1/5/2025

Προπτυχιακό μάθημα: **Μηχανική Μάθηση**
Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2024-25

2η Σειρά Ασκήσεων
Ημερομηνία παράδοσης: έως και 25/5/2024

Θέμα: Ομαδοποίηση δεδομένων με μείωση διάστασης

Ο σκοπός αυτής της εργασίας είναι η διερεύνηση του τρόπου με τον οποίο η **μείωση διαστάσεων** επηρεάζει την λύση της **ομαδοποίησης** σε δεδομένα υψηλής διάστασης. Θα δουλέψετε πάνω στο σύνολο δεδομένων εικόνας που δημιουργήσατε στην προηγούμενη άσκηση, μέσω του προεκπαιδευμένου μοντέλου CNN, αποτελούμενο από **D-διάστατα διανύσματα χαρακτηριστικών** και γνωστές ετικέτες κατηγοριών (τις οποίες θα χρησιμοποιήσετε μόνο για αξιολόγηση των μεθοδολογιών).

1. Μείωση Διαστάσεων

Εφαρμόστε τις παρακάτω μεθόδους μείωσης διαστασης και τον μετασχηματισμό των D-διάστατων δεδομένων σε M-διάστατα διανύσματα χαρακτηριστικών ($\boldsymbol{M \ll D}$):

- **PCA** (Ανάλυση Πρωτευουσών Συνιστωσών)
- **LLE** (Τοπική Γραμμική Ενσωμάτωση)
- **Autoencoders** με Βαθιά Νευρωνικά Δίκτυα (με 2 κρυμμένα επίπεδα των 256 των 128 νευρώνων)

Για κάθε μέθοδο:

- Πειραματιστείτε με διαφορετικές τιμές **M** ($M \in \{2, 3, 5, 10, 20\}$).
- Παρουσιάστε **οπτικοποίηση των δεδομένων σε 2D και 3D** (για $M = 2$ ή 3) δείχνοντας παράλληλα την διαμόρφωση των κατηγοριών χρησιμοποιώντας διαφορετικό χρώμα στα σημεία-δεδομένα ανάλογα με την πραγματική κατηγορία που ανήκουν, σχολιάζοντας τα αποτελέσματα.

2. Ομαδοποίηση στον Μειωμένο Χώρο

Εφαρμόστε ομαδοποίηση στον χώρο μειωμένων διαστάσεων M με τις εξής μεθόδους:

- **K-Means**, με:
  - **Ευκλείδεια απόσταση**
  - **Συνημίτονο απόστασης (Cosine Distance)**
- **Ιεραρχική Συνθετική Ομαδοποίηση (Agglomerative Clustering)**

3. Αξιολόγηση Ομαδοποίησης

Για κάθε συνδυασμό μεθόδου μείωσης διάστασης, διάστασης M και μεθόδου ομαδοποίησης:

- Εκτελέστε ομαδοποίηση με **K = 2 έως** $\boldsymbol{K_{max} = 10}$ αριθμό ομάδων
- Υπολογίστε τις εξής ποσότητες (δείκτες ποιότητας) πάνω σε σχετικό διάγραμμα:
    - **Καθαρότητα (Purity)**
    - **F-measure**
    - **Silhouette score**

Εκτιμήστε τον **βέλτιστο αριθμό συστάδων (**$\boldsymbol{k^*}$**)** με βάση τον δείκτη Silhouette.

Παραδώστε

- ένα **Notebook** με σχολιασμένο και κατανοητό κώδικα που να περιέχει διαγράμματα για προβολές 2D/3D των δεδομένων και δείκτες ποιότητας ως προς το K, καθώς επίσης πίνακες σύγκρισης μεθόδων και αποτελεσμάτων.
- μία **αναφορά** (report) με την διαδικασία, αποτελέσματα και σύγκριση των μεθοδολογιών.