ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΑΝΑΛΥΣΗΣ ΗΛΕΚΤΡΙΚΩΝ ΚΥΚΛΩΜΑΤΩΝ
ΜΑΘΗΜΑ: ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ – 9ο ΕΞΑΜΗΝΟ

Εργασία 3

(Προθεσμία: Κυριακή 02 Δεκεμβρίου 2023, 23:59)

1. Κατεβάστε το seeds Dataset από τη διεύθυνση https://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt . Το αρχείο περιέχει μορφολογικές μετρήσεις 210 σπόρων από τρεις ποικιλίες σιτηρών (ω1: Kama, ω2: Rosa, ω3: Canadian). Οι πρώτες 7 στήλες περιέχουν τις μετρηθείσες τιμές των μορφολογικών χαρακτηριστικών (λεπτομέρειες εδώ) και η τελευταία στήλη περιλαμβάνει την ετικέτα της ποικιλίας στην οποία ανήκει κάθε σπόρος.

Α. Απεικονίστε τον πίνακα αποστάσεων των δεδομένων για Ευκλείδεια και cosine μετρικές. Ποιες κλάσεις πιστεύετε ότι είναι ευκολότερο να διαχωριστούν μεταξύ τους; Γιατί;

Β. Να υπολογίσετε το Silhouette Coefficient για την ομαδοποίηση των 7-διάστατων δεδομένων σε k=2,3,..10 κλάσεις με τη μέθοδο k-means και Ευκλείδεια ή squared Euclidean μετρική. Απεικονίστε το διάγραμμα του Silhouette και σχολιάστε ποιος είναι ο βέλτιστος αριθμός κλάσεων σύμφωνα με το κριτήριο αυτό.

Γ. Να κανονικοποιηθούν τα δεδομένα ώστε κάθε χαρακτηριστικό να έχει μηδενική τιμή και μοναδιαία variance. Υπολογίσετε εκ νέου το Silhouette Coefficient στα κανονικοποιημένα δεδομένα για cosine μετρική. Απεικονίστε το νέο διάγραμμα του Silhouette. Τι παρατηρείτε?

Δ. Ομαδοποιείστε τα δεδομένα σε 3 κλάσεις με τη μέθοδο k-means και squared Euclidean μετρική. Υπολογίστε το Rand Index για τη σύγκριση της παραγόμενης ομαδοποίησης με τις ετικέτες του dataset. Επαναλάβετε 5 φορές (με τυχαία αρχικοποίηση κέντρων), και υπολογίστε την μέση τιμή και το variance του Rand Index.

Ε. Επαναλάβατε το (Δ) για cosine μετρική[1]. Ποια μετρική επιτυγχάνει ομαδοποίηση πιο πιστή στις πραγματικές ομάδες των δεδομένων;

Bonus: Έστω ότι σχεδιάζετε ένα σύστημα που χρειάζεται να εκτελεί ταξινόμηση με τη μέθοδο *Nearest Neighbor* χρησιμοποιώντας ένα πολύ μεγάλο σύνολο δεδομένων αναφοράς (εκπαίδευσης). Μπορείτε να σκεφτείτε έναν τρόπο να μειωθεί το υπολογιστικό

[1] Σε περίπτωση που η βιβλιοθήκη που χρησιμοποιείτε δεν υποστηρίζει k-means με cosine μετρική, μπορείτε να χρησιμοποιήσετε την προσέγγιση της κανονικοποίησης των διανυσμάτων σε μοναδιαίο μήκος ακολουθούμενο από k-means με Ευκλείδεια μετρική.

κόστος κάθε νέας ταξινόμησης, αξιοποιώντας τεχνικές ομαδοποίησης? Περιγράψτε το σκεπτικό σας και τα βήματα του αλγορίθμου.

(3 μονάδες)

2. Επιλέξτε έναν αλγόριθμο ιεραρχικής ομαδοποίησης της αρεσκείας σας (agglomerative ή divisive) και εφαρμόστε τον στο ίδιο σύνολο δεδομένων με τα προηγούμενα. Να απαντηθούν τα παρακάτω: (3 μονάδες)

Α. Περιγράψτε συνοπτικά την τεχνική που επιλέξατε και τις παραμέτρους που χρησιμοποιήσατε (π.χ. μετρική, linkage method κλπ.). σε μία παράγραφο.

Β. Κατασκευάστε το δενδρόγραμμα που προέκυψε από την ομαδοποίηση, και σχολιάστε πως σχετίζεται με τις ποικιλίες των σιτηρών (κλάσεις) των αντίστοιχων δεδομένων.

Γ. Χρησιμοποιήστε την ιεραρχική ομαδοποίηση που δημιουργήσατε ώστε να διαχωρίσετε τα δεδομένα σε 3 ομάδες. Χρησιμοποιήστε κάποιο κριτήριο εξωτερικής επικύρωσης (π.χ. rand index, adjusted rand index, mutual information κλπ) και συγκρίνετε την ομαδοποίηση αυτή με την αντίστοιχη που επιτυγχάνει ο k-means σε σχέση με τις ετικέτες των δεδομένων. Ποια μετρική επιτυγχάνει ομαδοποίηση πιο πιστή στις πραγματικές ομάδες των δεδομένων;

Δ. Σχολιάστε τα πλεονεκτήματα των ιεραρχικών τεχνικών ομαδοποίησης.

(3 μονάδες)

3. Χρησιμοποιώντας το ίδιο dataset, να υλοποιηθούν και να απαντηθούν τα παρακάτω:

Α. Να εφαρμοστεί η μέθοδος PCA στα δεδομένα. Ποιες είναι οι ελάχιστες κύριες συνιστώσες που πρέπει να κρατήσετε ώστε να εξηγείται τουλάχιστον το 90% της variance του αρχικού dataset στη νέα απεικόνιση, και πόσες για το 99% αυτής;

Β. Να υπολογισθεί το σφάλμα ανακατασκευής των δεδομένων χρησιμοποιώντας από 1 έως 7 κύριες συνιστώσες, και να αποτυπωθεί σε κατάλληλο διάγραμμα.

Γ. Να εφαρμοστεί η μέθοδος LDA για την απεικόνιση του dataset σε 2 διαστάσεις. Να συγκρίνετε την απεικόνιση αυτή με την αντίστοιχη που παράγεται από τη μέθοδο PCA. Ποια τα κυριότερα ποιοτικά χαρακτηριστικά των απεικονίσεων που παράγουν οι δύο μέθοδοι και σε τι οφείλονται? Εξηγήστε.

Δ. Με βάση τον πίνακα προβολής που παράγεται από την LDA στο προηγούμενο ερώτημα, ποια είναι τα δύο χαρακτηριστικά (features) που συνεισφέρουν περισσότερο στη διάκριση μεταξύ των κλάσεων και ποια τα δύο που συνεισφέρουν λιγότερο απ' όλα? Δημιουργείστε δυο δυσδιάστατες απεικονίσεις των δεομένων χρησιμοποιώντας το καθένα από τα δύο ζεύγη χαρακτηριστικών που καταδείξατε. Σχολιάστε.

(2 μονάδες)

4. Κατεβάστε από εδώ το dataset «Air Distances Between Cities in Statute Miles» που περιέχει την απόσταση μεταξύ πόλεων είτε του κόσμου είτε των Ηνωμένων πολιτειών μετρούμενων σε μίλια κατά μήκος μέγιστων κύκλων. Να απαντηθούν τα εξής:

Α. Χρησιμοποιείστε τη μέθοδο classical MDS για να δημιουργήσετε μία διανυσματική αναπαράσταση των πόλεων του κόσμου (Distance_Matrix_world) στις δύο και στις

τρεις διαστάσεις. Απεικονίστε τις αναπαραστάσεις αυτές σε κατάλληλο διάγραμμα και σχολιάστε το αποτέλεσμα.

Β. Δημιουργείστε για τις πόλεις μία διανυσματική αναπαράσταση με τις μέγιστες διαστάσεις *d* που μπορεί να σας επιστρέψει ο αλγόριθμος MDS. Εάν $\mathbf{Y} \in \mathbb{R}^{N \times d}$ ο πίνακας με τις αναπαραστάσεις για τις *N* πόλεις, να δημιουργήστε το διάγραμμα των ιδιοτιμών του πίνακα $\mathbf{Y} \cdot \mathbf{Y}^T$ σε φθίνουσα σειρά. Το διάγραμμα αυτό χρησιμοποιείται ως ένδειξη της βέλτιστης διάστασης αναπαράστασης, διατηρώντας τόσες διαστάσεις όσες οι σημαντικές ιδιοτιμές. Με βάση αυτό, πόσες διαστάσεις εκτιμάτε ότι είναι οι βέλτιστες για τα δεδομένα του αρχείου Distance_Matrix_world;

Bonus: Για ποιο λόγο υπάρχουν μη μηδενικές ιδιοτιμές για περισσότερες των 3 διαστάσεων στο παραπάνω πρόβλημα; Συγκρίνετε το αντίστοιχο διάγραμμα για τα δεδομένα του αρχείου Distance_Matrix_US. Που οφείλεται αυτή η διαφορά και πως σχετίζεται με τη φύση του προβλήματος και των δεδομένων;

(2 μονάδες)

Οδηγίες: *Το Silhouette Coefficient στις νεότερες εκδόσεις matlab μπορεί να υπολογιστεί μέσω της συνάρτησης* `evalclusters`. *Υποβάλετε τις απαντήσεις σας στην πλατφόρμα του e-class στην ενότητα εργασίες, πριν τη λήξη της προθεσμίας υποβολής. Η υποβολή σας πρέπει να αποτελείται από* ***ένα μόνο συμπιεσμένο αρχείο*** *(.rar, .zip κλπ.) το οποίο θα περιέχει όλα τα απαραίτητα αρχεία για την υποβολή σας. Καταγράφετε τις απαντήσεις σας σε μία αναφορά που θα αποστείλετε σε μορφή pdf ή docx ή εναλλακτικά ενσωματωμένες σε Python notebooks ή Matlab live scripts. Οι αποδεκτές γλώσσες προγραμματισμού είναι Matlab ή Python, όπου και θα πρέπει να συμπεριλάβετε στο συμπιεσμένο αρχείο και τα αρχεία του πηγαίου κώδικα (αρχεία .m, .py κλπ) με τα απαραίτητα σχόλια για την τεκμηρίωση εντός του κώδικα. υποβάλλοντας* ***το πολύ ένα αρχείο κώδικα για κάθε άσκηση****. Σε περίπτωση που χρησιμοποιήσετε Matlab, μπορείτε να συμπεριλάβετε τον κώδικα για όλες τις ασκήσεις σε ένα αρχείο, χωρίζοντας τις απαντήσεις σε διαφορετικά cells. Επίσης, για τις ασκήσεις που περιλαμβάνουν κώδικα, αντί απάντησης στην αναφορά μπορείτε να υποβάλετε Matlab live scripts ή Python notebooks κατάλληλα δομημένα ώστε να περιλαμβάνουν και τις απαντήσεις/σχολιασμούς που ζητούνται.*

Ξάνθη, 19/11/2023