Προπτυχιακό μάθημα: **Μηχανική Μάθηση,**
Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων,
(ακαδ. έτος 2020-21)

2η Σειρά Ασκήσεων

(Ημερομηνία παράδοσης : έως Τρίτη 1/6/2021)

3/5/2021

Θέμα: Ομαδοποίηση δεδομένων

Θα χρησιμοποιηθεί το ίδιο πειραματικό σύνολο δεδομένων με αυτό της πρώτης άσκησης, δηλ. το σύνολο fashion MNIST το οποίο περιέχει ασπρόμαυρες εικόνες (διάστασης 28 x 28) από 10 κατηγορίες ρούχων. Στόχος της παρούσας άσκησης είναι να πετύχουμε την (μη επιβλεπούμενη – unsupervised) ανάλυση της δομής τους μέσω της διαδικασίας της ομαδοποίησης, δηλ. του βέλτιστου διαχωρισμού του συνόλου δεδομένων σε $K=10$ ομάδες.

Θα χρησιμοποιήσετε 2 εναλλακτικούς τύπους δεδομένων για την αναπαράσταση κάθε εικόνας:

- **R1**: διάνυσμα μήκους 28x28=784 όπου κάθε συνιστώσα θα αντιπροσωπεύει την τιμή φωτεινότητας κάθε pixel της εικόνας (προτείνεται η κανονικοποίηση των τιμών διαιρώντας με το 255 που είναι η μέγιστη τιμή φωτεινότητας και αντιστοιχεί στο βαθύ μαύρο),
- **R2**: ιστόγραμμα φωτεινότητας χρησιμοποιώντας *M bins* (*M = 16 ή 32 ή 64 ή 128*).

Για την ομαδοποίηση (και για τους δύο τύπους μορφής δεδομένων, R1 ή R2) θα χρησιμοποιήσετε τις παρακάτω μεθόδους ομαδοποίησης τις οποίες **θα πρέπει να υλοποιήσετε** (και όχι να χρησιμοποιήσετε έτοιμες συναρτήσεις):

- ***K-means clustering:***

- Η κλασσική μορφή του αλγορίθμου χρησιμοποιώντας
  - **Ευκλείδια απόσταση** (*L2*), είτε
  - **Manhattan distance** (*L1*), είτε
  - **Συνημιτονοειδή** απόσταση (*cosine distance*)

  μεταξύ των διανυσματικών δεδομένων.

- Η γενική περίπτωση του αλγορίθμου ***K-medoid*** μόνο για την R2 μορφή αναπαράστασης χρησιμοποιώντας την **symmetric** (συμμετρική) **Kullback-Libler (KL) distance**, που είναι μία συνάρτηση απόστασης μεταξύ κατανομών (θα βρείτε περισσότερες λεπτομέρειες στο συνοδευόμενο αρχείο της άσκησης).

- ***Hierarchical clustering:*** χρησιμοποιώντας τη μέθοδο διχοτόμησης (*divisive* or *linkage*) μορφή κατασκευής του δέντρου των δεδομένων για όλες τις παραπάνω μορφές απόστασης.

Παρατηρήσεις

A) Το αρχικό σύνολο δεδομένων περιέχει 60,000 (train) και 10,000 (test) εικόνες. Δουλέψτε μόνο με τα training image-data. Σε περίπτωση που αντιμετωπίσετε κάποιο δυσκολία στον χειρισμό μεγάλου όγκου δεδομένων μπορείτε εναλλακτικά να μειώσετε το σύνολο δεδομένων, προσέχοντας μόνο να επιλέξετε ισοκαταμερισμένο αριθμό δεδομένων ανά κατηγορία. Επίσης, είναι δυνατόν να υπολογίσετε εξ'αρχής τις αποστάσεις, να τις αποθηκεύσετε σε μια δομή και να τις χρησιμοποιείται επαναληπτικά μέσα στον κώδικα.

B) Για τη σύγκριση των μεθόδων ομαδοποίησης χρησιμοποιήσετε τα δύο παρακάτω μέτρα αξιολόγησης:

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

- ***F-measure:***

  Για κάθε cluster, 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 F1-score (βλ. άσκηση1). Στο τέλος, η αξιολόγηση της μεθόδου clustering θα προκύπτει από το άθροισμα των F-measures για κάθε cluster.

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

Δώστε ένα **σύντομο *report*** με τον τρόπο κατασκευής των μεθόδων, τα αποτελέσματα των πειραμάτων ανά μέθοδο, όπως επίσης την βέλτιστη μέθοδο που θα προκύψει από την σύγκριση. Στο κείμενο θα πρέπει να υπάρχει και ο κώδικας που κατασκευάσατε ως παράρτημα.