Προπτυχιακό μάθημα: **Μηχανική Μάθηση**

Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2021-22

2η Σειρά Ασκήσεων

12/5/2022

Ημερομηνία παράδοσης : έως 22/6/2022

Θέμα: Μείωση διάστασης και Ομαδοποίηση εικόνων

Στην ιστοσελίδα: https://www.kaggle.com/c/11785-spring2021-hw2p2s1-face-classification υπάρχει ένα πειραματικό σύνολο δεδομένων που αποτελείται από πραγματικές εικόνες προσώπων, το οποίο χρησιμοποιήθηκε πρόσφατα ως αντικείμενο ενός διαγωνισμού στο περιβάλλον *Kaggle*. Συνολικά υπάρχουν έγχρωμες (RGB) εικόνες διάστασης 64x64 από 4,000 πρόσωπα και για κάθε ένα πρόσωπο υπάρχουν αρκετά αντίγραφα του ίδιου προσώπου (με διαφορετική οπτική γωνία, φωτεινότητα, κλπ.). Για τις ανάγκες της άσκησης χρησιμοποιήστε (τυχαία) 10 από τα υπάρχοντα πρόσωπα (δηλ. K=10 κατηγορίες) και για καθένα πρόσωπο (κατηγορία) χρησιμοποιήστε 50 εικόνες (τα περισσότερα από τα πρόσωπα έχουν περισσότερες από 50 εικόνες). Έτσι συνολικά, θα έχετε ένα σύνολο 10 x 50 = 500 εικόνων από K=10 διαφορετικά πρόσωπα.

Καθώς οι εικόνες είναι έγχρωμες υπάρχουν 3 κανάλια φωτεινότητας (**Red**, **Green**, **Blue**) και επομένως για κάθε *pixel* της εικόνας υπάρχουν 3 τιμές (3 στάθμες φωτεινότητας με τιμές μεταξύ 0-255). Έτσι κάθε εικόνα περιγράφεται με 3 διανύσματα (ένα για κάθε κανάλι) διάστασης 4,096 (64x64) το καθένα. Μία τεχνική μετατροπής μιας έγχρωμης εικόνας σε μονοχρωματική χωρίς απώλεια της πληροφορίας είναι να χρησιμοποιήσετε την παρακάτω γραμμική σχέση:

$$\mathbf{0.299 \cdot Red + 0.587 \cdot Green + 0.114 \cdot Blue}$$

Περισσότερες πληροφορίες μπορείτε να βρείτε κάνοντας μία απλή αναζήτηση στο διαδίκτυο ή στον επόμενο σύνδεσμο ή: https://www.kdnuggets.com/2019/12/convert-rgb-image-grayscale.html

Με τον τρόπο αυτό οι εικόνες περιγράφονται με ένα κανάλι και άρα ως ένα διάνυσμα 4,096 τιμών.

Στην συνέχεια, στο σύνολο δεδομένων που θα κατασκευάσετε (*N=500* διανυσμάτων διάστασης *d=4,096*) θα πρέπει να κάνετε τα εξής:

1. Μείωση διάστασης: (να μην χρησιμοποιήσετε την πληροφορία κατηγορίας)

α) Να μειώσετε τη διάσταση των δεδομένων χρησιμοποιώντας την μέθοδο *PCA*. Για την διάσταση του χώρου προβολής δοκιμάστε τις τιμές **M=100, 50, 25**.

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

2. Ομαδοποίηση Προσώπων: Να ομαδοποιήσετε τα παραδείγματα αυτού του συνόλου σε ***K=10*** ομάδες (θεωρώντας μία ομάδα ανά πρόσωπο) σύμφωνα με τις παρακάτω μεθόδους:

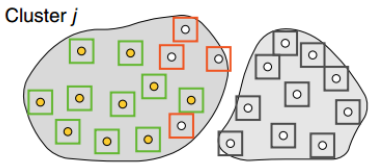
- **αλγόριθμος *k-means*** χρησιμοποιώντας είτε Ευκλείδια απόσταση, είτε συνημιτονοειδή απόσταση (*cosine distance*),
- ***agglomerative hierarchical clustering*** (συνθετική ιεραρχική ομαδοποίηση) χρησιμοποιώντας την στρατηγική ward για την εύρεση των ομάδων με την μικρότερη απόσταση και την συνένωσή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

Για την **αξιολόγηση της ομαδοποίησης** 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*.

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

Δώστε ένα **σύντομο *report*** με τον τρόπο κατασκευής των μεθόδων, τα αποτελέσματα των δοκιμών ανά μέθοδο, όπως επίσης την βέλτιστη μέθοδο που θα προκύψει από την σύγκριση. Στο *report* που θα παραδώσετε θα πρέπει να υπάρχει ο παρακάτω πίνακας συμπληρωμένος με τα αποτελέσματα που θα πάρετε:

Method	***dimension of data (M)***	***Purity***	***F-measure***
K-means (*Euclidean distance*)	*100*		
	50		
	25		
K-means (*Cosine distance*)	*100*		
	50		
	25		
Agglomerative Hierarchical Clustering	*100*		
	50		
	25		