3$.

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.

Λύση

Πιθανώς το **(β)**.

a) Σωστό. Ο ε-greedy επιλέγει την ενέργεια με τη μεγαλύτερη εκτιμώμενη ανταμοιβή με πιθανότητα $1-\varepsilon$, και με πιθανότητα ε, επιλέγει μια τυχαία ενέργεια για να εξασφαλίσει την εξερεύνηση.
b) Αυτό είναι κάπως διφορούμενο (αλλά μπορεί να θεωρηθεί σωστό). Ο ε-greedy αντιμετωπίζει όντως το μειονέκτημα του απλού άπληστου αλγορίθμου (ο οποίος εκμεταλλεύεται μόνο και δεν εξερευνά) εισάγοντας μια πιθανότητα «ε» επιλογής μιας τυχαίας ενέργειας, εξισορροπώντας έτσι την εξερεύνηση και την εκμετάλλευση.
c) Σωστό. Όταν $\varepsilon = 0$, συμπεριφέρεται ακριβώς όπως ο απλός άπληστος αλγόριθμος, επιλέγοντας πάντα την ενέργεια με την υψηλότερη εκτιμώμενη ανταμοιβή και χωρίς καθόλου εξερεύνηση.
d) Σωστό. Η αύξηση του «ε» (αύξηση της εξερεύνησης) μπορεί να οδηγήσει σε χειρότερη απόδοση εάν έχει ως αποτέλεσμα υπερβολική εξερεύνηση και όχι αρκετή εκμετάλλευση, ιδίως εάν οι εκτιμώμενες ανταμοιβές είναι ήδη κοντά στη βέλτιστη.

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.

Λύση

Το **(γ)**.

- Συνήθως υπάρχουν πολλά δέντρα απόφασης που μπορούν να ταξινομήσουν σωστά ένα σύνολο δεδομένων, ειδικά αν το σύνολο δεδομένων δεν είναι πολύ μεγάλο ή πολύπλοκο. Η δομή του δέντρου απόφασης μπορεί να ποικίλλει, με αποτέλεσμα διαφορετικά δέντρα που όλα δίνουν σωστές ταξινομήσεις.
- Αλγόριθμοι όπως οι ID3, C4.5 και CART έχουν σχεδιαστεί για την αποτελεσματική δημιουργία δέντρων απόφασης. Αυτοί οι αλγόριθμοι χρησιμοποιούν ευρετικές μεθόδους όπως το κέρδος πληροφορίας ή η ακαθαρσία Gini για να κάνουν διαχωρισμούς σε κάθε κόμβο, οδηγώντας σε σωστή ταξινόμηση των δεδομένων εκπαίδευσης.

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

Λύση

Το **(δ)**. Η μάθηση PAC είναι ένα πλαίσιο στη μηχανική μάθηση που παρέχει μια θεωρητική εγγύηση για την ικανότητα εκμάθησης ενός ταξινομητή. Ένας ταξινομητής είναι PAC learnable εάν, με μεγάλη πιθανότητα, μπορεί να παράγει μια υπόθεση που είναι κατά προσέγγιση σωστή, δεδομένου ενός επαρκούς αριθμού δειγμάτων εκπαίδευσης.

a) Αυτή η δήλωση υποδηλώνει ότι ο ταξινομητής είναι εκπαιδεύσιμος PAC μόνο για ένα συγκεκριμένο σύνολο εισόδου. Η εκμάθηση PAC συνήθως εξετάζει την ικανότητα γενίκευσης πέρα από ένα συγκεκριμένο σύνολο σημείων δεδομένων, οπότε αυτή η δήλωση είναι πιθανώς εσφαλμένη.
b) Παρόμοια με το (α), αυτή η δήλωση υπονοεί περιορισμένη γενίκευση, η οποία δεν συνάδει με τη γενική έννοια της μάθησης PAC. Ως εκ τούτου, είναι πιθανότατα εσφαλμένη.
c) Αυτή η δήλωση θα ήταν αληθής εάν ο ταξινομητής δεν μπορεί να γενικεύσει σε αόρατα δεδομένα, αλλά δεδομένου ότι μπορεί να κατακερματίσει οποιοδήποτε σύνολο σημείων στο επίπεδο επιλέγοντας κατάλληλες παραμέτρους, υποδηλώνει κάποιο επίπεδο ευελιξίας και learnability.

Ερώτημα 18

Σε ένα δέντρο απόφασης που έχει κατασκευαστεί από ένα σύνολο δεδομένων με τον αλγόριθμο ID3, εφαρμόζουμε τεχνική κλαδέματος 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.

Λύση

Το **(δ)**.

1) Λάθος. Το κλάδεμα γίνεται συνήθως για να βελτιωθεί η γενίκευση και να μειωθεί η υπερπροσαρμογή. Μπορεί να οδηγήσει σε εσφαλμένη ταξινόμηση ορισμένων στοιχείων, ιδίως εκείνων που είχαν ταξινομηθεί σωστά λόγω της υπερπροσαρμογής.
2) **Σωστό**. Πριν από το κλάδεμα, το δέντρο απόφασης μπορεί να ταξινομήσει τέλεια τα δεδομένα εκπαίδευσης, ενδεχομένως με υπερπροσαρμογή. Μετά το κλάδεμα, το δέντρο γίνεται απλούστερο και μπορεί να ταξινομήσει λανθασμένα ορισμένα δεδομένα εκπαίδευσης για να βελτιώσει τη γενίκευση.
3) Λάθος. Το κλάδεμα συνήθως αποσκοπεί στη μείωση της πολυπλοκότητας του δέντρου και στη βελτίωση της γενίκευσης με την αφαίρεση κλάδων που μπορεί να υπερπροσαρμόζουν τα δεδομένα εκπαίδευσης.
4) **Σωστό**. Το κλάδεμα μειώνει το μέγεθος του δέντρου με την αφαίρεση των περιττών κλάδων, γεγονός που συνήθως οδηγεί σε καλύτερη γενίκευση σε αόρατα δεδομένα.

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:

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

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.

Λύση

Το **(α)**. Υπολογίζουμε το δείκτη Gini για το σύνολο των δεδομένων. Υπάρχουν 4 εγγραφές: 3 "Yes" και 1 "No". Επομένως, η πιθανότητα του «Ναι» είναι $p_{Yes} = 3/4$ και η πιθανότητα του «Όχι» $p_{No} = 1/4$. Ο δείκτης Gini είναι:

$$Gini(D) = 1 - (p_{Yes}^2 + p_{No}^2) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = 1 - \left(\frac{9}{16} + \frac{1}{16}\right) = 1 - \frac{10}{16} = \frac{6}{16} = 0.375$$

Υπολογίζουμε το δείκτη Gini για κάθε χαρακτηριστικό:

- Outlook:
  - Sunny: 1 εγγραφή «Όχι»:

    $$Gini(Sunny) = 1 - (0^2 + 1^2) = 0$$

  - Overcast: 3 εγγραφές «Ναι»:

    $$Gini(Overcast) = 1 - (1^2 + 0^2) = 0$$

  - Συνολικά:

    $$Gini(Outlook) = \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0 = 0$$

- Temperature:
  - Hot: 2 "Ναι", 1 "Όχι":

    $$Gini(Hot) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} \approx 0.444$$

- Cool: 1 "Ναι":

$$Gini(Cool) = 1 - (1^2 + 0^2) = 0$$

- Συνολικά:

$$Gini(Temperature) = \frac{3}{4} \cdot 0.444 + \frac{1}{4} \cdot 0 = 0.333$$

- Windy:
  - False: 2 "Ναι", 1 "Όχι":

$$Gini(False) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} \approx 0.444$$

  - True: 1 "Ναι":

$$Gini(True) = 1 - (1\text{^}2 + 0\text{^}2) = 0$$

  - Συνολικά:

$$Gini(Windy) = \frac{3}{4} \cdot 0.444 + \frac{1}{4} \cdot 0 = 0.333$$

Δεδομένου ότι ο δείκτης Gini για το «Outlook» είναι ο μικρότερος, το «Outlook» θα επιλεγεί ως η ρίζα του δέντρου αποφάσεων.

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.

Λύση

Το **(γ)**. Η διάσταση VC είναι ένα μέτρο της ικανότητας ενός δυαδικού ταξινομητή, που αντιπροσωπεύει τον μεγαλύτερο αριθμό σημείων που μπορούν να διαλυθούν (δηλαδή να ταξινομηθούν σωστά με όλους τους δυνατούς τρόπους) από τον ταξινομητή. Για ένα απλό perceptron σε ένα δισδιάστατο χώρο χαρακτηριστικών, η διάσταση VC μπορεί να προσδιοριστεί εξετάζοντας τον αριθμό των σημείων που μπορούν να κατακερματιστούν από έναν γραμμικό διαχωριστή (μια γραμμή).

1) Ένα μεμονωμένο σημείο μπορεί πάντα να ταξινομηθεί σωστά ως οποιαδήποτε κλάση, οπότε μπορεί να κατακερματιστεί. Άρα VC ≥ 1.
2) Οποιαδήποτε δύο σημεία μπορούν να διαχωριστούν γραμμικά σε ένα επίπεδο 2D. Μπορούν να ταξινομηθούν με όλους τους δυνατούς τρόπους. Άρα VC ≥ 2.
3) Οποιαδήποτε τρία σημεία που δεν είναι κολλητά μπορούν να κατακερματιστούν από ένα γραμμικό διαχωριστή. Αυτό σημαίνει ότι μπορούμε να ταξινομήσουμε σωστά όλους τους πιθανούς 8 (2^3) συνδυασμούς τριών σημείων. Άρα VC ≥ 3.
4) Δεν μπορούν όλα τα σύνολα τεσσάρων σημείων να κατακερματιστούν από έναν μόνο γραμμικό διαχωριστή. Συγκεκριμένα, αν τέσσερα σημεία αποτελούν τις κορυφές ενός κυρτού τετράπλευρου, δεν υπάρχει μία και μόνη γραμμή που να μπορεί να διαχωρίσει όλες τις πιθανές δυαδικές επισημάνσεις αυτών των σημείων. Έτσι, τέσσερα σημεία δεν μπορούν πάντα να θρυμματιστούν από έναν γραμμικό διαχωριστή. Άρα VC < 4.

Επαναληπτική 23

Ερώτημα 1

Θεωρήστε το τυχαίο διάνυσμα $x = [x_1, x_2]^T$ που ακολουθεί την Caussian κατανομή, N(μ, Σ):

$$f(x) = (2\pi)^{-1}|\Sigma|^{-1/2} e^{\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]}$$

Με:

$$\Sigma = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}, \mu = [1,2]^T$$

Η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας $f(x_1|x_2 = 4)$ είναι:

a) $N\left(1, \frac{3}{4}\right)$
b) $N\left(0, \frac{3}{4}\right)$
c) $N\left(1, \frac{4}{3}\right)$
d) $N\left(0, \frac{4}{3}\right)$

Λύση

Ερώτημα 2

Δίνονται τα σημεία εκπαίδευσης $x_1 = -1, x_2 = -2, x_3 = 1, x_4 = 3$, από τα οποία τα 2 πρώτα ανήκουν στην ομάδα ω_1 και τα 2 τελευταία στην ομάδα ω_2. Αντιστοιχίζουμε την ετικέτα -1 στα σημεία της ομάδας ω_1 και την 1 στα σημεία της ομάδας ω_2. Αν χρησιμοποιήσουμε τα παραπάνω σημεία για να εκπαιδεύσουμε έναν ταξινομητή ελαχίστων τετραγώνων, το σημείο απόφασης που επιστρέφει ο ταξινομητής είναι:

a) 0,3
b) +0,25
c) 0,25
d) -0,3

Λύση

Το **(γ)**. Για να βρούμε το σημείο απόφασης που επιστρέφει ο ταξινομητής ελαχίστων τετραγώνων, πρέπει να χρησιμοποιήσουμε τα σημεία εκπαίδευσης και τις αντίστοιχες ετικέτες τους. Ο ταξινομητής ελαχίστων τετραγώνων επιδιώκει να βρει ένα γραμμικό όριο απόφασης που ελαχιστοποιεί το άθροισμα των τετραγωνικών σφαλμάτων μεταξύ των προβλεπόμενων ετικετών και των πραγματικών ετικετών.

Πρέπει να βρούμε τον γραμμικό ταξινομητή $f(x) = wx + b$ έτσι ώστε οι προβλεπόμενες ετικέτες να ταιριάζουν όσο το δυνατόν περισσότερο με τις πραγματικές ετικέτες.

Η αντικειμενική συνάρτηση ελαχίστων τετραγώνων είναι η ελαχιστοποίηση:

$$\sum_{i=1}^{4} \big(y_i - (wx_i + b)\big)^2$$

Μπορούμε να δημιουργήσουμε το σύστημα εξισώσεων με βάση τα σημεία εκπαίδευσης και τις ετικέτες τους:

$$-1 = w(-1) + b$$

$$-1 = w(-2) + b$$

$$1 = w(1) + b$$

$$1 = w(3) + b$$

Επομένως:

$$\begin{pmatrix} -1 & 1 \\ -2 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}$$

Έστω **A** ο πίνακας των εισόδων και **y** το διάνυσμα των ετικετών:

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 \\ -2 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}$$

Η λύση ελαχίστων τετραγώνων για w και b μπορεί να βρεθεί χρησιμοποιώντας:

$$\binom{w}{b} = (A^T A)^{-1} A^T y$$

Υπολογίζουμε το $(A^T A)$:

$$A^T A = \begin{pmatrix} -1 & -2 & 1 & 3 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ -2 & 1 \\ 1 & 1 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 15 & 1 \\ 1 & 4 \end{pmatrix}$$

Άρα, το $(A^T A)^{-1}$:

$$(A^T A)^{-1} = \frac{1}{15 \cdot 4 - 1 \cdot 1} \begin{pmatrix} 4 & -1 \\ -1 & 15 \end{pmatrix} = \frac{1}{59} \begin{pmatrix} 4 & -1 \\ -1 & 15 \end{pmatrix}$$

Υπολογίζουμε και το $(A^T y)$:

$$A^T y = \begin{pmatrix} -1 & -2 & 1 & 3 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 0 \end{pmatrix}$$

Λύνουμε ως προς $\binom{w}{b}$:

$$\binom{w}{b} = (A^T A)^{-1} A^T y = \frac{1}{59} \begin{pmatrix} 4 & -1 \\ -1 & 15 \end{pmatrix} \begin{pmatrix} 8 \\ 0 \end{pmatrix} = \frac{1}{59} \begin{pmatrix} 32 \\ -8 \end{pmatrix} = \begin{pmatrix} \frac{32}{59} \\ -\frac{8}{59} \end{pmatrix} \rightarrow$$

$$\rightarrow w = \frac{32}{59}, b = -\frac{8}{59}$$

Το σημείο απόφασης (όπου $f(x) = 0$) μπορεί να βρεθεί θέτοντας τη γραμμική εξίσωση ίση με μηδέν:

$$w\,x\, +\, b\, =\, 0 \rightarrow \frac{32}{59} x - \frac{8}{59} = 0 \rightarrow 32x\, =\, 8 \rightarrow$$

$$\rightarrow x = \frac{8}{32} = \frac{1}{4} = 0.25$$

Ερώτημα 3

Θεωρούμε μια παραλλαγή του αλγορίθμου k-means, όπου στο 1ο βήμα του αλγορίθμου, αντί της Ευκλείδειας απόστασης χρησιμοποιείται η l_1 απόσταση:

$$d(x_i, \theta_j) = \left|\left|x_i - \theta_j\right|\right|_1 = \sum_{k=1}^{l} \left|x_{ik} - \theta_{jk}\right|$$

Εφαρμόζουμε τον αλγόριθμο αυτό για τον διαχωρισμό των σημείων $x_1 = (0,1), x_2 = (1,2), x_3 = (2,1), x_4 = (0,-1), x_5 = (0,-2)$ σε 2 ομάδες. Αν τα κέντρα των ομάδων αρχικοποιούνται ως $\theta_1^{(0)} = (-1,0), \theta_2^{(0)} = (2,0)$ πόσες επαναλήψεις χρειάζεται ο αλγόριθμος για να συγκλίνει στις τελικές θέσεις των κέντρων των ομάδων;

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

Λύση

Το **(α)**. Υπολογίζουμε την απόσταση L_1 από κάθε σημείο προς τα αρχικά κέντρα θ_1 και θ_2.

- Αποστάσεις ως προς το $\theta_1^{(0)} = (-1,0)$:
  - $d\left((0,1), \theta_1^{(0)}\right) = |0 - (-1)| + |1 - 0| = 1 + 1 = 2$
  - $d\left((1,2), \theta_1^{(0)}\right) = |1 - (-1)| + |2 - 0| = 2 + 2 = 4$
  - $d\left((2,1), \theta_1^{(0)}\right) = |2 - (-1)| + |1 - 0| = 3 + 1 = 4$
  - $d\left((0,-1), \theta_1^{(0)}\right) = |0 - (-1)| + |-1 - 0| = 1 + 1 = 2$
  - $d\left((0,-2), \theta_1^{(0)}\right) = |0 - (-1)| + |-2 - 0| = 1 + 2 = 3$
- Αποστάσεις ως προς το $\theta_2^{(0)} = (2,0)$:
  - $d\left((0,1), \theta_2^{(0)}\right) = |0 - 2| + |1 - 0| = 2 + 1 = 3$
  - $d\left((1,2), \theta_2^{(0)}\right) = |1 - 2| + |2 - 0| = 1 + 2 = 3$
  - $d\left((2,1), \theta_2^{(0)}\right) = |2 - 2| + |1 - 0| = 0 + 1 = 1$
  - $d\left((0,-1), \theta_2^{(0)}\right) = |0 - 2| + |-1 - 0| = 2 + 1 = 3$
  - $d\left((0,-2), \theta_2^{(0)}\right) = |0 - 2| + |-2 - 0| = 2 + 2 = 4$

Επομένως τα x_1, x_4, x_5 είναι πιο κοντά στο θ_1 και τα x_2, x_3 είναι πιο κοντά στο θ_2. Θα υπολογίσουμε στη συνέχεια τα νέα κέντρα:

$$\theta_1^{(1)} = \left(\frac{0+0+0}{3}, \frac{1-1-2}{3}\right) = \left(0, -\frac{2}{3}\right)$$

$$\theta_2^{(1)} = \left(\frac{1+2}{2}, \frac{2+1}{2}\right) = (1.5, 1.5)$$

Αναθέτουμε ξανά τα σημεία με βάση τα νέα κέντρα.

- Αποστάσεις ως προς το $\theta_1^{(1)} = \left(0, -\frac{2}{3}\right)$:
  - $d\left((0,1), \theta_1^{(1)}\right) = |0-0| + |1 + \frac{2}{3}| = 0 + 1.67 = 1.67$
  - $d\left((1,2), \theta_1^{(1)}\right) = |1-0| + |2 + \frac{2}{3}| = 1 + 2.67 = 3.67$
  - $d\left((2,1), \theta_1^{(1)}\right) = |2-0| + |1 + \frac{2}{3}| = 2 + 1.67 = 3.67$
  - $d\left((0,-1), \theta_1^{(1)}\right) = |0-0| + |-1 + \frac{2}{3}| = 0 + 0.33 = 0.33$
  - $d\left((0,-2), \theta_1^{(1)}\right) = |0-0| + |-2 + \frac{2}{3}| = 0 + 1.33 = 1.33$
- Αποστάσεις ως προς το $\theta_2^{(1)} = (1.5, 1.5)$:
  - $d\left((0,1), \theta_2^{(1)}\right) = |0-1.5| + |1-1.5| = 1.5 + 0.5 = 2$
  - $d\left((1,2), \theta_2^{(1)}\right) = |1-1.5| + |2-1.5| = 0.5 + 0.5 = 1$
  - $d\left((2,1), \theta_2^{(1)}\right) = |2-1.5| + |1-1.5| = 0.5 + 0.5 = 1$
  - $d\left((0,-1), \theta_2^{(1)}\right) = |0-1.5| + |-1-1.5| = 1.5 + 2.5 = 4$
  - $d\left((0,-2), \theta_2^{(1)}\right) = |0-1.5| + |-2-1.5| = 1.5 + 3.5 = 5$

Επομένως τα x_1, x_4, x_5 είναι πιο κοντά στο θ_1 και τα x_2, x_3 είναι πιο κοντά στο θ_2. Εφόσον οι αναθέσεις δεν έχουν αλλάξει, ο αλγόριθμος έχει συγκλίνει.

Ερώτημα 4

Έστω $x_1, \dots, x_N \epsilon R$ ανεξάρτητες παρατηρήσεις που προέρχονται από τυχαίες μεταβλητές που ακολουθούν την ομοιόμορφη κατανομή στο διάστημα [α, 4]. Η εκτίμηση μέγιστης πιθανοφάνειας της παραμέτρου α της κατανομής δίνεται από τη σχέση:

a) $\alpha_{ML} = min[x_1, \dots, x_N]$
b) $\alpha_{ML} = max[x_1, \dots, x_N]$
c) $\alpha_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
d) $\alpha_{ML} = median[x_1, \dots, x_N]$

Λύση

Το **(α)**. Η συνάρτηση πυκνότητας πιθανότητας (ΣΠΠ) της ομοιόμορφης κατανομής στο $[\alpha, 4]$ είναι:

$$f(x) = \frac{1}{4-\alpha}, \alpha \le x \le 4$$

Για ανεξάρτητες παρατηρήσεις $(x_1, \dots, x_N)$, η συνάρτηση πιθανότητας $L(\alpha)$ δίνεται από:

$$L(\alpha) = \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \frac{1}{4-\alpha} = \left(\frac{1}{4-\alpha}\right)^N$$

Ο περιορισμός είναι ότι το «α» πρέπει να είναι μικρότερο ή ίσο με την ελάχιστη παρατήρηση (x_i), δηλαδή:

$$(x_i): [\alpha \le \min(x_1, \dots, x_N)]$$

Η συνάρτηση λογαριθμικής πιθανοφάνειας $\ell(\alpha)$ είναι:

$$\ell(\alpha) = \log L\,(\alpha) = \log\left(\left(\frac{1}{4-\alpha}\right)^N\right) = -N\log(4-\alpha)]$$

Για να μεγιστοποιήσουμε τη λογαριθμική πιθανότητα $\ell(\alpha)$, πρέπει να ελαχιστοποιήσουμε τον παρονομαστή $(4-\alpha)$ επειδή η $(\log)$ είναι μια μονότονα αυξανόμενη συνάρτηση. Όσο μεγαλύτερο είναι το «α» (μέχρι τον περιορισμό), τόσο μεγαλύτερη θα είναι η λογαριθμική πιθανότητα. Επομένως, η ΕΜΠ για το «α» είναι:

$$\alpha_{ML} = \min(x_1, \dots, x_N)$$

Ερώτημα 5

1) Ο αλγόριθμος δεν είναι κατάλληλος για την επεξεργασία δεδομένων σε χώρους υψηλής διάστασης.
2) Η επίδοση του αλγόριθμου επηρεάζεται σημαντικά όταν υπάρχουν ακραία σημεία (outliers) στα δεδομένα.
3) Ο αλγόριθμος παρουσιάζει πολύ υψηλή αριθμητική πολυπλοκότητα, της τάξης του $O(N^3)$, όπου N ο αριθμός των σημείων.
4) Ο αλγόριθμος δεν απαιτεί εκ των προτέρων γνώση του αριθμού των ομάδων στα δεδομένα.

Ποιοι από τους παραπάνω ισχυρισμούς δεν είναι αληθείς για τον αλγόριθμο ομαδοποίησης DBSCAN;

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

Λύση

Το **(α)**.

1) Σωστό. Ο DBSCAN μπορεί να δυσκολευτεί με δεδομένα υψηλών διαστάσεων λόγω της «κατάρας της διαστατικότητας». Σε χώρους υψηλών διαστάσεων, η έννοια της πυκνότητας μπορεί να γίνει λιγότερο σημαντική και οι μετρικές απόστασης που χρησιμοποιούνται μπορεί να γίνουν λιγότερο διακριτικές.
2) **Λάθος**. Ένα από τα δυνατά σημεία του DBSCAN είναι η ικανότητά του να χειρίζεται αποτελεσματικά το θόρυβο και τις ακραίες τιμές. Τα σημεία που θεωρούνται ακραία χαρακτηρίζονται ως θόρυβος και δεν επηρεάζουν το σχηματισμό συστάδων.
3) **Λάθος**. Η χρονική πολυπλοκότητα του DBSCAN είναι γενικά $O(N \log N)$ εάν χρησιμοποιείται μια αποτελεσματική δομή χωρικής ευρετηρίασης όπως ένα δέντρο KD ή ένα δέντρο R. Στη χειρότερη περίπτωση, μπορεί να είναι $O(N^2)$ εάν χρησιμοποιείται ένας απλός υπολογισμός απόστασης για κάθε ζεύγος σημείων.
4) Σωστό. Ο DBSCAN δεν απαιτεί να καθοριστεί εκ των προτέρων ο αριθμός των συστάδων, γεγονός που αποτελεί ένα από τα πλεονεκτήματά του έναντι αλγορίθμων όπως ο k-means.

Ερώτημα 6

Έστω νευρωνικό δίκτυο N_A με 2 εισόδους, ένα κρυμμένο στρώμα με 2 νευρώνες και 1 νευρώνα στο στρώμα εξόδου. Ποιο ισχύει;

a) Εάν οι συναρτήσεις ενεργοποίησης των νευρώνων είναι γραμμικές, τότε μπορούμε να βρούμε ένα νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα, το οποίο θα είναι ισοδύναμο με N_A.
b) Εάν οι συναρτήσεις ενεργοποίησης των νευρώνων είναι σιγμοειδείς, τότε μπορούμε να βρούμε ένα νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα, το οποίο θα είναι ισοδύναμο με N_A.
c) Ανεξάρτητα από τον τύπο των συναρτήσεων ενεργοποίησης των νευρώνων, μπορούμε να βρούμε ένα νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα, το οποίο θα είναι ισοδύναμο με N_A.
d) Κανένα από τα παραπάνω.

Λύση

---

a) **Σωστό**. Αν όλες οι συναρτήσεις ενεργοποίησης είναι γραμμικές, η έξοδος κάθε νευρώνα είναι ένας γραμμικός συνδυασμός των εισόδων του. Σε μια τέτοια περίπτωση, ολόκληρο το νευρωνικό δίκτυο είναι ισοδύναμο με ένα δίκτυο ενός στρώματος (γραμμικός μετασχηματισμός), επειδή η σύνθεση γραμμικών συναρτήσεων εξακολουθεί να είναι μια γραμμική συνάρτηση. Επομένως, ένα δίκτυο χωρίς κρυφά στρώματα (ένα μονό στρώμα) μπορεί να είναι ισοδύναμο με NA, εάν όλες οι συναρτήσεις ενεργοποίησης είναι γραμμικές.
b) Λάθος. Οι σιγμοειδείς συναρτήσεις ενεργοποίησης εισάγουν μη γραμμικότητα, επιτρέποντας στο δίκτυο να προσεγγίζει πιο σύνθετες συναρτήσεις. Ένα δίκτυο ενός στρώματος (χωρίς κρυφά στρώματα) με έναν μόνο νευρώνα δεν μπορεί γενικά να αναπαράγει τη συμπεριφορά ενός δικτύου πολλαπλών στρωμάτων με σιγμοειδείς ενεργοποιήσεις. Το κρυφό στρώμα επιτρέπει την αναπαράσταση πιο σύνθετων, μη γραμμικών συναρτήσεων που δεν μπορούν να αναπαρασταθούν από ένα δίκτυο ενός στρώματος.
c) Λάθος. Η ικανότητα αναπαραγωγής της λειτουργίας του αρχικού δικτύου χωρίς κρυφά στρώματα εξαρτάται από τον τύπο των συναρτήσεων ενεργοποίησης. Όπως αναφέρθηκε, οι μη γραμμικές ενεργοποιήσεις, όπως οι σιγμοειδείς συναρτήσεις, δεν μπορούν να αναπαραχθούν από ένα δίκτυο χωρίς κρυφά στρώματα, ενώ οι γραμμικές ενεργοποιήσεις μπορούν.

Ερώτημα 7

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

a) Αν αφαιρέσουμε από το σύνολο δεδομένων ένα σημείο που ταξινομείται ορθά και βρίσκεται κοντά στο όριο απόφασης τότε το όριο απόφασης (βέλτιστο υπερεπίπεδο διαχωρισμού) δεν θα επηρεαστεί.
b) Αν έχουμε ένα πρόβλημα ταξινόμησης 3 κλάσεων, τότε το πλήθος των δυαδικών ΣΜ που θα πρέπει να εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-one είναι μεγαλύτερο από το πλήθος των δυαδικών ΣΜ που θα πρέπει να εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-all.
c) Με τη χρήση συναρτήσεων πυρήνα (kernel functions) γίνεται απεικόνιση των δεδομένων σε μη γραμμικό χώρο χωρίς να εμφανίζεται πουθενά στις πράξεις η συνάρτηση Μ/Σ Φ() μόνη της.
d) Η εισαγωγή των μεταβλητών χαλαρότητας (slack variables) $\boldsymbol{\xi}_i$ μπορεί να οδηγήσει στην εύρεση λύσης σε περιπτώσεις μη γραμμικά διαχωρίσιμων κλάσεων εξασφαλίζοντας πάντα ορθές ταξινομήσεις.

Λύση

---

a) Λάθος. Εάν αφαιρεθεί ένα σημείο που βρίσκεται κοντά στο όριο απόφασης, μπορεί να επηρεάσει τη θέση του ορίου απόφασης, καθώς τα διανύσματα υποστήριξης είναι τα σημεία δεδομένων που βρίσκονται πλησιέστερα στο όριο και επηρεάζουν άμεσα τη θέση του.
b) Λάθος. Σε ένα πρόβλημα ταξινόμησης 3 κλάσεων:
    i. Μέθοδος "ένας εναντίον ενός": Εκπαιδεύουμε $\binom{3}{2} = 3$ δυαδικές SVMs.
    ii. Μέθοδος μία εναντίον όλων: Εκπαιδεύουμε 3 δυαδικές SVM (μία για κάθε κλάση έναντι των υπολοίπων).
    iii. Επομένως, ο αριθμός των δυαδικών SVMs για τη μέθοδο ένα εναντίον ενός είναι ίσος με τον αριθμό για τη μέθοδο ένα εναντίον όλων.
c) **Σωστό**. Οι συναρτήσεις πυρήνα επιτρέπουν στις SVM να λειτουργούν σε έναν χώρο χαρακτηριστικών υψηλών διαστάσεων έμμεσα, χωρίς να χρειάζεται να υπολογίζονται ρητά οι συντεταγμένες των δεδομένων στον εν λόγω χώρο. Το τέχνασμα του πυρήνα επιτρέπει στο SVM να χρησιμοποιεί το γινόμενο τελείας στο χώρο χαρακτηριστικών χωρίς ποτέ να υπολογίζει ρητά το μετασχηματισμό.
d) Λάθος. Η εισαγωγή χαλαρών μεταβλητών επιτρέπει στα SVM να χειρίζονται μη γραμμικά διαχωρίσιμα δεδομένα επιτρέποντας κάποιες λανθασμένες ταξινομήσεις (soft margin SVM). Δεν εξασφαλίζει πάντα σωστές ταξινομήσεις αλλά στοχεύει στην ελαχιστοποίηση των λανθασμένων ταξινομήσεων, ενώ βρίσκει μια ισορροπία μεταξύ του εύρους περιθωρίου και του σφάλματος ταξινόμησης.

Ερώτημα 8

Έστω ότι πρόκειται να εκπαιδεύσουμε γραμμικό ταξινομητή SVM με χρήση μεταβλητών χαλάρωσης (slack variables) $\boldsymbol{\xi_i}$. Τα δεδομένα του συνόλου εκπαίδευσης προέρχονται από μετρήσεις αισθητήρων που δυνητικά είναι επιρρεπείς σε σφάλματα. Τι από τα παρακάτω ισχύει;

a) Για το συγκεκριμένο πρόβλημα ταξινόμησης φαίνεται να είναι καταλληλότερη μία μικρή τιμή για την παράμετρο ποινής C.
b) Για το συγκεκριμένο πρόβλημα ταξινόμησης φαίνεται να είναι καταλληλότερη μία μεγάλη τιμή για την παράμετρο ποινής C.
c) Η τιμή της παραμέτρου ποινής C είναι αδιάφορη για αυτό το πρόβλημα.
d) Η επιλογή της τιμής της παραμέτρου ποινής C εξαρτάται αποκλειστικά από το μέγεθος του συνόλου εκπαίδευσης.

Λύση

Το **(α)**. Η παράμετρος ποινής C σε ένα SVM με μεταβλητές χαλάρωσης ελέγχει τον συμβιβασμό μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του σφάλματος ταξινόμησης στα δεδομένα εκπαίδευσης. Συγκεκριμένα:

- Μια **μεγάλη τιμή του C** δίνει μεγαλύτερη έμφαση στην ελαχιστοποίηση του σφάλματος ταξινόμησης, πράγμα που σημαίνει ότι το SVM θα προσπαθήσει να ταξινομήσει σωστά όλα τα παραδείγματα εκπαίδευσης, γεγονός που μπορεί να οδηγήσει σε υπερπροσαρμογή, ιδίως εάν τα δεδομένα εκπαίδευσης είναι θορυβώδη ή επιρρεπή σε σφάλματα.
- Μια **μικρή τιμή του C** επιτρέπει περισσότερες λανθασμένες ταξινομήσεις, με αποτέλεσμα μεγαλύτερο περιθώριο. Αυτό μπορεί να είναι επωφελές όταν τα δεδομένα εκπαίδευσης είναι θορυβώδη, καθώς οδηγεί σε έναν πιο εύρωστο ταξινομητή που γενικεύει καλύτερα σε νέα δεδομένα.

Δεδομένου ότι τα δεδομένα στο σύνολο εκπαίδευσης προέρχονται από μετρήσεις αισθητήρων που είναι δυνητικά επιρρεπείς σε σφάλματα, θα θέλαμε να αποφύγουμε την υπερβολική προσαρμογή στο θόρυβο των δεδομένων. Επομένως, μια μικρότερη τιμή του C θα ήταν καταλληλότερη, καθώς θα επέτρεπε κάποιες λανθασμένες ταξινομήσεις και θα βοηθούσε στη δημιουργία ενός μοντέλου που γενικεύει καλύτερα.

Ερώτημα 9

Ποια από τις παρακάτω προτάσεις σχετικά με τον ε-άπληστο αλγόριθμο στην ενισχυτική μάθηση είναι λανθασμένη;

a) Στη διάρκεια της εκτέλεσης του αλγορίθμου, σε κάποιες περιπτώσεις επιλέγεται η ενέργεια με τη μέγιστη εκτιμώμενη ανταμοιβή και σε άλλες κάποια άλλη ενέργεια με τυχαίο τρόπο
b) Δεν αποκλείεται η αύξηση της εξερεύνησης σε κάποιες περιπτώσεις να μην βελτιώνει την επίδοση του αλγορίθμου.
c) Αντιμετωπίζει το πρόβλημα της ελλιπούς εκμετάλλευσης που διακρίνει τον απλό άπληστο αλγόριθμο.
d) Για κάποια οριακή τιμή του ε εκφυλίζεται στον απλό άπληστο αλγόριθμο.

Λύση

Πιθανώς το **(γ)**.

a) Σωστό. Ο ε-greedy επιλέγει την ενέργεια με τη μεγαλύτερη εκτιμώμενη ανταμοιβή με πιθανότητα $1-\varepsilon$, και με πιθανότητα ε, επιλέγει μια τυχαία ενέργεια για να εξασφαλίσει την εξερεύνηση.
b) Σωστό. Η αύξηση του «ε» (αύξηση της εξερεύνησης) μπορεί να οδηγήσει σε χειρότερη απόδοση εάν έχει ως αποτέλεσμα υπερβολική εξερεύνηση και όχι αρκετή εκμετάλλευση, ιδίως εάν οι εκτιμώμενες ανταμοιβές είναι ήδη κοντά στη βέλτιστη.
c) Αυτό είναι κάπως διφορούμενο (αλλά μπορεί να θεωρηθεί σωστό). Ο ε-greedy αντιμετωπίζει όντως το μειονέκτημα του απλού άπληστου αλγορίθμου (ο οποίος εκμεταλλεύεται μόνο και δεν εξερευνά) εισάγοντας μια πιθανότητα «ε» επιλογής μιας τυχαίας ενέργειας, εξισορροπώντας έτσι την εξερεύνηση και την εκμετάλλευση.
d) Σωστό. Όταν $\varepsilon = 0$, συμπεριφέρεται ακριβώς όπως ο απλός άπληστος αλγόριθμος, επιλέγοντας πάντα την ενέργεια με την υψηλότερη εκτιμώμενη ανταμοιβή και χωρίς καθόλου εξερεύνηση.

Ερώτημα 10

Ποια από τις παρακάτω προτάσεις σχετικά με την αρχικοποίηση βαρών σε ένα Multilayer Perceptron (MLP) είναι αληθής;

a) Όσο πιο μεγάλες είναι οι αρχικές τιμές που θα επιλεγούν, τόσο πιο γρήγορα θα συγκλίνει το δίκτυο σε λύση.
b) Οι πολλαπλές εκπαιδεύσεις ξεκινώντας από διαφορετικές αρχικές τιμές κάθε φορά μπορούν να συνεισφέρουν στην αποφυγή παγίδευσης σε τοπικά ελάχιστα.
c) Συνιστάται η αρχικοποίηση να γίνεται σε μικρές ενιαίες τιμές.
d) Συνιστάται η αρχικοποίηση να γίνεται σε διαφορετικές και σχετικά μεγάλες τιμές.

Λύση

---

a) Λάθος. Τα μεγάλα αρχικά βάρη μπορεί να οδηγήσουν σε πολύ μεγάλες τιμές στις ενεργοποιήσεις και τις κλίσεις των νευρώνων κατά τη διάρκεια της εκπαίδευσης, προκαλώντας προβλήματα όπως εκρηκτικές κλίσεις ή αστάθεια στη διαδικασία εκπαίδευσης. Αυτό μπορεί στην πραγματικότητα να επιβραδύνει ή ακόμα και να αποτρέψει τη σύγκλιση.
b) **Σωστό**. Η εκπαίδευση του δικτύου πολλές φορές με διαφορετικά τυχαία αρχικά βάρη μπορεί να συμβάλει στην εύρεση μιας καλύτερης λύσης αποφεύγοντας ενδεχομένως τα τοπικά ελάχιστα. Αυτή η προσέγγιση αυξάνει τις πιθανότητες εύρεσης ενός πιο βέλτιστου συνόλου βαρών.
c) Λάθος. Η αρχικοποίηση όλων των βαρών στην ίδια μικρή τιμή (ή στο μηδέν) μπορεί να εμποδίσει το δίκτυο να μάθει σωστά, επειδή μπορεί να οδηγήσει σε συμμετρία, όπου κάθε νευρώνας σε ένα στρώμα μαθαίνει το ίδιο πράγμα. Αντ' αυτού, χρησιμοποιούνται συνήθως μικρές τυχαίες τιμές για να σπάσει η συμμετρία.
d) Λάθος. Ενώ είναι σημαντικό να αρχικοποιούνται τα βάρη σε διαφορετικές τιμές για να σπάσει η συμμετρία, δεν συνιστώνται σχετικά μεγάλες τιμές για τους λόγους που αναφέρονται στο (α). Αντ' αυτού, τα βάρη συνήθως αρχικοποιούνται σε μικρές τυχαίες τιμές, συχνά χρησιμοποιώντας μεθόδους που λαμβάνουν υπόψη τον αριθμό των νευρώνων στα επίπεδα (π.χ. αρχικοποίηση Xavier ή He).

Ερώτημα 11

Για να εκτιμήσουμε την επαγωγική μεροληψία ενός ταξινομητή θα χρειαστούμε:

a) Μόνο το σύνολο υποθέσεων.
b) Μόνο το σύνολο δεδομένων εκπαίδευσης.
c) Το σύνολο υποθέσεων και τον ιδανικό ταξινομητή (όχι το σύνολο δεδομένων).
d) Το σύνολο υποθέσεων, τον ιδανικό ταξινομητή και το σύνολο δεδομένων.

Λύση

Το **(δ)**. Η επαγωγική προκατάληψη αναφέρεται στο σύνολο των υποθέσεων που κάνει ένας αλγόριθμος μάθησης για να προβλέψει τις εξόδους δεδομένων εισόδων που δεν έχει συναντήσει. Είναι αυτό που επιτρέπει στον αλγόριθμο να γενικεύει πέρα από τα δεδομένα εκπαίδευσης. Η εκτίμηση της επαγωγικής προκατάληψης περιλαμβάνει την κατανόηση της σχέσης μεταξύ του συνόλου υποθέσεων, των δεδομένων εκπαίδευσης και του ιδανικού ταξινομητή.

Ερώτημα 12

Για έναν ταξινομητή σε χώρο 2 διαστάσεων που χρησιμοποιεί το σύνολο υποθέσεων όλων των νευρώνων τύπου perceptron ισχύει ότι:

a) Είναι PAC learnable αφού οι νευρώνες τύπου perceptron έχουν άπειρη διάσταση VC.
b) Είναι PAC learnable αφού οι νευρώνες τύπου perceptron έχουν διάσταση VC $k = 3$.
c) Είναι PAC learnable αφού οι νευρώνες τύπου perceptron έχουν διάσταση VC $k = 4$.
d) Καμία από τις παραπάνω.

Λύση

Το **(β)**. Η διάσταση VC είναι ένα μέτρο της χωρητικότητας μιας κλάσης υποθέσεων, που ορίζεται ως ο μεγαλύτερος αριθμός σημείων που μπορούν να κατακερματιστούν (δηλαδή να ταξινομηθούν σωστά με όλους τους δυνατούς τρόπους) από τις υποθέσεις της κλάσης. Για ένα perceptron σε ένα χώρο d διαστάσεων, η διάσταση VC είναι d+1. Αυτό το αποτέλεσμα προκύπτει από το γεγονός ότι ένα perceptron μπορεί να συντρίψει d+1 σημεία σε χώρο d διάστασης.

Ερώτημα 13

Δίνεται ένα σύνολο δεδομένων με n στοιχεία που περιγράφονται από χαρακτηριστικά που αποτιμώνται ως Αληθή ή ως Ψευδή και ανήκουν σε μία από τις 2 κατηγορίες εξόδου. Έστω ένας αλγόριθμος εύρεσης δέντρου απόφασης που ταξινομεί τα στοιχεία του συνόλου δεδομένων που είναι παραλλαγή του ID3 εφαρμόζοντας την εξής τεχνική: αν σε κάποιο βήμα το σύνολο δεδομένων έχει λιγότερα από k στοιχεία ($n \gg k$), δεν παράγει νέο κόμβο στο δέντρο. Ισχύει ότι:

a) Ο αλγόριθμος χωρίς να είναι γενικά αποδοτικός, υπολογίζει 1 δέντρο το οποίο δεν ταξινομεί πάντα σωστά τα στοιχεία.
b) Ο αλγόριθμος χωρίς να είναι γενικά αποδοτικός, υπολογίζει 1 δέντρο το οποίο ταξινομεί πάντα σωστά τα στοιχεία.
c) Ο αλγόριθμος υπολογίζει αποδοτικά 1 δέντρο, το οποίο ταξινομεί πάντα σωστά τα στοιχεία.
d) Ο αλγόριθμος υπολογίζει αποδοτικά 1 δέντρο, το οποίο δεν ταξινομεί πάντα σωστά τα στοιχεία.

Λύση

Το **(δ)**.

- **Αποτελεσματικότητα**: Ο αλγόριθμος δεν συνεχίζει να χωρίζει τα δεδομένα μόλις το μέγεθος του υποσυνόλου πέσει κάτω από k. Αυτό μπορεί να κάνει τον αλγόριθμο πιο αποδοτικό, αποτρέποντας περιττούς υπολογισμούς σε πολύ μικρά υποσύνολα.
- **Ακρίβεια**: Το δέντρο απόφασης μπορεί να μην ταξινομεί πάντα σωστά όλα τα στοιχεία, ειδικά εκείνα που βρίσκονται σε υποσύνολα μικρότερα από k, καθώς το δέντρο σταματά να χωρίζει και έτσι μπορεί να μην καταγράφει όλες τις αποχρώσεις των δεδομένων.

Ερώτημα 14

Δίνεται ένα σύνολο δεδομένων με στοιχεία που περιγράφονται από χαρακτηριστικά που αποτιμώνται ως Αληθή η Ψευδή και ανήκουν σε μία από τις 2 κατηγορίες εξόδου. Έστω ότι χρησιμοποιούμε για την εύρεση ενός δέντρου απόφασης τον αλγόριθμο ID3 και τον αλγόριθμο CART. Ισχύει ότι:

a) Ο αλγόριθμος ID3 υπολογίζει ευκολότερα το δέντρο από τον αλγόριθμο CART (χωρίς να είναι πιο αποδοτικός), χωρίς να διαφοροποιούνται ιδιαίτερα ως προς την επιλογή χαρακτηριστικού απόφασης.
b) Ο αλγόριθμος CART υπολογίζει ευκολότερα το δέντρο από τον αλγόριθμο ID3 (χωρίς να είναι πιο αποδοτικός), χωρίς να διαφοροποιούνται ιδιαίτερα ως προς την επιλογή χαρακτηριστικού απόφασης.
c) Οι 2 αλγόριθμοι διαφοροποιούνται σημαντικά ως προς την επιλογή χαρακτηριστικού απόφασης.
d) Οι 2 αλγόριθμοι διαφοροποιούνται σημαντικά ως προς την απόδοση τους (ο ένας είναι αποδοτικότερος).

Λύση

Το **(γ)**.

- Αλγόριθμος ID3:
  - **Επιλογή χαρακτηριστικών απόφασης**: Ο ID3 χρησιμοποιεί το κέρδος πληροφορίας ως κριτήριο για την επιλογή του καλύτερου χαρακτηριστικού για τον διαχωρισμό των δεδομένων σε κάθε κόμβο.
  - **Αποτελεσματικότητα**: Ο ID3 μπορεί να είναι υπολογιστικά εντατικός, καθώς υπολογίζει την εντροπία και το κέρδος πληροφορίας για κάθε χαρακτηριστικό σε κάθε κόμβο.
- Αλγόριθμος CART:
  - **Επιλογή χαρακτηριστικών απόφασης**: Ο CART χρησιμοποιεί την Gini impurity (ή μερικές φορές την εντροπία) για την επιλογή του καλύτερου χαρακτηριστικού για το διαχωρισμό των δεδομένων.
  - **Αποτελεσματικότητα**: Ο CART, όπως και ο ID3, αξιολογεί κάθε χαρακτηριστικό για να βρει την καλύτερη διάσπαση, αλλά υποστηρίζει επίσης δέντρα παλινδρόμησης (χρησιμοποιώντας διαφορετικά κριτήρια όπως το μέσο τετραγωνικό σφάλμα για τις διασπάσεις).

Ερώτημα 15

Δίνεται το παρακάτω σύνολο δεδομένων. Κατά τη διαδικασία εκτέλεσης του αλγορίθμου ID3, ποιο χαρακτηριστικό επιλέγεται στη ρίζα του δέντρου απόφασης;

a) Temperature
b) Outlook
c) Επειδή περισσότερα από 1 χαρακτηριστικά έχουν το ίδιο κέρδος πληροφορίας, ο αλγόριθμος θα τερματίσει.
d) Επειδή περισσότερα από 1 χαρακτηριστικά έχουν το ίδιο κέρδος πληροφορίας, ο αλγόριθμος θα επιλέξει 1 τυχαία ένα από αυτά.

Outlook	Temperature	Play Tennis
Sunny	Hot	No
Sunny	Cool	Yes
Overcast	Hot	Yes
Overcast	Hot	No

Λύση

Το **(α)**. Υπολογίζουμε αρχικά την εντροπία για όλα τα δεδομένα:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Εδώ έχουμε:

$$p(\text{Yes}) = \frac{2}{4} = 0.5, \qquad p(\text{No}) = \frac{2}{4} = 0.5$$

Άρα:

$$\text{Entropy}(S) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = -0.5 \cdot (-1) - 0.5 \cdot (-1) = 1$$

Στην συνέχεια, υπολογίζουμε το κέρδος πληροφορίας για κάθε χαρακτηριστικό:

- Outlook:
    - Sunny: 2 instances (Hot, No), (Cool, Yes)

- Overcast: 2 instances (Hot, Yes), (Hot, No)

Entropy for Sunny:

- Play Tennis = Yes: 1 instance

- Play Tennis = No: 1 instance

\[\text{Entropy}(Sunny) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1 \]

Entropy for Overcast:

- Play Tennis = Yes: 1 instance

- Play Tennis = No: 1 instance

$$ \text{Entropy}(Overcast) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1 $$

$$ \text{Weighted Entropy}(Outlook) = \frac{2}{4} \cdot 1 + \frac{2}{4} \cdot 1 = 1 $$

$$ \text{Information Gain}(Outlook) = \text{Entropy}(S) - \text{Weighted Entropy}(Outlook) $$

$$ \text{Information Gain}(Outlook) = 1 - 1 = 0 $$

Attribute: Temperature

Subsets for Temperature:

- Hot: 3 instances (Sunny, Hot, No), (Overcast, Hot, Yes), (Overcast, Hot, No)

- Cool: 1 instance (Sunny, Cool, Yes)

Entropy for Hot:

- Play Tennis = Yes: 1 instance

- Play Tennis = No: 2 instances

$$ p(\text{Yes}) = \frac{1}{3}, \, p(\text{No}) = \frac{2}{3} $$

$$ \text{Entropy}(Hot) = -\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \left(\frac{2}{3}\right) $$

$$ \text{Entropy}(Hot) = -\frac{1}{3} \cdot -1.585 - \frac{2}{3} \cdot -0.585 $$

$$ \text{Entropy}(Hot) \approx 0.918 $$

Entropy for Cool:

- Play Tennis = Yes: 1 instance

- Play Tennis = No: 0 instances

$$ \text{Entropy}(Cool) = -1 \log_2(1) - 0 \log_2(0) = 0 $$

$$ \text{Weighted Entropy}(Temperature) = \frac{3}{4} \cdot 0.918 + \frac{1}{4} \cdot 0 = 0.6885 $$

$$ \text{Information Gain}(Temperature) = \text{Entropy}(S) - \text{Weighted Entropy}(Temperature) $$

$$ \text{Information Gain}(Temperature) = 1 - 0.6885 = 0.3115 $$

Conclusion:

- **Outlook** has an information gain of 0
- **Temperature** has an information gain of 0.3115

The attribute with the highest information gain is **Temperature**.

Therefore, the correct answer is:

a) **Temperature**