Τμήμα
Εφαρμοσμένη Πληροφορική

Ακαδημαϊκό Έτος
2022 – 2023

Συγγραφείς
Λύσανδρος-Ιωάννης Φάσσος
(dai16011)
Σωτήριος Πασχάλης
(dai16003)

Μάθημα
Μηχανική Μάθηση

Ακαδημαικό Εξάμηνο
7ο

Επιβλέπων Καθηγητής
Ευτύχιος Πρωτοπαπαδάκης

Εργασία
3η

Καταληκτική Ημερομηνία Παράδοσης
12 Δεκεμβρίου 2022

Τύπος Εργασίας
Προγραμματιστική με φύλλο Αναφοράς

Περιεχόμενα

Εισαγωγή

Στην συγκεκριμένη εργασία εξετάζονται τεχνικές συσταδοποίησης (clustering) δεδομένων, αρχικά πάνω σε ακατέργαστα (raw) δεδομένα και σε δεδομένα που έχουν υποστεί ανάλυσης κυρίων συνιστωσών (PCA = Principal Component Analysis). Έπειτα οι τεχνικές αυτές αξιολογούνται βάσει τεσσάρων μετρικών έκαστη. Αυτές είναι οι: Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Fowlkes-Mallows Score.

Μέθοδοι που χρησιμοποιήθηκαν

Για το ερώτημα 1, φορτώνεται το σύνολο δεδομένων (dataset) fashion-MNIST από την βιβλιοθήκη TensorFlow, το οποίο αφορά είδη ρουχισμού. Περιέχει 70.000 στοιχεία, τα οποία στοιχεία είναι εικόνες 28x28px. Η κάθε εικόνα ανήκει σε μία από 10 classes, οι οποίες μας πληροφορούν για το τι είδος ρουχισμού είναι το κάθε στοιχείο.

Στη συνέχεια, στο ίδιο block κώδικα υλοποιείται το ερώτημα 2. Γίνεται ο διαχωρισμός train - validation δεδομένων, καθώς το dataset ήδη έρχεται με διαχωρισμό train-test δεδομένων, 60.000 - 10.000 αντίστοιχα. Από τα 60.000 train δεδομένα, επιλέγεται το 10% αυτών για να χρησιμοποιηθούν ως δεδομένα validation. Άρα, ο τελικός διαχωρισμός των train - test - validation δεδομένων είναι 54.000 - 10.000 - 6.000 αντίστοιχα.

Στα ερωτήματα 3 και 4, σε διαφορετικά block κώδικα έκαστο, εκτελείται η PCA (με ποσοστό διασποράς 95%). Στο πρώτο block, του ερωτήματος 3, γίνεται ο μετασχηματισμός των δεδομένων στο training data set, αφού αναπαρασταθούν τα δεδομένα του στον 2D χώρο μέσω της reshape συνάρτησης. Στο δεύτερο block, γίνεται η ίδια διαδικασία για το validation data set, με την διαφορά ότι επιπροσθέτως πραγματοποιείται και ο αντίστροφος μετασχηματισμός μέσω της inverse_transform συνάρτησης.

Στο block κώδικα του ερωτήματος 5, επιλέγονται από το αρχικό validation data set 10 ρούχα από κάθε είδος (class) ρουχισμού, καθώς και 10 αντίστοιχα από το ανακατασκευασμένο validation data set που έχει υποστεί PCA. Η αναζήτηση γίνεται τυχαία και σαρώνει έως 30 στοιχεία στην τρέχουσα υλοποίηση. (index < 30)

Στο ερώτημα 6 γίνεται ο μετασχηματισμός με PCA πάνω στο test data set, όπως αυτός εφαρμόστηκε προηγουμένως και στο train data set.

Τα επόμενα τρία block κώδικα είναι η εφαρμογή ισάριθμων τεχνικών συσταδοποίησης (clustering) πάνω στα ακατέργαστα και στα PCA-επεξεργασμένα δεδομένα, και τύπωση των αποτελεσμάτων τους σε γραφήματα, δηλαδή όλα τα υπόλοιπα ερωτήματα. Οι επιλεγμένες τεχνικές συσταδοποίησης είναι οι:

- K-μέσοι (K-means)
- Φασματική (Spectral)
- Γκαουσιανή μίξη (Gaussian Mixture)

Για κάθε τεχνική, υπολογίζονται τέσσερις μετρικές αποτελεσματικότητας της συσταδοποίησης. Αυτές είναι οι:

- Silhouette Score (όσο πιο κοντά στο 1, τόσο καλύτερη συσταδοποίηση)

- Calinski-Harabasz Score (όσο υψηλότερη τιμή, τόσο καλύτερη συσταδοποίηση)
- Davies-Bouldin Score (όσο χαμηλότερη τιμή, τόσο καλύτερη συσταδοποίηση)
- Fowlkes-Mallows Score. (όσο υψηλότερη τιμή, τόσο καλύτερη συσταδοποίηση)

Για την τεχνική K-Means σε ακατέργαστα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette: -0.095927104764886
- C-H: 2003.9746088016102
- D-B: 7.281398077445287
- F-M: 0.4219812291157372 7

Για την τεχνική K-Means σε PCA-επεξεργασμένα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette: -0.0959376785568457 2
- C-H: 1996.8836646944133
- D-B: 7.598250946074073
- F-M: 0.4228663866675373

Για την τεχνική Spectral σε ακατέργαστα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette: -0.0489025155340924 1
- C-H: 2.498612198444206 5
- D-B: 422.11796257473964
- F-M: 0.10078872255863454

Για την τεχνική Spectral σε PCA-επεξεργασμένα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette:-0.03651168893602093 5
- C-H: 1.2510479377632602
- D-B: 487.99550220485935
- F-M: 0.10085965603525981

Για την τεχνική Gaussian Mixture σε ακατέργαστα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette: -0.057971123414160095
- C-H: 1751.977723756689
- D-B: 13.451800927178112
- F-M: 0.4381485027174446 5

Για την τεχνική Gaussian Mixture σε PCA-επεξεργασμένα δεδομένα, τα υπολογισθέντα σκόρ είναι:

- Silhouette: -0.08568071867921485
- C-H: 2302.5314846224105
- D-B: 74.8833623557068
- F-M: 0.44356668468104576

Συμπεράσματα

Από τα αποτελέσματα των σκόρ των τεχνικών συσταδοποίησης μπορούμε να εξάγουμε τα εξής συμπεράσματα:

- Η τεχνική K-Means παρουσίασε ελαφρώς χειρότερα αποτελέσματα σε όλες τις μετρικές, πάνω στα PCA-επεξεργασμένα δεδομένα, αλλά τα σκόρ και στις δύο περιπτώσεις ήταν επαρκώς καλά, διαγραμματικά 7/10 συστάδες που δημιουργήθηκαν είναι σχετικά διακριτές, με τις υπόλοιπες τρείς να είναι σχηματισμένες αλλά σχετικά δυσδιάκριτες.
- Η Spectral τεχνική όχι μόνο απέδωσε χειρότερα σε όλες τις μετρικές στα PCA-επεξεργασμένα δεδομένα, αλλά και στις δύο περιπτώσεις οι τιμές που υπολογίστηκαν ήταν εξαιρετικά κακές, η τεχνική απέτυχε να αποδώσει ξεκάθαρες συστάδες, και αυτό φαίνεται διαγραμματικά, με τις

παρατηρήσεις να τοποθετούνται φαινομενικά τυχαία στις κλάσεις τους.

- Η Gaussian Mixture τεχνική παρουσίασε ενδιαφέροντα αποτελέσματα, καθώς δύο μετρικές βελτίωσαν το αποτέλεσμα τους (C-H και F-M) και δύο χειροτέρεψαν (Silhouette και D-B) όταν η GM εφαρμόστηκε στα PCA-επεξεργασμένα δεδομένα. Η D-B χειροτέρεψε έντονα, καθώς η τιμή της υπερπενταπλασιάστηκε. Το σκόρ σιλουέτας επίσης χειροτέρεψε. Η μετρική C-H όμως καλυτέρεψε πολύ έντονα (~2302 από ~1751). Αυτό διαγραμματικά αποτυπώνεται με την δημιουργία συστάδων που μοιάζουν με λωρίδες, αλλά μόνο 6/10 δείχνουν ξεκάθαρες. Οι υπόλοιπες τέσσερις δημιουργούνται διασκορπισμένες και είναι εξαιρετικά δυσδιάκριτες.

Για το συγκεκριμένο dataset, η τεχνική K-Means παράγει πρακτικά τα καλύτερα αποτελέσματα, καθώς οι συσταδοποιήσεις που

δημιουργούνται είναι οι πιο ευδιάκριτες και συνεπώς αποδεκτές για χρήση. Οι συστάδες που παράγονται από τις άλλες δύο τεχνικές δεν είναι αποδεκτές για χρήση, ειδικά αυτές της Spectral τεχνικής.