3η Εργασία (μάθηση χωρίς επίβλεψη – συσταδοποίηση)

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 θα χρησιμοποιήσετε τις ακόλουθες δύο (2) τεχνικές dimensionality reduction: 1.
Principal component analysis (PCA).
2. Stacked autoencoder (SAE).

Επίσης θα πρέπει να χρησιμοποιήσετε δύο (2) διαφορετικές τεχνικές clustering. Εξ' αυτών η πρώτη θα είναι υποχρεωτικά ο Minibatch kmeans. Η 2η τεχνική επαφίεται στην επιλογή σας.

Η αξιολόγηση της καταλληλότητας των cluster θα γίνει με χρήση των ακόλουθων τριών (3) μετρικών απόδοσης.

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

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 αρχιτεκτονική 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 εσείς κρίνετε χρήσιμη, για να δείξετε ότι η τεχνική για το dimensionality reduction μάλλον θα δουλέψει

6. Θα χρησιμοποιεί την τεχνική για το dimensionality reduction πάνω στα test data και θα τα κωδικοποιεί.
7. Θα χρησιμοποιεί δύο (2) διαφορετικές τεχνικές clustering για να δημιουργήσει τα αντίστοιχα clusters.
8. Θα υπολογίζει τις τρείς (3) μετρικές απόδοσης.
9. θα 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)
10. Επιλέξτε την τεχνική clustering που σας έδωσε τα καλύτερα αποτελέσματα. Επιλέξτε, τυχαία, δύο clusters από και τυπώστε 10 τυχαίες εικόνες ανά cluster.

Γενικές παρατηρήσεις:

1. Τα βήματα 3 έως και 9 θα εκτελεστούν δύο (2) 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 τις όποιες εικόνες προκύψουν κατά την εκτέλεση 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;

Οδηγίες:

Α. Οι εργασίες είναι σε ομάδες από ένα (1) μέχρι τρία (3) άτομα. Κάθε άτομο μπορεί να υποβάλει εργασία σε μία μόνο ομάδα κάθε φορά.

Β. Οι εργασίες θα πρέπει να αναρτώνται στο eClass (όχι e-mail, weTransfer, dropbox, ή άλλο link) σε ένα αρχείο zip (όχι rar) εντός της προβλεπόμενης προθεσμίας. **Προσοχή**: Κάθε ομαδική εργασία θα υποβάλλεται μόνο από ένα μέλος της ομάδας (εσείς επιλέγετε ποιος/ποια).

Γ. Εκπρόθεσμές υποβολές χάνουν το 2% του βαθμού, για κάθε μέρα που περνά, μετά την προθεσμία υποβολής.

Δ. Κάθε εργασία πρέπει να συνοδεύεται από:

- Τα **αρχεία .py** που περιέχουν τις απαντήσεις στα ερωτήματα
- Μια **αναφορά** σε pdf **(και μόνο σε pdf)**, που θα ακολουθεί όλες τις καλές πρακτικές, όπως αυτές περιγράφονται στο έντυπο: «Καλές πρακτικές checklist.pdf»

Να θυμάστε ότι:

- Ο κώδικας *πρέπει* να συνοδεύεται από κατάλληλα σχόλια, στα αγγλικά.
- Αν κάτι δεν διευκρινίζεται, έχετε το δικαίωμα να κάνετε όποια υλοποίηση σας βολεύει.
- Οι βιβλιοθήκες που θα χρησιμοποιήσετε *πρέπει* να μπορούν να εγκατασταθούν μέσω του pip.
- Ο κώδικας *πρέπει* να τρέχει σε Google Colab.
- Η υποβολή γίνεται αποκλειστικά μέσω του eclass.
- Δεν θα δοθεί παράταση στην προθεσμία υποβολής.

Σημαντική παρατήρηση: εργασία που *δεν* συνοδεύεται από γραπτή αναφορά βαθμολογείται με μηδέν (0).

Ημερομηνία Παράδοσης: **Βλέπε μενού Εργασίες, στο eclass**.