ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

7ο εξάμηνο — Ακαδημαϊκό έτος 2024-2025

2η Σειρά Ασκήσεων

Ημερ. Παράδ.: 08.01.2025

Γενικές Οδηγίες: Οι αναλυτικές σειρές ασκήσεων είναι ατομικές, και οι λύσεις που θα δώσετε πρέπει να αντιπροσωπεύουν μόνο την προσωπική σας εργασία. Εξηγήστε επαρκώς την εργασία σας. Αν χρησιμοποιήσετε κάποια άλλη πηγή εκτός των σημειώσεων για την λύση σας, πρέπει να το αναφέρετε. Η παράδοση των λύσεων των αναλυτικών ασκήσεων της σειράς αυτής θα γίνει ηλεκτρονικά στην HELIOS ιστοσελίδα του μαθήματος και θα πρέπει να την υποβάλετε ως ένα ενιαίο αρχείο PDF με το εξής filename format χρησιμοποιώντας μόνο λατινικούς χαρακτήρες: ML24_hwk2_AM_LastnameFirstname.pdf, όπου AM ειναι ο 8-ψήφιος αριθμός μητρώου σας. Σκαναριμένες χειρόγραφες λύσεις επιτρέπονται αρκεί να είναι καθαρογραμμένες και ευανάγνωστες. Επίσης στην 1η σελίδα των λύσεων θα αναγράφετε το ονοματεπώνυμο, A.M., και email address σας. Συμπεριλάβετε και τον κώδικα προγραμμάτων, π.χ. Matlab ή Python, που χρησιμοποιήσατε για αριθμητική επίλυση. Να σημειωθεί ότι η καταληκτική ημερομηνία παράδοσης είναι τελική και δεν θα δοθεί παράταση.

Άσκηση 2.1 (Decision Trees - Random Forests)

Δίνονται οι παρατηρήσεις που απεικονίζονται στο παρακάτω Σχήμα, όπου X_1, X_2, X_3, X_4 τα χαρακτηριστικά και Y η κλάση ταξινόμησης.

(α) Να υπολογίσετε το δέντρο απόφασης που προκύπτει με τη χρήση του κριτηρίου Gini. Αν δύο χαρακτηριστικά οδηγούν στην ίδια μείωση του κριτηρίου επιλέξτε το πρώτο αλφαβητικά. Εκτελέστε τον αλγόριθμο εκπαίδευσης έως ότου κάθε φύλλο περιέχει δείγματα μίας μόνο κλάσης (ακρίβεια ίση με 1).

(β) Μπορείτε να βρείτε με το χέρι ένα μικρότερο δέντρο με ακρίβεια ίση με 1; Υπόδειξη: Εξετάστε πώς σχετίζονται τα χαρακτηριστικά X_1 και X_3.

(γ) Εκπαιδεύστε ένα Random Forest το οποίο αποτελείται από 3 δέντρα απόφασης. Χρησιμοποιήστε την 1η, 4η και 7η παρατήρηση για την εκπαίδευση του 1ου δέντρου,την 2η, 5η, και 8η για τη εκπαίδευση του 2ου δέντρου και την 3η, 6η, και 9η για την εκπαίδευση του 3ου δέντρου. Η εκπαίδευση του κάθε δέντρου γίνεται με τον τρόπο που υποδηλώνει το 1ο υποερώτημα.

X_1	X_2	X_3	X_4	Y
1	1	0	0	1
1	1	0	1	1
0	0	1	0	1
0	1	1	1	1
1	0	0	0	1
1	0	1	1	-1
0	1	0	1	-1
1	1	1	1	-1
0	0	0	1	-1

Άσκηση 2.2 (Εφαρμογή του k-Means σε συνθετικά και πραγματικά δεδομένα)

Συνθετικά δεδομένα: Δίνεται το σύνολο προτύπων στο $\mathbb{R}^2$,

$$X = \{(2,3), (3,2), (1,2), (4,5), (5,4), (3,4), (6,4), (6,5)\}.$$

(α) Δώστε αναλυτικά (με όλους τους σχετικούς υπολογισμούς) τα διαδοχικά βήματα του αλγόριθμου k-means μέχρι την σύγκλισή του, για $m = 2$ κλάσεις και αρχικοποίηση των κέντρων των κλάσεων στα σημεία $(3,3)$ και $(4,4)$. Ποιες είναι οι θέσεις των τελικών κέντρων και πως κατανέμονται τα σημεία στις δύο κλάσεις;

Πραγματικά δεδομένα: Δίνεται (στην ιστοσελίδα του μαθήματος στο helios) το ευρέως διαδεδομένο σύνολο δεδομένων κρίνων (Iris dataset) του Fisher. Τα δεδομένα αποτελούνται από 3 κλάσεις (για τους 3 διαφορετικούς τύπους κρίνου), καθεμιά από τις οποίες περιλαμβάνει 50 δείγματα. Τα δεδομένα περιγράφονται από 4 διαφορετικά χαρακτηριστικά:

- μήκος σεπάλων σε εκ.
- πλάτος σεπάλων σε εκ.
- μήκος πετάλων σε εκ.
- πλάτος πετάλων σε εκ.
- τύπος κρίνου (Iris Setosa/Iris Versicolour/Iris Virginica)

(β) Υλοποιήστε (σε Python ή Matlab) τον αλγόριθμο k-means με τυχαία αρχικοποίηση και εφαρμόστε τον στο παραπάνω σύνολο δεδομένων. Θέστε τον αριθμό κλάσεων σε $m = 3$. Ο αλγόριθμος τερματίζει όταν ικανοποιείται το κριτήριο που δίνεται στη σελίδα 32 της 10ης σειράς διαφανειών με $\varepsilon = 10^{-5}$. Με βάση την πληροφορία για τον τύπο των κρίνων που υπάρχει στα δεδομένα και την ομαδοποίηση στην οποία καταλήγει ο k-means, υπολογίστε τον confusion matrix και το success rate του αλγόριθμου.
(γ) Επαναλάβετε τα ζητούμενα του ερωτήματος (β) για την περίπτωση που το διάνυσμα χαρακτηριστικών περιλαμβάνει μόνο δύο χαρακτηριστικά, το 'μήκος πετάλων' και το 'πλάτος πετάλων'. Επιπλέον, δώστε γραφική απεικόνιση της ομαδοποίησης στην οποία καταλήγει ο k-means με βάση τα δύο παραπάνω χαρακτηριστικά και των κέντρων των κλάσεων που εκτιμά ο αλγόριθμος.
(δ) Συγκρίνετε τα αποτελέσματα που πήρατε στα ερωτήματα (β) και (γ) και σχολιάστε.

Άσκηση 2.3 (Ιεραρχική Ομαδοποίηση)

(α) Έστω δύο διανύσματα $\mathrm{x}, \mathrm{y} \in \mathbb{R}^l$, και θ_{xy} η μεταξύ τους γωνία. Ορίζουμε την συνάρτηση εγγύτητας $d_c(\mathrm{x}, \mathrm{y})$ ως εξής:

$$d_c(\mathrm{x}, \mathrm{y}) = 1 - \cos\theta_{xy}.$$

Να αποδείξετε ότι η $d_c(\mathrm{x}, \mathrm{y})$ είναι μέτρο ανομοιότητας. Είναι η $d_c(\mathrm{x}, \mathrm{y})$ μετρική ανομοιότητας;
(β) Δίνεται το σύνολο προτύπων στο $\mathbb{R}^2$,

$$X = \{(1,5), (3,4), (0,2), (5,4), (2,6), (3,3), (2,3), (4,2)\}.$$

Δώστε τον πίνακα προτύπων $D(X)$ και προσδιορίστε τον πίνακα εγγύτητας $P(X)$ με βάση την μετρική d_c.
(γ) Με βάση τον πίνακα εγγύτητας $P(X)$ που υπολογίσατε στο ερώτημα (β), περιγράψτε αναλυτικά τις διαδοχικές ομαδοποιήσεις που θα προκύψουν από την εφαρμογή του ιεραρχικού αλγόριθμου απλού δεσμού, καθώς και το αντίστοιχο δενδρόγραμμα εγγύτητας.
(δ) Επαναλάβετε τα ζητούμενα στο ερώτημα (γ) για τον ιεραρχικό αλγόριθμο πλήρους δεσμού.
(ε) Συγκρίνετε και σχολιάστε τα αποτελέσματα που πήρατε στα δύο προηγούμενα ερωτήματα και προσδιορίστε τη βέλτιστη ομαδοποίηση σε κάθε περίπτωση από τα αντίστοιχα δενδρογράμματα.

Άσκηση 2.4 (Θεωρία PCA)

Δίνεται μια ακολουθία δεδομένων (τυχαία διανύσματα με μηδενικό μέσο) $x_n \in \mathbb{R}^l$, $n = 1, ..., N$, με δειγματικό πίνακα συνδιασποράς (sample covariance matrix):

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T.$$

Μεταξύ των γραμμικών υπόχωρων διάστασης $m \leq l$, θέλουμε να βρούμε εκείνον στον οποίο η μεταβλητότητα (διασπορά) των προβολών των δεδομένων μεγιστοποιείται. Ο υπόχωρος ορίζεται με βάση m αμοιβαία ορθογώνια μοναδιαία διανύσματα $\mathrm{u}_i, i = 1, \ldots m$ που αντιστοιχούν σε m άξονες ή κατευθύνσεις.
(α) Για $m = 1$, δείξτε ότι το διάνυσμα u_1 (που καθορίζει τον βέλτιστο υπόχωρο διάστασης 1) είναι το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή λ_1 του $\hat{\Sigma}$.
(β) Για $m = 2$, δείξτε ότι τα $\mathrm{u}_1, \mathrm{u}_2$ είναι τα ιδιοδιανύσματα που αντιστοιχούν στις δύο μεγαλύτερες ιδιοτιμές λ_1, λ_2 του $\hat{\Sigma}$.
(γ) Γενικεύστε τα (α) και (β) και αποδείξτε ότι τα βέλτιστα $\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_m$ είναι τα ιδιοδιανύσματα που αντιστοιχούν στις m μεγαλύτερες ιδιοτιμές $\lambda_1, \lambda_2, \ldots, \lambda_m$ του $\hat{\Sigma}$.
(δ) Εξηγείστε γιατί το ποσοστό p της διασποράς των δεδομένων που 'ερμηνεύεται' από m κύριες συνιστώσες δίνεται από τη σχέση:

$$p = N\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{n=1}^{N}\|\mathrm{x}_n\|^2}.$$

Υπόδειξη: Στα (α),(β),(γ), να προσδιορίσετε την διασπορά της προβολής των δεδομένων σε ένα μοναδιαίο διάνυσμα u και να ορίσετε κατά περίπτωση κατάλληλα προβλήματα βελτιστοποίησης λαμβάνοντας υπόψη τους περιορισμούς $\|\mathrm{u}_i\| = 1, \forall i$ και $\mathrm{u}_i^T\mathrm{u}_j = 0, i \neq j$. Στη συνέχεια να επιλύσετε τα προβλήματα αυτά με χρήση της μεθόδου των πολλαπλασιαστών Lagrange.

Άσκηση 2.5 (Markov Decision Processes - Reinforcement Learning)

Θεωρούμε ότι ένα ρομπότ κινείται σε ένα πλέγμα 4×4, όπου το τετράγωνο (3,3) είναι ο στόχος και το τετράγωνο (2,2) είναι παγίδα. Το ρομπότ μπορεί να κινηθεί πάνω, κάτω, αριστερά ή δεξιά και λαμβάνει ανταμοιβές ως εξής:

- +10 αν φτάσει στον στόχο (3,3)
- -1 για κάθε κίνηση
- -10 αν καταλήξει στην παγίδα (2,2)

Οι κινήσεις είναι ντετερμινιστικές (δηλαδή το ρομπότ πάντα κινείται προς την κατεύθυνση που επιλέγει, εκτός αν βγαίνει εκτός πλέγματος).

Ζητούνται:

(α) Να καθορίσετε τα στοιχεία μιας MDP για αυτό το πρόβλημα: καταστάσεις, ενέργειες, ανταμοιβές και πιθανότητες μετάβασης.
(β) Να διατυπώσετε την εξίσωση Bellman για την αξία V_s^π μιας κατάστασης s υπό μια πολιτική π.
(γ) Χρησιμοποιώντας την πολιτική που οδηγεί το ρομπότ πάντα προς τον στόχο, υπολογίστε την τιμή της αρχικής κατάστασης (0,0) αν το ρομπότ ακολουθεί αυτή την πολιτική. Υποθέστε ότι $\gamma = 0.9$.
(δ) Πώς επηρεάζει η ύπαρξη της παγίδας (2,2) την τιμή των καταστάσεων γύρω της;