Πυθαρούλιος Κωνσταντίνος
sdi1600142

Ιγιάμου Περισανίδης Σίμων
sdi1600051

Project 2019-2020

Εργασία 3

Το πρόγραμμα αναπτύχθηκε στα πλαίσια του μαθήματος Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα 2019-2020 για το Τμήμα Πληροφορικής και Τηλεπικοινωνιών του ΕΚΠΑ.

Αρχεία

Αρχεία

- **evaluate.py:** Κώδικας για το ερώτημα Α.
- **firstlayer.py:** Κώδικας για το ερώτημα Β.
- **preprocessinput.py:** Τροποποιεί αρχεία για να έρθουν σε μορφή αναγνώσιμη από τον κώδικα της εργασίας 2 (π.χ. nn_presentations.csv).
- **results*.csv:** Αποτελέσματα clustering.

Εκτέλεση

$ python3 evaluate.py –i [inputFile] –o [outputFile] ή
$ python3 evaluate.py –-input='inputFile' –output='outputFile'

Το ίδιο ισχύει και για τα firstlayer.py, preprocessinput.py.

Παράδειγμα:
$ ./python3 evaluate.py –input='nn_representation.csv' –output='output.csv'
cluster –i vectors_dataset_small.csv –o outputfile.txt –c cluster.conf

Περιγραφή τρόπου προσέγγισης

Υπάρχουν επεξηγηματικά σχόλια σε όλη την έκταση του κώδικα, για να είναι εύκολη η ανάγνωση του. Ως αποτέλεσμα οι λεπτομέρειες την υλοποίησης είναι προφανείς από αυτά. Ωστόσο στη συνέχεια υπάρχουν κάποια γενικά σχόλια.

Ερώτημα Α(evaluate.py)

Οι συναρτήσεις για τον υπολογισμό του MSE και MAE, υπάρχουν έτοιμες από την sklearn. Στην υλοποίηση της MAPE, αποφεύγουμε την διαίρεση με το 0, γιατί διαφορετικά το αποτέλεσμα ενδέχεται να ήταν inf.

Ερώτημα Β(firstlayer.py)

Φτιάχνουμε ένα νέο μοντέλο (model_new) το οποίο αποτελείται από το πρώτο layer του WindDenseNN, μέσω της συνάρτησης Model() του keras.

Ερώτημα Γ

Τα αποτελέσματα των 4 εκτελέσεων βρίσκονται στα αρχεία results*.csv.

Συγκρίσεις

Τρέχουμε το πρόγραμμα της 2^{ης} εργασίας , βάζοντας σαν είσοδο τα διανύσματα 128 διαστάσεων του nn_representations.csv και τα διανύσματα 64 διαστάσεων που προκύπτουν από το ερώτημα Β. Οι μετρήσεις έγιναν με τους αλγορίθμους K-means++ για το initialization, Lloyd για τα assignments και Mean centroids για τα update.

Για k=4:

Διανύσματα 128 διαστάσεων

```
Algorithm:  Init:k-means++ Assignment:lloyd Update:mean

CLUSTER-0 {size: 7415, centroid: newCentroid}

CLUSTER-1 {size: 3096, centroid: newCentroid}

CLUSTER-2 {size: 5712, centroid: newCentroid}

CLUSTER-3 {size: 7765, centroid: newCentroid}

clustering_time: 11.7174 seconds

Silhouette: si:[0.279487, 0.226372, 0.278954, 0.329632] stotal: 0.278611
```

Διανύσματα 64 διαστάσεων

```
Algorithm:  Init:k-means++ Assignment:lloyd Update:mean

CLUSTER-0 {size: 10693, centroid: newCentroid}

CLUSTER-1 {size: 4604, centroid: newCentroid}

CLUSTER-2 {size: 1828, centroid: newCentroid}

CLUSTER-3 {size: 6863, centroid: newCentroid}

clustering_time: 5.20109 seconds

Silhouette: si:[0.64898, 0.350831, 0.288794, 0.438421] stotal: 0.431757
```

Παρατηρούμε μια σημαντική βελτίωση του δείκτη αξιολόγησης Silhouette για τα 64-διάστατα διανύσματα, έναντι αυτών με 128 διαστάσεις. Επίσης, όπως και ήταν αναμενόμενο, παρατηρούμε μια βελτίωση στον χρόνο εκτέλεσης της συσταδοποίησης. Η πρώτη εκτέλεση χρειάστηκε σχεδόν τον διπλάσιο χρόνο απ' την δεύτερη. Συμπερασματικά, η μείωση διάστασης είχε πολύ θετική επιρροή στην συσταδοποίηση για k=4.

Για k=12:

Διανύσματα 128 διαστάσεων

Algorithm: Init:k-means++ Assignment:lloyd Update:mean

CLUSTER-0 {size: 2242, centroid: newCentroid}

CLUSTER-1 {size: 1197, centroid: newCentroid}

CLUSTER-2 {size: 702, centroid: newCentroid}

…

CLUSTER-11 {size: 3618, centroid: newCentroid}

clustering_time: 16.8542 seconds

Silhouette: si:[0.172206, 0.368743, 0.319568, 0.319034, 0.298597, 0.291796, 0.354568, 0.209583, 0.291191, 0.279379, 0.315956, 0.265978] stotal: 0.29055

Διανύσματα 64 διαστάσεων

Algorithm: Init:k-means++ Assignment:lloyd Update:mean

CLUSTER-0 {size: 3864, centroid: newCentroid}

CLUSTER-1 {size: 530, centroid: newCentroid}

CLUSTER-2 {size: 220, centroid: newCentroid}

…

CLUSTER-11 {size: 7603, centroid: newCentroid}

clustering_time: 7.3656 seconds

Silhouette: si:[0.377829, 0.356887, 0.0589094, 0.19, 0.277933, 0.302266, 0.252006, 0.321299, 0.202147, 0.306918, 0.266984, 0.61184] stotal: 0.293751

Παρατηρούμε ενώ ο δείκτης αξιολόγησης Silhouette για τα 64-διάστατα και για τα 128-διάστατα είναι σχεδόν ίδιος, υπάρχει σημαντική βελτίωση στον χρόνο εκτέλεσης της συσταδοποίησης. Η πρώτη εκτέλεση χρειάστηκε σχεδόν 17 ενώ η δεύτερη μόνο 7 δευτερόλεπτα. Συμπερασματικά, η μείωση διάστασης είχε θετική επιρροή και σε αυτό το πείραμα.