Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα
Άσκηση 3

Δημήτριος Μπούσουλας 1115201500106
Άγγελος Λάλουσης 1115201500081

Κατάλογος αρχείων:

- erwtima1.py : Κάνει load το προεκπαιδευμένο μοντέλο από το WindDenseNN.h5 ώστε να κάνει πρόβλεψη με input data το nn_representations.csv. Έπειτα υπολογίζει τα MAE, MSE και MAPE και τα γράφει το predict.csv, όπως ζητά η εκφώνηση.

- erwtima2.py : Φορτώνει επίσης το μοντέλο WindDenseNN.h5 και στην συνέχεια παίρνει το layer[0] στο οποίο δίνει σαν είσοδο το nn_representations.csv και βγάζει την μορφοποίηση που τα διανύσματα είναι 64 διαστάσεων. Τέλος γράφει αυτήν την μορφοποίηση στο new_representation.csv όπως ζητά η εκφώνηση.

- WindDenseNN.h5 : Το προεκπαιδευμένο μοντέλο.

- nn_representations.csv : Το input dataset.

- actual.csv : Τα actual αποτελέσματα.

- predicted.csv : Η έξοδος του πρώτου ερωτήματος.

- new_representations.csv : Η έξοδος του δεύτερου ερωτήματος.

- output_128_4.txt, output_128_12.txt, output_64_4.txt, output_64_12.txt : Τα 4 output του clustering με αντίστοιχο representation και k (clustering)

- Clustering/ : Φάκελος που περιέχει τον κώδικα της δεύτερης άσκησης σε C για το clustering. Καθώς επίσης το configuration file αλλά και τα 2 representations (64 και 128) που χρησιμοποιούνται ως inputs. Συγκεκριμένα τα αρχεία της δεύτερης άσκησης είναι τα εξής:

  main.c, input.c, write_output.c, initialization.c, assignment.c, update.c, manhattan_distance.c, lsh.c, lsh_functions.c, silhouette.c, structs.h, functions.h και Makefile

Το εκτελέσιμο ονομάζεται cluster και η εντολή εκτέλεσης είναι:
./cluster –i <input file> –c <configuration file> -o <output file>

Αλλαγές Κώδικα Άσκησης 2:

Στον κώδικα του clustering δεν συμπεριλάβαμε όσα αρχεία και συναρτήσεις αφορούσαν τα curves. Επίσης αλλάξαμε το input.c ώστε να διαβάζει σωστά τα csv που δίνονται σαν είσοδο (καθώς αντί για id δίνονται ημερομηνίες)

Παραδοχές:

Στο configuration δίνεται το k του clustering, το k και το L του lsh (αν επιλεχθεί το lsh ως assignment), καθώς και η επιλογή των συναρτήσεων για initialization, assignement και update του clustering. Default τιμές ως πιο γρήγορο και ακριβή συνδυασμό έπειτα από δοκιμές επιλέξαμε: (k-means++, Lloyds Assignment, PAMean)

Αποτελέσματα μοντέλου:

Μέσο Απόλυτο Σφάλμα (MAE) : 0.0526
Μέσο Απόλυτου Ποσοστού Σφάλμα (MAPE) : 37.70%
Μέσο Τετραγωνικού Σφάλματος (MSE) : 0.0047

Τα ακριβή αποτελέσματα βρίσκονται στην πρώτη γραμμή του predicted.csv

Σύγκριση Αποτελεσμάτων Clustering:

- new_representation.csv(64 διαστάσεις):
  - k_clusters = 4:
    Clustering time: 3 seconds
    Silhouette: [s1:0.288794, s2:0.313159, s3:0.267603, s4:0.648980]
    stotal: 0.452256

  - k_clusters = 12:
    Clustering time: 26 seconds
    Silhouette: [s1:0.152439, s2:0.171717, s3:0.215451, s4:0.602500, s5:0.160702, s6:0.216948, s7:0.161280, s8:0.129758, s9:0.246653, s10:0.159947, s11:0.219594, s12:0.152985]
    stotal: 0.306584

- nn_representations.csv(128 διαστάσεις):
  - k_clusters = 4:
    Clustering time: 9 seconds
    Silhouette: [s1:0.225813, s2:0.278553, s3:0.228538, s4:0.199729]
    stotal: 0.236717

- k_clusters = 12:
  Clustering time: 28 Seconds
  Silhouette: [s1:0.342898, s2:0.298567, s3:0.227174, s4:0.219791, s5:0.198722, s6:0.248005, s7:0.115340, s8:0.201929, s9:0.210779, s10:0.133278, s11:0.207455, s12:0.176455]
  stotal: 0.208018

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι ο χρόνος της συσταδοποίησης είναι παρόμοιος ανάμεσα στα δύο αρχεία όταν έχουν το ίδιο k_cluster. Αντίθετα υπάρχει αύξηση του χρόνου όταν στο ίδιο αρχείο ο αριθμός των cluster από 4 γίνει 12.

Επίσης παρατηρώντας σε κάθε περίπτωση τα αποτελέσματα της silhouette καταλαβαίνουμε ότι γίνεται καλύτερη συσταδοποίηση όταν πρόκειται για το αρχείο των 64 διαστάσεων (new_representation.csv) σε σχέση με αυτό των 128 (nn_representations.csv). Συγκεκριμένα η 128 βγάζει stotal περίπου 0.24 και 0.21 για κ=4 και κ=12 αντίστοιχα. Ενώ η 64 βγάζει stotal περίπου 0.45 και 0.31 για κ=4 και κ=12 αντίστοιχα.