ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Επικοινωνιών, Ηλεκτρονικής & Συστημάτων Πληροφορικής
Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων - NETMODE
Ηρώων Πολυτεχνείου 9, Ζωγράφου, 157 80 Αθήνα, Τηλ: 210-772.2503, Fax: 210-772.1452
e-mail: maglaris@netmode.ntua.gr, URL: http://www.netmode.ntua.gr

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

(ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση)

ΕΝΔΕΙΚΤΙΚΕΣ ΑΣΚΗΣΕΙΣ 2, ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2022

Ο Αλγόριθμος K-Means

Να ομαδοποιήσετε τα παρακάτω ζεύγη σε δύο ομάδες με τη χρήση του Αλγορίθμου K-Means

Ζεύγη	Χαρακτηριστικό 1	Χαρακτηριστικό 2
Α	1.0	1.0
Β	1.5	2.0
Γ	3.0	4.0
Δ	5.0	7.0
Ε	3.5	5.0
ΣΤ	4.5	5.0
Ζ	3.5	4.5

Για τον ορισμό των δύο ομάδων επιλέγουμε αυθαίρετα τα σημεία Α (Ομάδα 1) και το σημείο Δ (Ομάδα 2). Σε αυτή την περίπτωση θα έχουμε τα εξής:

	Αρχικά σημεία	Mean Vector (Centroid)
Ομάδα 1	Α	(1.0,1.0)
Ομάδα 2	Δ	(5.0,7.0)

Για να ομαδοποιήσουμε και τα υπόλοιπα σημεία, εξετάζουμε για κάθε ένα από αυτά την απόσταση από το τα δύο centroids των ομάδων με βάση την ευκλείδεια απόσταση

$$\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad (1)$$

Κάθε φορά που ομαδοποιείται ένα νέο σημείο σε μία ομάδα, θα πρέπει να επανυπολογίσουμε το centroid της συγκεκριμένης ομάδας. Με βάση τα παραπάνω θα έχουμε 6 επαναλήψεις μέχρι να φράσουμε στην τελική μορφή των δύο ομάδων:

	Ομάδα 1	Ομάδα 2

Βήμα	Σημεία	Mean Vector (centroid)	Σημεία	Mean Vector (centroid)
1ο	Α	(1.0,1.0)	Δ	(5.0,7.0)
2ο	Α,Β	(1.2,1.5)	Δ	(5.0,7.0)
3ο	Α,Β,Γ	(1.8,2.3)	Δ	(5.0,7.0)
4ο	Α,Β,Γ	(1.8,2.3)	Δ,Ε	(4.2,6.0)
5ο	Α,Β,Γ	(1.8,2.3)	Δ,Ε,ΣΤ	(4.3,5.7)
6ο	Α,Β,Γ	(1.8,2.3)	Δ,Ε,ΣΤ,Ζ	(4.1,5.4)

Άρα οι ομάδες που θα δημιουργηθούν είναι οι εξής:

	Σημεία	Mean Vector (Centroid)
Ομάδα 1	Α,Β,Γ	(1.8,2.3)
Ομάδα 2	Δ,Ε,ΣΤ,Ζ	(4.1,5.4)

Πριν όμως καταλήξουμε ότι οι παραπάνω ομάδες είναι οι τελικές, θα πρέπει να εξετάσουμε την απόσταση κάθε σημείου από τα δύο centroids με βάση την ευκλείδεια απόσταση. Σε αυτή την περίπτωση θα έχουμε τα εξής:

Ζεύγη	Απόσταση από το Centroid 1	Απόσταση από το Centroid 2
Α	1.5	5.4
Β	0.4	4.3
Γ	2.1	1.8
Δ	5.7	1.8
Ε	3.2	0.7
ΣΤ	3.8	0.6
Ζ	2.8	1.1

Με βάση τα παραπάνω αποτελέσματα, προκύπτει ότι το σημείο Γ απέχει λιγότερο από το κέντρο της Ομάδας 2. Συνεπώς θα πρέπει να επανυπολογιστούν τα mean vectors για κάθε ομάδα όπως και πριν, με βάση αυτή την αλλαγή. Σε αυτή την περίπτωση θα έχουμε τα εξής:

	Σημεία	Mean Vector (Centroid)
Ομάδα 1	Α,Β	(1.3,1.5)
Ομάδα 2	Γ,Δ,Ε,ΣΤ,Ζ	(3.9,5.1)

Στα πλαίσια του συγκεκριμένου παραδείγματος ακόμα και μετά τον έλεγχο προκύπτουν ότι η τελική ομαδοποίηση είναι η παραπάνω. Σε κάθε περίπτωση όμως θα πρέπει να πραγματοποιείται ο έλεγχος, ώστε να επιβεβαιωθεί το τελικό αποτέλεσμα.

Principal Component Analysis (PCA)

Να εφαρμόσετε τη μέθοδο Principal Component Analysis (PCA) για να μειώσετε τη διάσταση του παρακάτω dataset:

$$D = \begin{pmatrix} 14.23 & 1.71 & 2.43 \\ 13.2 & 1.78 & 2.14 \\ 13.16 & 2.36 & 2.67 \end{pmatrix}$$

όπου η πρώτη στήλη αντιστοιχεί στο feature X, η δεύτερη στήλη αντιστοιχεί στο feature Y και η Τρίτη στο feature Z.

Αρχικά, βρίσκουμε τους μέσους όρους για κάθε feature:

$$\mu_X = \frac{14.23 + 13.2 + 13.16}{3} = 13.53$$

$$\mu_Y = 1.95$$

$$\mu_Z = 2.41$$

Στη συνέχεια, κανονικοποιούμε τα δεδομένα αφαιρώντας από τις τιμές τους κάθε feature τον αντίστοιχο μέσο όρο.

$$D_N = \begin{pmatrix} 0.7 & -0.24 & 0.02 \\ -0.33 & -0.17 & -0.27 \\ -0.37 & 0.41 & 0.26 \end{pmatrix}$$

Έπειτα, υπολογίζουμε τον πίνακα συδιακύμανσης:

$$Cov = \begin{pmatrix} 0.3679 & -0.1318 & 0.0035 \\ -0.1318 & 0.1273 & 0.0739 \\ 0.0035 & 0.0739 & 0.0705 \end{pmatrix}$$

Ενδεικτικά:

$$Cov(X, X) = \frac{1}{3-1}\big(0.7 \cdot 0.7 + (-0.33) \cdot (-0.33) + (-0.37) \cdot (-0.37)\big) = 0.3679$$

$$Cov(X, Y) = \frac{1}{3-1}\big(0.7 \cdot (-0.24) + (-0.33) \cdot (-0.17) + (-0.37) \cdot (0.41)\big) = -0.1318$$

Υπολογίζουμε τις ιδιοτιμές του πίνακα συδιακύμανσης:

$$\det(Cov - \lambda I) = \begin{vmatrix} 0.3679 - \lambda & -0.1318 & 0.0035 \\ -0.1318 & 0.1273 - \lambda & 0.0739 \\ 0.0035 & 0.0739 & 0.0705 - \lambda \end{vmatrix} = 0$$

$$\lambda_1 = 0.4281, \lambda_2 = 0.1375, \lambda_3 = 5.1 \cdot 10^{-18}$$

Παρατηρούμε ότι η ιδιοτιμή λ_3 είναι πολύ μικρότερη από τις άλλες δύο ιδιοτιμές. Έτσι, θα θεωρήσουμε ως Principal Components τα δύο πρώτα features. Υπολογίζουμε τα ιδιοδιανύσματα που τους αντιστοιχούν:

$$\Xi = \begin{pmatrix} -0.9061 & 0.3454 & -0.2442 \\ 0.4159 & 0.6221 & -0.6633 \end{pmatrix}$$

Τέλος, μετασχηματίζουμε το dataset ώστε να υπάρχει η συνεισφορά μόνο από τα δύο principal components:

$$D' = \Xi \cdot D_N = \begin{pmatrix} -0.6579 & 0.0586 & -0.1749 \\ 0.3313 & -0.4775 & -0.3321 \end{pmatrix}$$

Συντάχθηκε από τους υπεύθυνους εργαστηριακής υποστήριξης του μαθήματος Νίκο Κωστόπουλο και Δημήτρη Πανταζάτο, Υποψήφιους Διδάκτορες Ε.Μ.Π.