Πανεπιστήμιο Μακεδονίας
Τμήμα Εφαρμοσμένης Πληροφορικής
7ο εξάμηνο | Ακαδημαϊκό Έτος: 2022 – 2023

Μηχανική Μάθηση

3η Εργασία (Μη επιβλεπόμενη μάθηση – Συσταδοποίηση)

Στην εργασία αυτή θα αναπτύξετε συνδυαστικά μοντέλα προβολής σε νέο χώρο (PCA) / τεχνικών συσταδοποίησης (clustering), τα οποία θα χρησιμοποιήσετε πάνω στα δεδομένα του dataset **fashion-mnist**. Παρέχονται στο ακόλουθο link: https://github.com/zalandoresearch/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 δεδομένα, όταν κάνετε προβολή σε νέο χώρο, μέσω της ανάλυσης σε κύριες συνιστώσες.

Ο κώδικας που θα αναπτύξετε πρέπει να κάνει τα ακόλουθα:

1. Θα φορτώνει τα δεδομένα του fashion-mnist.
2. Θα διαχωρίζει τα δεδομένα σε τρία σύνολα: train, validation & test data.
3. Θα εκτελεί ανάλυση σε κύριες συνιστώσες (PCA) πάνω στα δεδομένα του train set. Προσοχή: κατά την PCA επιλέξτε όποιο ποσοστό διασποράς επιθυμείτε να διαφυλάξετε, αρκεί να βρίσκεται στο διάστημα $[85\%, 100\%]$.
4. Θα χρησιμοποιεί τον υπάρχοντα μετασχηματισμό (που προήλθε από τα train data) και θα τον εφαρμόσετε πάνω στα validation data. Τώρα χρησιμοποιείστε τον αντίστροφο μετασχηματισμό, για να τις επαναφέρετε στον αρχικό χώρο.
5. Τυπώστε τυχαία εικόνες από το validation set (τουλάχιστον μία από κάθε κλάση) καθώς και τις ανακατασκευασμένες, όπως αυτές προήλθαν από τον αντίστροφο μετασχηματισμό.
6. Θα χρησιμοποιεί τον υπάρχοντα μετασχηματισμό (που προήλθε από τα train data) και θα τον εφαρμόσετε πάνω στα test data.
7. Θα χρησιμοποιεί τρεις διαφορετικές τεχνικές συσταδοποίησης (οποίες επιθυμείτε εσείς) για να δημιουργήσει υποομάδες, πάνω στα:
    a. Χωρίς επεξεργασία test data
    b. PCA transformed test data
8. Θα υπολογίζει τους δείκτες απόδοσης που παρουσιάστηκαν στο εργαστήριο και έναν ακόμα της επιλογή σας.
9. Θα παρουσιάζει ενδεικτικά αποτελέσματα ομαδοποίησης για τυχαίες εικόνες.

Χρησιμοποιώντας τα αποτελέσματα του αλγορίθμου, και γραφικές παραστάσεις που θα φτιάξετε στο excel, θα συντάξετε μια έκθεση στην οποία θα παρουσιάζετε τα συμπεράσματα σας, θα κάνετε συγκριτικές αξιολογήσεις.

Η αναφορά πρέπει να απαντάει στα ακόλουθα ερωτήματα:

Πότε έχετε καλύτερες τιμές στα clustering performance scores; όταν έχετε raw data ή όταν έχετε pca projected data;

Με βάση τα clustering performance scores, ποιος είναι ο ιδανικός αριθμός clusters για το dataset που έχετε; Παίρνετε το ίδιο αποτέλεσμα όταν έχετε raw και pca projected data;

Ποια clustering τεχνική σας δίνει τα καλύτερα αποτελέσματα; Τελικά ποιος είναι ο καλύτερος συνδυασμός PCA ή raw data / clustering approach;

Παρουσιάστε τα αποτελέσματα ομαδοποίησης του καλύτερου δυνατού συνδυασμού, με βαση τα clustering metrics. Πρακτικά μιλώντας είναι τα αποτελέσματα αποδεκτά; Αν όχι, τι θα μπορούσατε να βελτιώσετε/κάνετε διαφορετικά;

Οδηγίες:

Α. Οι εργασίες είναι σε ομάδες μέχρι τέσσερα (4) άτομα. Κάθε άτομο μπορεί να υποβάλει εργασία σε μία μόνο ομάδα κάθε φορά.

Β. Οι εργασίες θα πρέπει να αναρτώνται στο eClass σε ένα αρχείο zip (όχι rar) εντός της προβλεπόμενης προθεσμίας. Δεν θα δοθεί παράταση. **Προσοχή**: Κάθε ομαδική εργασία θα υποβάλλεται μόνο από ένα μέλος της ομάδας (εσείς επιλέγετε ποιος/ποια).

Γ. Κάθε εργασία πρέπει να συνοδεύεται από:

- Ένα και μόνο ένα αρχείο .py θα περιέχει τις απαντήσεις στα ερωτήματα
- Μια **αναφορά** σε pdf με τα ακόλουθα στοιχεία:
    - Εξώφυλλο: 1 σελίδα, περιλαμβάνει τα στοιχεία των φοιτητών της ομάδος, όνομα μαθήματος, ημερομηνία, τμήμα και λοιπά σχετικά στοιχεία.
    - Συγκεντρωτικός πίνακας περιεχομένων, εικόνων, και λοιπών γραφημάτων που παραθέτετε στην αναφορά.
    - Ενότητα εισαγωγή: 1 σελίδα, περιγράφετε το πρόβλημα (*χωρίς* να αντιγράψετε αυτούσια την εκφώνηση της άσκησης)
    - Μέθοδοι που εφαρμόστηκαν: από 2 μέχρι 10 σελίδες, περιγράφετε τις μεθόδους που χρησιμοποιήσατε και παραθέτετε τα σχετικά αποτελέσματα. Φροντίστε να είναι ξεκάθαρο στο ποιο ερώτημα αναφέρεστε.
    - Συμπεράσματα: 1 σελίδα, με βάση τα αποτελέσματα τι προτείνετε, ποιο μοντέλο αποδίδει καλύτερα, τι θα μπορούσε να γίνει για περαιτέρω βελτίωση στην απόδοση.
    - Η αναφορά θα περιέχει γραφικές παραστάσεις κάθε είδους και πίνακες αξιολόγησης των αποτελεσμάτων που πρέπει να συνοδεύονται (έκαστο) από μια τουλάχιστον παράγραφο με σχολιασμό.

Φροντίστε ώστε:

- Ο κώδικας να συνοδεύεται απαραιτήτως από κατάλληλα σχόλια.
- Να έχει γίνει συντακτικός και ορθογραφικός έλεγχος στην αναφορά που θα υποβάλετε.
- Οι προτάσεις να είναι κατανοητές και μικρές σε έκταση.
- Οι εικόνες να ***μην*** έχουν προκύψει από print screen. Αν το πρόγραμμα δημιουργεί μια εικόνα αποθηκεύστε την κανονικά (jpg ή png), πριν την χρησιμοποιήσετε.
- Οι γραφικές παραστάσεις να περιλαμβάνουν ονόματα στους άξονες και λεζάντα. Σκοπός είναι να γίνεται κατανοητό τι δείχνει, με μια ματιά.
- Αν κάτι δεν διευκρινίζεται, έχετε το δικαίωμα να κάνετε όποια υλοποίηση σας βολεύει. Φροντίστε να μπορείτε να εξηγήσετε τι ακριβώς κάνατε στον κώδικα.
- Οι βιβλιοθήκες που θα χρησιμοποιήσετε *πρέπει* να μπορούν να εγκατασταθούν μέσω του pip.
- Ο κώδικας *πρέπει* να τρέχει σε Google Colab.

Καταληκτική Ημερομηνία Παράδοσης: **23 Ιανουαρίου 2023**. Δεν θα δοθεί παράταση.