# Μηχανική Μάθηση

## 2η Εργασία – Clustering problems

Το πρόβλημα που θα αντιμετωπίσετε αφορά στην αξιολόγηση διαφορετικών συνδυαστικών μοντέλων μείωσης διάστασης του χώρου των μεταβλητών (dimensionality reduction) και συσταδοποίησης (clustering) πάνω σε εικόνες.

Τα δεδομένα που θα χρησιμοποιήσετε αφορούν στο **fashion-mnist** dataset**.** Πληροφορίες στο ακόλουθο link: https://keras.io/api/datasets/fashion_mnist/.

Για την συγκεκριμένη άσκηση θα παραδώσετε ένα αρχείο σε python στο οποίο θα παρουσιάζετε τις διαφορές στα αποτελέσματα συσταδοποίησης όταν χρησιμοποιώντας ακατέργαστα (raw) δεδομένα και όταν χρησιμοποιούνται σύνθετα περιγραφικά χαρακτηριστικά (features), τα οποία προέκυψαν μέσω τεχνικών dimensionality reduction.

Συνολικά πρέπει να χρησιμοποιήσετε τρείς (3) διαφορετικές τεχνικές dimensionality reduction. Οι δύο (2) εκ των οποίων *πρέπει* να είναι:

1. Principal component analysis.
2. Stacked autoencoder.

Η τρίτη τεχνική επαφίεται στην επιλογή σας.

Επίσης θα πρέπει να χρησιμοποιήσετε τρείς (3) διαφορετικές τεχνικές clustering. Οι δύο (2) εκ των οποίων πρέπει να είναι:

1. Minibatch kmeans
2. DBSCAN

Η τρίτη τεχνική επαφίεται στην επιλογή σας.

Η αξιολόγηση της καταλληλότητας των cluster θα γίνει με χρήση τεσσάρων (4) μετρικών απόδοσης. Οι τρείς (3) εξ’ αυτών είναι οι:

1. Calinski–Harabasz index
2. Davies–Bouldin index
3. Silhouette score

Η 4η μετρική είναι της επιλογής σας.

Ο κώδικας που θα παραδώσετε, σε γλώσσα Python, πρέπει να υλοποιεί τα ακόλουθα:

1. Θα φορτώνει τα δεδομένα του fashion-mnist.
2. Θα διαχωρίζει τα δεδομένα σε τρία σύνολα: train, validation & test data.
3. Θα τρέχει μια τεχνική dimensionality reduction, πάνω στα train data.

Παρατηρήσεις: α) η αρχιτεκτονική για τον SAE θα καθοριστεί από εσάς. β) Τα NN-based models, κατά την διάρκεια του fit, θα χρειαστούν και τα validation data. γ) Προσοχή: στην SAE αρχιτεκτονικές πρέπει να απομονώσετε το κομμάτι του encoder για να κάνετε το dimensionality reduction.

4. Θα τυπώνει τυχαία εικόνες από το dataset (μία από κάθε κλάση) καθώς και τις ανακατασκευασμένες, εφόσον η τεχνική το επιτρέπει.
5. Θα τυπώνει μια γραφική παράσταση, όποια κρίνετε χρήσιμη, για να δείξετε ότι η τεχνική για το dimensionality reduction μάλλον θα δουλέψει
6. Θα χρησιμοποιεί την τεχνική πάνω στα test data και θα τα κωδικοποιεί.
7. Θα χρησιμοποιεί τρεις (3) διαφορετικές τεχνικές clustering για να δημιουργήσει τα αντίστοιχα clusters.
8. Θα υπολογίζει τις τέσσερεις (4) μετρικές απόδοσης.
9. Να καταχωρεί σε ένα dataframe (Pandas) σε μια νέα γραμμή τις ακόλουθες πληροφορίες:
    a. Dimensionality reduction technique name (str). Use "Raw" if no technique was used.
    b. Clustering algorithm (str).
    c. Training time for the dim. red. tech. in seconds (double)
    d. Execution time for the clustering tech. in seconds (double)
    e. Number of suggested clusters (int)
    f. Calinski–Harabasz index (double)
    g. Davies–Bouldin index (double)
    h. Silhouette score (double)
    i. The metric of your choice (double)
10. Θα παρουσιάζει ενδεικτικά αποτελέσματα ομαδοποίησης για τυχαίες εικόνες για τέσσερις (4) κατηγορίες (classes) της επιλογής σας.

Γενικές παρατηρήσεις:

1. Τα βήματα 3 έως και 9 θα εκτελεστούν τρείς (3) φορές, όσες δηλαδή και οι τεχνικές μείωσης διάστασης.
2. Οι τεχνικές clustering θα εφαρμοστούν στο test set. Είναι σημαντικά μικρότερο σε πλήθος παρατηρήσεων, άρα θα τρέξει ποιο γρήγορα.
3. Η παραπάνω διαδικασία θα επαναληφθεί δύο (2) φόρες χρησιμοποιώντας α) τις τιμές των pixel των εικόνων (κανονικοποιημένες στο [0,1]) και β) τις τιμές των εικόνων που παράγει η τεχνική του dimensionality reduction.

Χρησιμοποιώντας τα αποτελέσματα, τις γραφικές παραστάσεις και εικόνες του κώδικα, καθώς και γραφικές παραστάσεις που θα φτιάξετε στο excel, θα συντάξετε μια έκθεση στην οποία θα παρουσιάζετε τα συμπεράσματα σας, θα κάνετε συγκριτικές αξιολογήσεις και θα προτείνετε ποιος είναι ο καλύτερος δυνατός συνδυασμός τεχνικών για την συγκεκριμένη περίπτωση.

Υπήρξε περίπτωση στην οποία ένας συνδυασμός πέτυχε τα καλύτερα αποτελέσματα σε όλες τις μετρικές;

**Σημαντική παρατήρηση:** Αξιολογείστε με βάση την αναφορά που θα παραδώσετε, εργασία που *δεν* συνοδεύεται από γραπτή αναφορά βαθμολογείται με 0.

Καταληκτική Ημερομηνία Παράδοσης: **Δευτέρα 22 Ιανουαρίου 2024, 23:55**

Προσοχή! Δεν θα δοθεί παράταση. Κατανείμετε προσεκτικά τις ώρες εργασίας σας.