13/5/2024

Προπτυχιακό μάθημα: **Μηχανική Μάθηση**
Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2023-24

2η Σειρά Ασκήσεων
Ημερομηνία παράδοσης : έως και 31/5/2024

Θέμα: Μείωση διάστασης και Ομαδοποίηση δεδομένων

Χρησιμοποιήστε το γνωστό σύνολο δεδομένων ***fashion MNIST*** επιλέγοντας ***n*** εικόνες ανά κατηγορία (π.χ. *n=1000* ή *2000* μετά από τυχαία επιλογή του υπάρχοντος συνόλου των δεδομένων - συνολικά υπάρχουν *50,000* δεδομένα για εκπαίδευση) σε διανυσματική (κανονικοποιημένη) μορφή με διάσταση *d=784*. Στη συνέχεια να κάνετε τα εξής βήματα:

1. Μείωση διάστασης:
α) Να μειώσετε τη διάσταση των δεδομένων χρησιμοποιώντας την μέθοδο *PCA* διατηρώντας το 90% της διακύμανσης των δεδομένων εκπαίδευσης. Έστω ***M*** η διάσταση του χώρου προβολής των δεδομένων.

β) Να μειώσετε την διάσταση των δεδομένων εκπαιδεύοντας έναν ***Autoencoder*** με **αρχιτεκτονική *d – d/4 - M – d/4 – d***, θεωρώντας την τιμή του **M** που βρήκατε στο ερώτημα (α).

γ) Χρησιμοποιήστε την καλύτερη μέθοδο ταξινόμησης που βρήκατε στην 1η εργασία (στο ερώτημα [Α]) πάνω στον χώρο μειωμένης διάστασης των δεδομένων και παρατηρήστε αν βελτιώνεται η επίδοσή της.

2. Ομαδοποίηση:
Σύμφωνα με την παραπάνω διαδικασία μείωσης διάστασης, έχετε δύο σύνολα εκπαίδευσης που αποτελούνται από διανύσματα διάστασης Μ, ένα για την μέθοδο PCA και ένα για την μέθοδο Autoencoder. Τα παρακάτω βήματα θα τα εφαρμόσετε ανεξάρτητα και στα δύο σύνολα (ξεχωριστά).

α) Να ομαδοποιήσετε τα παραδείγματα του συνόλου σε ***K*** ομάδες σύμφωνα με τον **αλγόριθμο *k-means*** με Ευκλείδια απόσταση, χρησιμοποιώντας ως αριθμό των ομάδων την τιμή ***K=10* έως 20**. Στη συνέχεια υπολογίστε το ***silhouette score*** για κάθε μια τιμή και κάντε ένα κατάλληλο διάγραμμα ώστε να βρείτε την βέλτιστη τιμή ***K****.

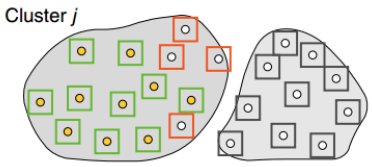
β) Να βρείτε το **κέντρο** κάθε *cluster* και να κατασκευάσετε και να τυπώσετε την αντίστοιχη εικόνα (πραγματικής διάστασης) σε κάθε περίπτωση.

γ) Να αξιολογήσετε την ποιότητα της ομαδοποίησης υποθέτοντας το πλήθος των ομάδων ίσο με ***K**** χρησιμοποιήσετε τα δύο παρακάτω μέτρα:

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

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

Όπως και στην πρώτη εργασία, θα πρέπει να παραδώσετε τόσο το report (αρχείο pdf), όσο και τον κώδικα (αρχείο .ipynb) σε περιβάλλον ***Jupyter Notebook***, δίνοντας παράλληλα ικανοποιητικά σχόλια (φτιάξτε ένα αρχείο zip που θα περιλαμβάνει τα 2 αυτά αρχεία). Επίσης μην ξεχάσετε να γράψετε το όνομα των μελών της ομάδας (σε σχόλιο) στην αρχή του κειμένου. Το αρχείο θα το παραδώσετε με την γνωστή διαδικασία turnin. Συγκεκριμένα, για την 1η σειρά ασκήσεων θα γίνει στέλνοντας το παρακάτω:

turnin Homework2@mye002 *your-filename*