ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

7ο εξάμηνο — Ακαδημαϊκό έτος 2022-2023

2η Σειρά Ασκήσεων

Ημερ. Παράδ.: 22.01.2023

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: ML22_hwk2_AM_LastnameFirstname.pdf, όπου ΑΜ 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, Α.Μ., 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 (Θεωρία Μηχανικής Μάθησης)

Να δείξετε ότι η κλάση $\mathrm{H}^n_{\text{rec}}$ των παράλληλων στους άξονες υπερ-παραλληλογράμμων του $\mathbb{R}^n$ είναι PAC εκπαιδεύσιμη.

Υπόδειξη. Μπορείτε να συμβολίσετε την κλάση ως $\mathrm{H}^n_{\text{rec}} = \{h_{(a_1,a_2,...,a_n,b_1,b_2,...,b_n)} : a_1 \leq b_1, a_2 \leq b_2, ..., a_n \leq b_n\}$, όπου $h_{(a_1,a_2,...,a_n,b_1,b_2,...,b_n)}(x_1, x_2, ..., x_n) = 1$ για εκείνα τα $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$ για τα οποία $a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2, ..., a_n \leq x_n \leq b_n$. Για την απόδειξη, μπορείτε να ακολουθήσετε (με τις απαραίτητες επεκτάσεις) τη μεθοδολογία που ακολουθήσαμε για την αντίστοιχη απόδειξη για $n = 2$. Εναλλακτικά μπορείτε να υπολογίσετε τη διάσταση VC της κλάσης $\mathrm{H}^n_{\text{rec}}$ και να δείξετε ότι είναι πεπερασμένη για κάθε n.

Άσκηση 2.2 (Σύγκριση των Αλγορίθμων k-Means και Fuzzy c-Means)

(α) Όπως αναφέρεται στη σελίδα 38 στη σειρά διαφανειών #9, ο αλγόριθμος fuzzy c-means ελαχιστοποιεί τη συνάρτηση κόστους

$$J(U, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{m} u_{ij}^q d(\mathrm{x}_i, \boldsymbol{\theta}_j), \quad q > 1$$

με τους ακόλουθους περιορισμούς

$$\sum_{j=1}^{m} u_{ij} = 1, \ i = 1, 2, \ldots, N.$$

Υπενθυμίζεται ότι για τις συναρτήσεις συμμετοχής u_{ij} ισχύει επιπλέον $0 < u_{ij} < 1$. Έστω ότι η απόσταση $d(\mathrm{x}_i, \boldsymbol{\theta}_j) = ||\mathrm{x}_i - \boldsymbol{\theta}_j||^2$, δηλαδή ισούται με το τετράγωνο της Ευκλείδειας απόστασης του σημείου x_i από το κέντρο $\boldsymbol{\theta}_j$. Χρησιμοποιώντας τη μέθοδο των πολλαπλασιαστών Lagrange, να δείξετε ότι στο πλαίσιο μιας διαδικασίας εναλλασσόμενης ελαχιστοποίησης της συνάρτησης $J(U, \boldsymbol{\theta})$, μπορούμε να καταλήξουμε στις ακόλουθες αναδρομικές σχέσεις για τα u_{ij} και τις εκτιμήσεις των κέντρων των κλάσεων $\boldsymbol{\theta}_j$

$$u_{ij} = \frac{1}{\sum_{k=1}^{m} \left(\frac{d(\mathrm{x}_i, \boldsymbol{\theta}_j)}{d(\mathrm{x}_i, \boldsymbol{\theta}_k)} \right)^{\frac{1}{q-1}}}, \quad \boldsymbol{\theta}_j = \frac{\sum_{i=1}^{N} u_{ij}^q \mathrm{x}_i}{\sum_{i=1}^{N} u_{ij}^q}.$$

Θα πρέπει να σημειωθεί ότι οι δύο αυτές σχέσεις αποτελούν τα βασικά βήματα του αλγόριθμου fuzzy c-means.
(β) Έστω ένα πρόβλημα ομαδοποίησης δεδομένων σε $m = 2$ κλάσεις, τα σημεία των οποίων ακολουθούν την κανονική κατανομή με μέσα διανύσματα $\boldsymbol{\mu}_1 = [1, 1]^T$, $\boldsymbol{\mu}_2 = [2.5, 2.5]^T$ και πίνακες συμμεταβλητότητας

$$\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 1 & -0.4 \\ -0.4 & 1 \end{bmatrix}.$$

Δημιουργήστε ένα σύνολο δεδομένων με 300 σημεία από κάθε κλάση. Δώστε μια κατάλληλη γραφική αναπαράσταση του συνόλου των δεδομένων και σχολιάστε.

(γ) Υλοποιήστε (σε Python ή Matlab) τον αλγόριθμο k-means με τυχαία αρχικοποίηση και εφαρμόστε τον στο παραπάνω σύνολο δεδομένων. Ο αλγόριθμος τερματίζει όταν ικανοποιείται το κριτήριο που δίνεται στη σελίδα 32 της 9ης σειράς διαφανειών με $\varepsilon = 10^{-4}$. Δώστε γραφική απεικόνιση της ομαδοποίησης στην οποία καταλήγει ο k-means και των κέντρων των κλάσεων που εκτιμά ο αλγόριθμος.
(δ) Επαναλάβετε τα ζητούμενα του ερωτήματος (γ) για τον αλγόριθμο fuzzy c-means με $q = 2$. Ο αλγόριθμος αρχικοποιείται με τα ίδια διανύσματα $\boldsymbol{\theta}_j, j = 1, 2, \ldots, m$ με τα οποία αρχικοποιήθηκε και ο k-means στο προηγούμενο ερώτημα. Μετά τον τερματισμό του fuzzy c-means, το διάνυσμα x_i καταχωρείται στην κλάση j με το μέγιστο $u_{ij}, j = 1, 2, \ldots, m$.
(ε) Με βάση τα αποτελέσματα στα δύο προηγούμενα ερωτήματα, συγκρίνετε τους δύο αλγόριθμους ως προς i) τον αριθμό των επαναλήψεων που απαιτούνται μέχρι τον τερματισμό τους, ii) τη μέση απόσταση των εκτιμώμενων κέντρων από τα πραγματικά κέντρα και iii) το ρυθμό επιτυχίας (success rate). Μπορείτε να τρέξετε τους αλγόριθμους για πολλά διαφορετικά σύνολα δεδομένων [παρόμοια με αυτό του ερωτήματος (β)] για να επιβεβαιώσετε τα συμπεράσματά σας.

Άσκηση 2.3 (Ιεραρχική Ομαδοποίηση)

(α) Θεωρήστε δύο σημεία στο $\mathbb{R}^l$, $\mathrm{x} = [x_1, x_2, ..., x_l]^T$ και $\mathrm{y} = [y_1, y_2, ..., y_l]^T$ και έστω $|x_i - y_i| = \max_{j=1,2,...,l}\{|x_j - y_j|\}$. Ορίζουμε την απόσταση $d_n(\mathrm{x}, \mathrm{y})$ ως εξής

$$d_n(\mathrm{x}, \mathrm{y}) = |x_i - y_i| + \frac{1}{l/2 + 1} \sum_{j=1, j\neq i}^{l} |x_j - y_j|.$$

Να αποδείξετε ότι η d_n είναι μετρική. (Μπορεί να δειχτεί ότι η d_n αποτελεί μια προσέγγιση της Ευκλείδειας απόστασης με χαμηλότερη υπολογιστική πολυπλοκότητα.)
(β) Δίνεται το σύνολο προτύπων στο $\mathbb{R}^2$,

$$X = \{(0, 3), (1.4, 2.6), (-1.5, 3.4), (-0.2, -0.4), (1, -1), (2, -1.5), (2.6, -1.8), (3, -2)\}.$$

Δώστε τον πίνακα προτύπων $D(X)$ και προσδιορίστε τον πίνακα εγγύτητας $P(X)$ με βάση την μετρική d_n.
(γ) Με βάση τον πίνακα εγγύτητας $P(X)$ που υπολογίσατε στο ερώτημα (β), δώστε 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 και SVD)

Μας δίνεται μια ακολουθία δεδομένων (τυχαία διανύσματα με μηδενικό μέσο) $\mathrm{x}_n \in \mathbb{R}^d$, $n = 1, ..., N$, και θέλουμε να βρούμε μια κατεύθυνση (μοναδιαίο διάνυσμα) $\mathrm{e} \in \mathbb{R}^d$ και σταθερές a_n έτσι ώστε, αν προσεγγίσουμε κάθε δεδομένο μας (διάνυσμα στήλης) x_n με ένα διάνυσμα $a_n\mathrm{e}$, το συνολικό μέσο τετραγωνικό λάθος J να είναι ελάχιστο:

$$J(a_1, ..., a_n, \mathrm{e}) = \frac{1}{N} \sum_{n=1}^{N} \|\mathrm{x}_n - a_n\mathrm{e}\|^2, \quad \|\cdot\| = \text{Euclidean norm}$$

(α) Θεωρήστε γνωστό το e και αποδείξετε ότι τα βέλτιστα a_n που ελαχιστοποιούν το J είναι

$$a_n = \langle \mathrm{x}_n, \mathrm{e} \rangle = \mathrm{x}_n^T \mathrm{e}$$

(β) Αντικαθιστώντας τα βέλτιστα a_n στο J, αποδείξετε ότι αποκτούμε ένα λάθος

$$J_1(\mathrm{e}) = -\mathrm{e}^T \mathrm{R}_x \mathrm{e} + (1/N) \sum_{n=1}^{N} \|\mathrm{x}_n\|^2, \quad \mathrm{R}_x = \frac{1}{N} \sum_{n=1}^{N} \mathrm{x}_n \mathrm{x}_n^T,$$

όπου R_x είναι ο εμπειρικός πίνακας αυτοσυσχέτισης των δεδομένων.
(γ) Αποδείξετε ότι το βέλτιστο e που ελαχιστοποιεί το J_1 είναι το ιδιοδιάνυσμα του R_x που αντιστοιχεί στην μέγιστη ιδιοτιμή λ_1. (Υπόδειξη: ελαχιστοποιήστε το J_1 εκμεταλλευόμενοι τον περιορισμό $\|\mathrm{e}\| = 1$ με Lagrange πολλαπλασιαστή.)
(δ) Τι σχέση έχει η ανωτέρω λύση με PCA (Principal Component Analysis)?
(ε) Να βρεθεί αναλυτικά πως σχετίζεται η ανωτέρω λύση με την SVD (Singular Value Decomposition) του $N \times d$ πίνακα X που σχηματίζεται στιβάζοντας τα διανύσματα x_n, $n = 1, ..., N$, ως γραμμές?

Άσκηση 2.5 (Αριθμητική εφαρμογή του PCA)

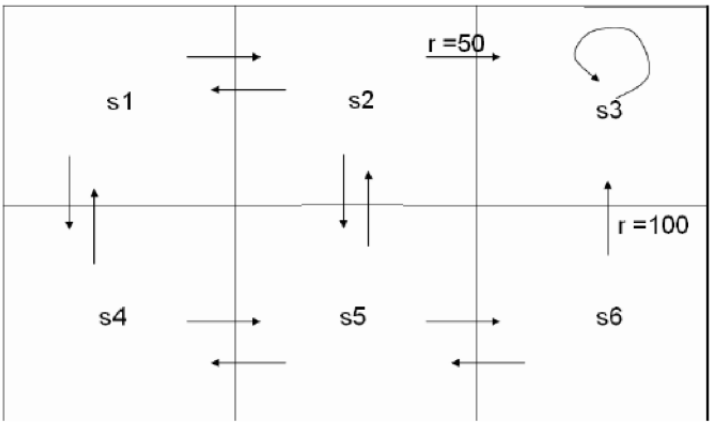
Ζητείται η εφαρμογή της Principal Component Analysis (PCA) πάνω στο ευρέως διαδεδομένο σύνολο δεδομένων κρίνων (Iris dataset) του Fisher, προκειμένου να μετασχηματιστούν τα δεδομένα σε ένα χώρο χαμηλότερων διαστάσεων. Τα δεδομένα αποτελούνται από 3 κλάσεις (για τους 3 διαφορετικούς τύπους κρίνου), καθεμιά 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)

(α) Κατεβάστε το σύνολο δεδομένων που έχει ανεβεί στην Helios ιστοσελίδα του μαθήματος (αρχείο PCA_iris_data).
(β) Προεπεξεργαστείτε τα δεδομένα αφαιρώντας τη μέση τιμή και διαιρώντας με την τυπική απόκλιση του κάθε χαρακτηριστικού ξεχωριστά. Τα προκύπτοντα δεδομένα θα πρέπει να έχουν μέση τιμή 0 και διασπορά 1.
(γ) Υπολογίστε τον δειγματικό πίνακα συνδιασπορών (sample covariance matrix) $\mathrm{C} = (1/N)\sum_{n=1}^{N} \mathrm{x}_i \mathrm{x}_i^T$, όπου x_n είναι το n-στό δείγμα και N το πλήθος των δειγμάτων.
(δ) Παραγοντοποιήστε τον πίνακα συνδιασπορών κάνοντας χρήση του Singular Value Decomposition (SVD) και βρείτε τις αντίστοιχες ιδιοτιμές και ιδιοδιανύσματα. Προσέξτε εάν η συγκεκριμένη υλοποίηση του SVD δίνει τα αποτελέσματα με φθίνουσα ή αύξουσα σειρά ιδιοτιμών. Ο μετασχηματισμός SVD ενός πίνακα C είναι μια παραγοντοποίηση της μορφής $\mathrm{C} = \mathrm{U}\Sigma\mathrm{V}^T$. Για έναν συμμετρικό θετικά-ορισμένο πίνακα, C, οι πίνακες $\mathrm{U} = \mathrm{V}$ περιέχουν τα ιδιοδιανύσματα και Σ είναι ένας διαγώνιος πίνακας με τις αντίστοιχες ιδιοτιμές.
(ε) Προβάλετε τα δεδομένα πάνω στις δύο πρώτες κύριες συνιστώσες και σχεδιάστε τα αποτελέσματα που προκύπτουν.
(στ) Ποιος είναι ο ελάχιστος απαιτούμενος αριθμός από κύριες συνιστώσες ώστε να "ερμηνεύεται" το 95% της διασποράς των τιμών;

Άσκηση 2.6 (Ενισχυτική Μάθηση)

Θεωρήστε το πρόβλημα μετακίνησης με ντετερμινιστικό τρόπο ενός ρομπότ στον διδιάστατο κόσμο της εικόνας. Σε όσες μεταβάσεις δεν αναγράφεται ανταμοιβή (r) θεωρήστε μηδενική τιμή ανταμοιβής. Επίσης, θεωρήστε συντελεστή έκπτωσης $\gamma = 0.8$.



(α) Για κάθε κατάσταση s υπολογίστε την τιμή $v^*(s)$.
(β) Σημειώστε στην εικόνα τα βέλη μεταβάσεων καταστάσεων-ενεργειών που αντιστοιχούν σε βέλτιστη πολιτική. Εάν υπάρχει ισοπαλία, να το αναφέρετε.
(γ) Πόσες πλήρεις επαναλήψεις ως προς την αξία είναι επαρκείς για να είναι εγγυημένη η εύρεση της βέλτιστης πολιτικής; Θεωρήστε ότι οι αρχικές τιμές αξίας είναι μηδενικές και ότι οι καταστάσεις λαμβάνονται με τυχαία σειρά σε κάθε επανάληψη.
(δ) Είναι δυνατό να μεταβληθούν οι ανταμοιβές, ώστε η v^* να αλλάξει, αλλά η βέλτιστη πολιτική π^* να παραμείνει αμετάβλητη; Εάν ναι, δώστε ένα παράδειγμα και περιγράψτε την αλλαγή που θα συμβεί στην v^*. Εάν όχι, εξηγήστε σύντομα γιατί είναι αδύνατο.