5/5/2023

Προπτυχιακό μάθημα: **Μηχανική Μάθηση**

Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2022-23

2η Σειρά Ασκήσεων

Ημερομηνία παράδοσης : έως 8/6/2023

Θέμα: Μέθοδοι μείωσης Ομαδοποίησης δεδομένων και μείωσης διάστασης (*Clustering and Dimension Reduction methods*)

(*δίνεται επιπλέον bonus 10%*)

Στην εργασία αυτή θα χρησιμοποιήσετε τα πρώτο πειραματικό σύνολο της πρώτης σειράς ασκήσεων:

- **Mobile Price Classification** – 2,000 δεδομένα 20 διάστασης σε 4 κατηγορίες (0, 1, 2, 3)

https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

Για το σύνολο δεδομένων υπάρχουν δύο αρχεία για training και test.

Πρώτη φάση

Στόχος της εργασίας είναι να μελετήσετε αρχικά τις επιδόσεις δύο γνωστών αλγορίθμων ομαδοποίησης (clustering):

- **αλγόριθμος *k-means*** χρησιμοποιώντας είτε Ευκλείδια απόσταση,
- ***agglomerative hierarchical clustering*** (συνθετική ιεραρχική ομαδοποίηση) χρησιμοποιώντας την στρατηγική ward για την εύρεση των ομάδων με την μικρότερη απόσταση και την συνένωσή τους (*merge*).

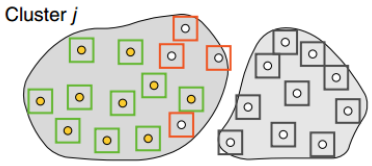
Τα δεδομένα αυτά προέρχονται από δύο κατηγορίες. Παρόλ' αυτά στο πρόβλημα της ομαδοποίησης δεν είναι γνωστή εκ των προτέρων η κατηγορία. Θα υποθέσετε διαφορετικό αριθμό ομάδων (K) και συγκεκριμένα τις τιμές: K={2, 4, 6, 8, 10}.

Η αξιολόγηση της επίδοσης των μεθόδων (για κάθε τιμή του K και για κάθε μέθοδο) θα γίνει στο εκάστοτε σύνολο ελέγχου (*testing set*) με βάση τα γνωστά μέτρα αξιολόγησης ***accuracy*** (ακρίβεια ή ποσοστό επιτυχίας) και ***F1-score***, τροποποιημένα ως εξής για το πρόβλημα της ομαδοποίησης:

- ***Purity***: Η κατηγορία κάθε ομάδας (c_j) καθορίζεται, μετά το τέλος της ομαδοποίησης, από την πλειοψηφούσα πραγματική κατηγορία (ω_k) μεταξύ των μελών της ομάδας. Τότε η ακρίβεια (*purity*) υπολογίζεται μετρώντας το μέσο των σωστά ταξινομημένων σημείων. Δηλ.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

- ***F-measure:***



		Truth	
		P	N
Hypothesis	P	TP (a)	FP (b)
	N	FN (c)	TN (d)

Precision: $\frac{a}{a+b}$ Recall: $\frac{a}{a+c}$ F-measure: $F_\alpha = \frac{1+\alpha}{\frac{1}{precision} + \frac{\alpha}{recall}}$ $\alpha = 1$ $\alpha \in (0;1)$ $\alpha > 1$

Για κάθε cluster *j*, αφού καθορίσετε την πλειοψηφούσα κατηγορία ως κατηγορία *cluster*, βρείτε τα TP (*true positive*), FP (*false positive*) και FN (*false negative*) και στη συνέχεια το *F-measure* $F_a^{(j)}$ χρησιμοποιώντας την τιμή $\alpha=1$. Συνολικά, η αξιολόγηση μιας μεθόδου *clustering* θα γίνεται από το άθροισμα των *F-measures* για κάθε *cluster*.

$$Total\ \ F-measure = \sum_{j=1}^{K} F_1^{(j)}$$

Για κάθε τιμή αριθμού ομάδων, K, κάντε 10 εκτελέσεις (μόνο για τον αλγόριθμο kmeans καθώς η ιεραρχική ομαδοποίηση δεν εξαρτάται από την αρχικοποίηση) και υπολογίστε τον μέσο όρο για κάθε μέτρο αξιολόγησης.

Δεύτερη φάση

Στη δεύτερη φάση θα μελετήσετε την αξία του μετασχηματισμού του αρχικού χώρου των δεδομένων μέσω της τεχνικής νευρωνικών δικτύων **Autoencoder**. Συγκεκριμένα, θα εκπαιδεύσετε αρχικά πάνω στα δεδομένα (του συνόλου εκπαίδευσης) έναν **Autoencoder** με την εξής αρχιτεκτονική νευρωνικού δικτύου:

20-100-M-100-20

υποθέτοντας τις εξής τιμές του M (διάσταση του μετασχηματισμένου χώρου): M={2, 10, 50}.

Παρατηρήσεις :

- για M=50 δεν γίνεται μείωση διάστασης, αλλά αύξηση διάστασης και λεπτομερέστερη περιγραφή των δεδομένων,
- για M=2 τα δεδομένα μετασχηματίζονται σε έναν δυσδιάστατο χώρο με το σημαντικό πλεονέκτημα της δυνατότητας οπτικοποίησής τους κάνοντας ένα plot και (πιθανώς) καλύτερης αντίληψής τους. Να κάνετε ένα τέτοιο plot και να το δείξετε.

Στη συνέχεια και μετά την μάθηση του Autoencoder, να χρησιμοποιήσετε μόνο τον encoder (1ο μισό του δικτύο) για να προβάλλετε όλα τα δεδομένα (και των δύο αρχείων) στον νέο χώρο Μ-διάστασης. Δουλέψτε πάνω σε αυτόν τον χώρο και επαναλάβετε την διαδικασία της πρώτης φάσης.
Παρόμοια, για κάθε τιμή της διάστασης του μετασχηματισμένου χώρου, Μ, και του αριθμού των ομάδων, Κ, κάντε 10 εκτελέσεις (και στους δύο αλγορίθμους ομαδοποίησης στην περίπτωση αυτή λόγω εξάρτησης από την αρχικοποίηση του autoencoder) και υπολογίστε τον μέσο όρο για κάθε μέτρο αξιολόγησης.

Δώστε ένα **σύντομο *report* (*pdf*** μορφή αρχείου***)*** το οποίο θα στείλετε ηλεκτρονικά (μέσω turnin):

turnin Homework2@mye002 filename1 filename2 ...

περιγράφοντας εν συντομία
α) την διαδικασία κατασκευής των μεθόδων,
β) τα αποτελέσματα των δοκιμών ανά μέθοδο και ανά περίπτωση και για τις δύο φάσεις συμπληρώνοντας τους παρακάτω δύο πίνακες,
γ) την βέλτιστη μέθοδο που θα προκύψει από την σύγκρισή τους και τα συμπεράσματά σας, και
δ) ενσωματώστε στο κείμενο τον κώδικα που κατασκευάσατε ως παράρτημα.

Method	***Number of Clusters (K)***	***Purity***	***F-measure***
K-means (*Euclidean distance*)	*2*		
	4		
	6		
	8		
	10		
Agglomerative Hierarchical Clustering	*2*		
	4		
	6		
	8		
	10		

Πίνακας 1. Πειραματικά αποτελέσματα της **πρώτης φάσης** (τα αποτελέσματα προέκυψαν από τον μέσο όρο 10 ανεξάρτητων εκτελέσεων ανά περίπτωση)

Method	*Dimension (M)*	*Number of Clusters (K)*	*Purity*	*F-measure*
K-means (*Euclidean distance*)	***2***	*2*		
		4		
		6		
		8		
		10		
	10	*2*		
		4		
		6		
		8		
		10		
	50	*2*		
		4		
		6		
		8		
		10		
Agglomerative Hierarchical Clustering	***2***	*2*		
		4		
		6		
		8		
		10		
	10	*2*		
		4		
		6		
		8		
		10		
	50	*2*		
		4		
		6		
		8		
		10		

Πίνακας 2. Πειραματικά αποτελέσματα της **δεύτερης φάσης** (τα αποτελέσματα προέκυψαν από τον μέσο όρο 10 ανεξάρτητων εκτελέσεων ανά περίπτωση)

Καλή επιτυχία!