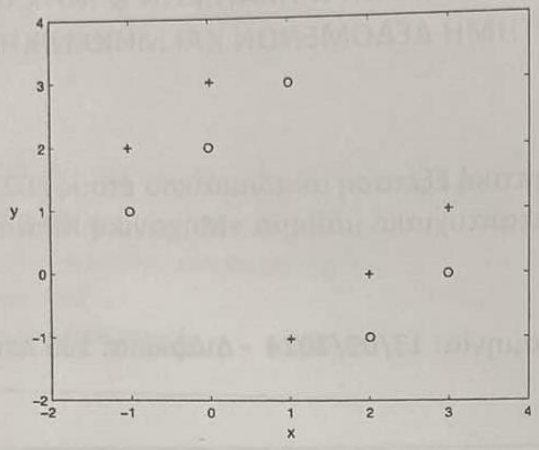
Θέματα ML ΕΔΕΜΜ 2023-2024 ΣΕΠΤΕΜΒΡΙΟΣ

Μαζί με λύσεις, οι οποίες δεν είναι σίγουρα σωστές

ΘΕΜΑ Ι. [8 μονάδες] Έστω ότι εφαρμόζουμε τον ταξινομητή k-πλησιέστερων γειτόνων (k-NN) χρησιμοποιώντας Ευκλείδεια απόσταση για το σύνολο δεδομένων (10 δείγματα) του παρακάτω διαγράμματος.



Σχήμα 1: Δεδομένα για εφαρμογή k-NN

(α) Ποιο θα είναι το σφάλμα στο leave one out cross validation (10-fold cross validation) εάν εφαρμόσουμε 1-NN;

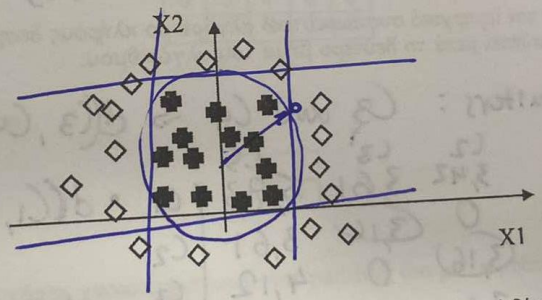
Το σφάλμα θα είναι 100% γιατί για όλα τα δείγματα ο κοντινότερος γείτονας ανήκει στην αντίθετη κλάση, άρα όποιο δείγμα κάνουμε leave out θα ταξινομηθεί σε λάθος κλάση.

(β) Ποια από τις παρακάτω τιμές του k οδηγεί σε μικρότερη τιμή σφάλματος: 3, 5 ή 9; Ποια είναι η τιμή του σφάλματος για αυτό το k;

Γενικά καλοί κανόνας είναι $k < \sqrt{N} = \sqrt{10}$ άρα όλα

ΘΕΜΑ ΙΙ. [12 μονάδες] Σας ζητείται να σχεδιάσετε νευρωνικό δίκτυο MLP (το κατά το δυνατόν πιο απλό) για ταξινόμηση διδιάστατων δεδομένων $X = (x_1, x_2)$ σε δύο κλάσεις. Δίνονται τα δεδομένα εκπαίδευσης που φαίνονται στο Σχήμα.

Περιγράψτε τη δομή του δικτύου, αναφερθείτε στις εισόδους, εξόδους και στο πλήθος νευρώνων κάθε στρώματος, εξηγώντας (ποιοτικά) το ρόλο (λειτουργία) κάθε νευρώνα.



Σχήμα 2: Δεδομένα εκπαίδευσης για το νευρωνικό δίκτυο

ΘΕΜΑ ΙΙΙ. [10 μονάδες] (α) Δίνεται ο παρακάτω πίνακας εγγύτητας 5 σημείων $\mathbf{x}_i \in \mathbb{R}^2, i = 1, 2, \ldots, 5$:

$$P(X) = \begin{matrix} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \end{matrix} \begin{matrix} \begin{matrix} C_1 & C_2 & C_3 & C_4 & C_5 \end{matrix} \\ \begin{bmatrix} 0 & 3.42 & 2.24 & 3.61 & 5.83 \\ 3.42 & 0 & 2.83 & 3.16 & 3.61 \\ 2.24 & 2.83 & 0 & 1.41 & 4.12 \\ 3.61 & 3.16 & 1.41 & 0 & 2.15 \\ 5.83 & 3.61 & 4.12 & 2.15 & 0 \end{bmatrix} \end{matrix}$$

Αν με βάση τον $P(X)$ εφαρμόσουμε τον ιεραρχικό συσσωρευτικό αλγόριθμο πλήρους δεσμού (complete link - MAX), να δείξετε ποια θα είναι η συσταδοποίηση που προκύπτει μετά το δεύτερο βήμα του αλγόριθμου.

1st step : Merge 2 closest clusters : C_3 and C_4 $\rightarrow$ $d(C_3, C_4) = 1,41 \rightarrow C_3'$ cluster

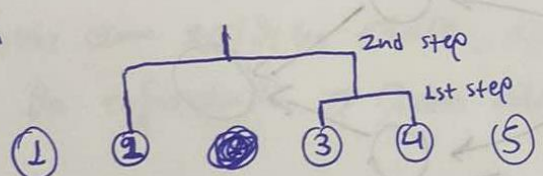
Recompute $P(x)$ =

	C_1	C_2	C_3'	C_5
C_1	0	3,42	3,61	5,83
C_2	3,42	0	3,16	3,61
C_3'	3,61	3,16	0	4,12
C_5	5,83	3,61	4,12	0

- $d(C_1, C_3') = \max(d(C_1, C_3), d(C_1, C_4)) = \max(2,24, 3,61) = 3,61$
- $d(C_2, C_3') = \max(2,83, 3,16) = 3,16$
- $d(C_5, C_3') = \max(4,12, 2,15) = 4,12$

- 2nd step : 2 closest clusters : C_2 and C_3' $\rightarrow$ $d(C_2, C_3') = 3,16$

Dendrogram



(β) Εφαρμόζουμε τον αλγόριθμο k-means για τον διαχωρισμό των σημείων $\mathbf{x}_1 = (2, 3), \mathbf{x}_2 = (3, 3), \mathbf{x}_3 = (6, 7), \mathbf{x}_4 = (8, 8), \mathbf{x}_5 = (7, 5), \mathbf{x}_6 = (9, 6)$ σε δύο ομάδες. Αν τα κέντρα των ομάδων αρχικοποιούνται ως $\theta_1^{(0)} = (2, 3)$ και $\theta_2^{(0)} = (8, 8)$, να υπολογίσετε και να δείξετε ποιες θα είναι θέσεις των δύο κέντρων μετά την ολοκλήρωση της πρώτης επανάληψης του αλγόριθμου.

d	θ_1	θ_2
x_1	0	>0
x_2	1	>1
x_3	5,66	2,24
x_4	>0	0
x_5	5,39	3,16
x_6	7,62	2,23

Clusters

$C_1 = \{x_1, x_2\}$

$C_2 = \{x_3, x_4, x_5, x_6\}$

New centroids : $\theta_1^{(1)} = \left(\frac{2+3}{2}, \frac{3+3}{2}\right) = (2,5, 3)$

$\theta_2^{(1)} = \left(\frac{6+8+7+9}{4}, \frac{7+8+5+6}{4}\right) = (7,5, 6,5)$

ΘΕΜΑ IV. [8 μονάδες] Έστω ότι έχουμε τα παρακάτω δεδομένα 2 διαστάσεων:

1	0	1	0	0
0	1	1	1	1
1	0	1	1	0
0	0	1	0	0
1	1	0	1	0

Να εφαρμόσετε συνέλιξη με το φίλτρο

4	1
3	5

και να υπολογίσετε τον παραγόμενο χάρτη χαρακτηριστικών για padding ίσο με το μηδέν και για stride ίσο με το 1 και στις δύο διαστάσεις.

$P=0$, $S=1$

$$Lout = \frac{Lin - F}{S} + 1 = \frac{5-2}{1} + 1 = 4$$

Output

9	9	12	8
4	10	13	8
4	6	8	4
8	4	9	3

$C_{1,1} = (4\cdot1)+(1\cdot0)+(3\cdot0)+(5\cdot1) = 9$

$C_{1,2} = 4\cdot0 + 1\cdot1 + 3\cdot1 + 5\cdot1 = 9$

$C_{1,3} = 4\cdot1 + 1\cdot0 + 3\cdot1 + 5\cdot1 = 12$

$C_{1,4} = 4\cdot0 + 1\cdot0 + 3\cdot1 + 5\cdot1 = 8$

$C_{2,1} = 1+3 = 4$

$C_{2,2} = 4+1+5 = 10$

$C_{2,3} = 4+1+3+5 = 13$

$C_{2,4} = 4+1+3 = 8$

$C_{3,1} = 4$

$C_{3,2} = 1+5 = 6$

$C_{3,3} = 4+1+3 = 8$

$C_{3,4} = 4$

$C_{4,1} = 3+5 = 8$

$C_{4,2} = 1+3 = 4$

$C_{4,3} = 4+5 = 9$

$C_{4,4} = 3$

ΘΕΜΑ V. [30 μονάδες] Στις ακόλουθες ερωτήσεις επιλέξτε ευχρινώς ποια (μοναδική) απάντηση (Α, Β, Γ ή Δ) θεωρείτε ότι είναι η ορθή. Ορθή απάντηση: 3 μονάδες. Λάθος απάντηση: -0.75 μονάδα. Κενή απάντηση: 0 μονάδες.

1. Δίνεται το σύνολο δεδομένων του Πίνακα 1. Το πλήθος των διαφορετικών δέντρων απόφασης που ταξινομούν σωστά τα δεδομένα είναι:

   (Α′) 2.
   (Β′) 4.
   (Γ′) 8.
   (Δ′) Καμία από τις άλλες απαντήσεις δεν είναι σωστή.

Πίνακας 1: Σύνολο δεδομένων $\mathbb{D}$

X1	X2	Y
TRUE	TRUE	No
TRUE	FALSE	Yes
FALSE	TRUE	Yes
FALSE	FALSE	No

---

2. Δίνονται οι παρακάτω προτάσεις:

   (Α′) Κάθε δέντρο απόφασης μπορεί να μετασχηματιστεί άμεσα σε μία λογική πρόταση σε κανονική διαζευκτική μορφή. Σ
   (Β′) Το πρόβλημα εύρεσης του ελάχιστου δέντρου απόφασης δεν μπορεί πάντα να επιλυθεί. NP-complete Σ
   (Γ′) Ο αλγόριθμος CART χρησιμοποιεί το μέτρο της εντροπίας για την επιλογή του χαρακτηριστικού απόφασης. Λ

   Πόσες από τις παραπάνω προτάσεις ισχύουν;

   (Α′) Καμία
   (Β′) 1
   (Γ′) 2
   (Δ′) 3

---

3. Έστω ένα σύνολο δεδομένων και δύο διαφορετικά δέντρα απόφασης που ταξινομούν σωστά όλα τα δεδομένα, χωρίς σφάλμα. Ισχύει ότι:

   (Α′) Τα δύο δέντρα έχουν την ίδια ικανότητα γενίκευσης, ανεξάρτητα από το ύψος τους.
   (Β′) Το υψηλότερο δέντρο έχει καλύτερη γενίκευση.
   (Γ′) Το υψηλότερο δέντρο έχει χειρότερη γενίκευση.
   (Δ′) Καμία από τις άλλες προτάσεις δεν είναι σωστή.

---

4. Ένα τυχαίο δάσος κατασκευάζεται με τμηματοποίηση χαρακτηριστικών (feature bagging). Ισχύει ότι:

   (Α′) Τα δέντρα που το απαρτίζουν έχουν γενικά κοινούς κόμβους χαρακτηριστικών (παρατηρούμε ίδιους κόμβους σε διαφορετικά δέντρα).
   (Β′) Τα δέντρα που το απαρτίζουν έχουν διαφορετικούς κόμβους χαρακτηριστικών (δεν παρατηρούμε ίδιους κόμβους σε διαφορετικά δέντρα).
   (Γ′) Τα δέντρα που το απαρτίζουν έχουν τους ίδιους κόμβους χαρακτηριστικών (παρατηρούμε ακριβώς τους ίδιους κόμβους σε όλα τα δέντρα, πιθανά σε διαφορετική διάταξη).
   (Δ′) Καμία από τις άλλες προτάσεις δεν είναι σωστή.

---

5. 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 πιθανοφάνειες παρατήρησης ενός χαρακτηριστικού $x \in \mathbb{R}$ στις δύο κλάσεις είναι $P(x|\omega_1)$ και $P(x|\omega_2)$, αντίστοιχα. Σύμφωνα με το θεώρημα Bayes, ποιες από τις παρακάτω εκφράσεις δίνουν την εκ των υστέρων πιθανότητα $P(\omega_1|x)$;

I. $P(\omega_1|x) = \frac{P(x|\omega_1)P(\omega_1)}{P(x)}$

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

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

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

(Α′) I και II
(Β′) I και III
(Γ′) I και IV
(Δ′) III και IV

---

6. Ποιοι από τους παρακάτω ισχυρισμούς είναι λανθασμένοι σχετικά με τον αλγόριθμο ομαδοποίησης k-means;
(I): Ο k-means βρίσκει εγγυημένα την ολικά βέστιστη λύση για τα κέντρα των κλάσεων.
(II): Ο k-means αναθέτει κάθε σημείο στην εγγύτερή του κλάση με βάση την Ευκλείδεια απόσταση.
(III): Ο k-means μπορεί να χειριστεί αποτελεσματικά μη-σφαιρικές κλάσεις, ανεξάρτητα από το σχήμα τους.
(IV): Η υπολογιστική πολυπλοκότητα του k-means είναι $O(nklt)$, όπου n είναι ο αριθμός των σημείων, k ο αριθμός των κλάσεων, l ο αριθμός των χαρακτηριστικών και t ο αριθμός των επαναλήψεων του αλγόριθμου.

(Α′) I και II
(Β′) I και III
(Γ′) III και IV
(Δ′) I, III και IV

(Ε) I, II, III

---

7. Έστω Multi Layer Perceptron $MLP1$ με γραμμικές συναρτήσεις ενεργοποίησης. Τι από τα παρακάτω ισχύει;

(Α′) Μπορούμε να βρούμε ένα ισοδύναμο νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα μόνο εάν πληρούνται συγκεκριμένες προϋποθέσεις για τη δομή του $MLP1$.
(Β′) Μπορούμε να βρούμε ένα ισοδύναμο νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα μόνο εάν πληρούνται συγκεκριμένες προϋποθέσεις για τις εισόδους του $MLP1$.
(Γ′) Μπορούμε να βρούμε ένα ισοδύναμο νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα ανεξαρτήτως της δομής και των εισόδων του $MLP1$.
(Δ′) Δεν μπορούμε να βρούμε ισοδύναμο νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα.

---

8. Ποιοι από τους παρακάτω ισχυρισμούς για τα SVMs είναι ορθοί;

(I): Αν αφαιρέσουμε ένα σημείο που ταξινομείται ορθά και βρίσκεται μακριά από το όριο απόφασης, τότε το όριο απόφασης (βέλτιστο υπερεπίπεδο διαχωρισμού) δεν θα επηρεαστεί.
(II): Με τη χρήση συναρτήσεων πυρήνα (kernel functions) γίνεται έμμεσα απεικόνιση των δεδομένων σε μη γραμμικό χώρο χωρίς να εμφανίζεται πουθενά στις πράξεις η συνάρτηση μετασχηματισμού $\Phi()$ μόνη της.
(III): Αν έχουμε ένα πρόβλημα ταξινόμησης τριών (3) κλάσεων, τότε το πλήθος των δυαδικών SVM που θα πρέπει να εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-one είναι μικρότερο από το πλήθος των δυαδικών SVM που θα πρέπει να εκπαιδεύσουμε αν ακολουθήσουμε τη μέθοδο one-against-all.

(Α′) I και II
(Β′) I και III
(Γ′) II και III
(Δ′) I, II και III

---

9. 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;

(Α΄) Το πρόβλημα είναι γραμμικά διαχωρίσιμο.

(Β΄) Το πρόβλημα, αν και όχι αυστηρά γραμμικά διαχωρίσιμο, είναι σχεδόν γραμμικά διαχωρίσιμο, γι΄ αυτό και ένα δίκτυο ADALINE μπορεί να συγκλίνει επιτρέποντας μικρό αριθμό λανθασμένων ταξινομήσεων.

(Γ΄) Η συνάρτηση δεν μπορεί να αναπαρασταθεί από απλό perceptron, αλλά μπορεί να αναπαρασταθεί από πολυστρωματικό perceptron (MLP) με ένα κρυμμένο στρώμα δύο νευρώνων και στρώμα εξόδου ενός νευρώνα.

(Δ΄) Καμία από τις παραπάνω.

---

10. Τι από τα παρακάτω θα συμβεί εάν αυξήσουμε την τιμή της υπερπαραμέτρου C σε ένα Support Vector Machine (SVM);

(Α΄) Το σφάλμα εκπαίδευσης τείνει να μειωθεί.

(Β΄) Το περιθώριο διαχωρισμού margin θα παραμείνει σίγουρα αμετάβλητο.

(Γ΄) Και οι δύο παραπάνω προτάσεις είναι σωστές.

(Δ΄) Καμία από τις παραπάνω προτάσεις δεν είναι σωστή.
