

Τεχνικές Εξόρυξης Δεδομένων Μεγάλης Κλίμακας

Χειμερινό Εξάμηνο 2017-2018

1η Άσκηση, Ημερομηνία παράδοσης: Έναρξη Εξεταστικής Χειμερινού Εξαμήνου
Ομαδική Εργασία (2 Ατόμων)

Σκοπός της εργασίας

Σκοπός της εργασίας είναι η εξοικείωσή σας με τα βασικά στάδια της διαδικασίας που ακολουθούνται για την εφαρμογή τεχνικών εξόρυξης δεδομένων, ήτοι: συλλογή, προ-επεξεργασία / καθαρισμός, μετατροπή, εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού *Python* με την χρήση του εργαλείου *SciKit Learn*.

Περιγραφή των Δεδομένων

Η εργασία σχετίζεται με την κατηγοριοποίηση χωροχρονικών δεδομένων (τροχιές λεωφορείων). Τα Dataset δίνονται σε αρχεία CSV. Οι διαφορετικές στήλες είναι διαχωρισμένες με τον χαρακτήρα ','.

Δίνονται δυο αρχεία:

1. **train_set.csv**:

Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας και περιέχει τα εξής πεδία:

- *JourneyPatternId*: Η γραμμή που ακολουθούσε το λεωφορείο, η τιμή αυτή εμπεριέχει τις διαφορετικές κατευθύνσεις κάθε γραμμής (πχ. 224-0: ΕΛ.ΒΕΝΙΖΕΛΟΥ - ΚΑΙΣΑΡΙΑΝΗ και 224-1: ΚΑΙΣΑΡΙΑΝΗ - ΕΛ.ΒΕΝΙΖΕΛΟΥ).
- *VehicleID*: Ένας unique αριθμός ο οποίος προσδιορίζει κάθε όχημα.
- *Timestamp*: Ο χρόνος που καταγράφηκε το στίγμα GPS του λεωφορείου.
- *Longitude*: Το γεωγραφικό μήκος του στίγματος.
- *Latitude*: Το γεωγραφικό πλάτος του στίγματος.

2. **test_set.csv**

test_set_a1.csv

test_set_a2.csv:

Τα αρχεία αυτά θα χρησιμοποιηθούν για να κάνετε προβλέψεις για δεδομένα διαφορετικά από αυτά που χρησιμοποιήσατε για την εκπαίδευση των μοντέλων σας. Το αρχείο αυτά περιέχουν σε κάθε γραμμή τη διαδρομή μιας τροχιάς. Το Format της τροχιάς στα αρχεία είναι το παρακάτω:

$$[[t_1, lon_1, lat_1], [t_2, lon_2, lat_2], ..., [t_N, lon_N, lat_N]]$$

- t_1 : η χρονική στιγμή που καταγράφηκε το στίγμα.
- lon_1 : Το γεωγραφικό μήκος του στίγματος.
- lat_1 : Το γεωγραφικό πλάτος του στίγματος.
- $[t_i, lon_i, lat_i]$ είναι το i -στο σημείο της τροχιάς.
- N : το πλήθος των σημείων της τροχιάς.

Ερώτημα 1

(Α) Προ-επεξεργασία των Δεδομένων

Στο σημείο αυτό καλείστε να εξάγετε από το αρχείο εκπαίδευσης τις διαφορετικές διαδρομές (trips) που ακολούθησε κάθε όχημα. Η κάθε διαδρομή αποτελείται από το ταξινομημένο σύνολο των στιγμάτων που στάλθηκαν από κάθε όχημα κατά τη διάρκεια μιας διαδρομής (JourneyPatternId). Πιο συγκεκριμένα θα πρέπει να εντοπίσετε για κάθε όχημα πότε ξεκινάει και ολοκληρώνεται μια διαδρομή (εναλλαγή του πεδίου JourneyPatternId).

Στο παρακάτω παράδειγμα φαίνονται τρεις διαφορετικές διαδρομές για το όχημα 41564.

JourneyPatternId	VehicleID	Timestamp	Longitude	Latitude
224-1	41564	t_0	lon_0	lat_0
224-1	41564	t_1	lon_1	lat_1
224-0	41564	t_2	lon_2	lat_2
224-0	41564	t_3	lon_3	lat_3
224-0	41564	t_4	lon_4	lat_4
224-0	41564	t_5	lon_5	lat_5
224-1	41564	t_6	lon_6	lat_6
224-1	41564	t_7	lon_7	lat_7

Πίνακας 1: Στίγματα των λεωφορείων

Τελικά θα πρέπει να αποθηκεύσετε τις διαφορετικές διαδρομές που εξάγατε με το όνομα "trips.csv". Θα πρέπει να ακολουθήσετε το παρακάτω format:

$TripId$; $JourneyPatternId$; $[[t_1, lon_1, lat_1], [t_2, lon_2, lat_2], \dots, [t_N, lon_N, lat_N]]$

- $TripId$: Μοναδικός αριθμός της τροχιάς που εξήχθη.
- $JourneyPatternId$: Η γραμμή στην οποία υπάγεται η τροχιά.

- $[[t_1, lon_1, lat_1], [t_2, lon_2, lat_2], \dots, [t_N, lon_N, lat_N]]$: Η ακολουθία των σημείων της τροχιάς συνοδευόμενα από τον χρόνο.

Σύμφωνα με τον Πίνακα 1 θα πρέπει να εξαχθούν και να αποθηκευτούν στο αρχείο “trips.csv” οι παρακάτω διαδρομές:

0	224-1	$[[t_0, lon_0, lat_0], [t_1, lon_1, lat_1]]$
1	224-0	$[[t_2, lon_2, lat_2], [t_3, lon_3, lat_3], [t_4, lon_4, lat_4], [t_5, lon_5, lat_5]]$
2	224-1	$[[t_6, lon_6, lat_6], [t_7, lon_7, lat_7]]$

Πίνακας 2: Διαφορετικές διαδρομές που εντοπίστηκαν από τα στίγματα του Πίνακα 1

(B) Καθαρισμός των Δεδομένων

Στο σημείο αυτό καλείστε να απομακρύνετε προβληματικές-θορυβώδεις τροχιές από το αρχείο “trips.csv”. Για να το κάνετε αυτό θα πρέπει για κάθε τροχία να εξάγετε τις παρακάτω τιμές:

- TotalDistance: Το συνολικό μήκος της διαδρομής (σε χιλιόμετρα).
- MaxDistance: Η μέγιστη απόσταση μεταξύ δύο διαδοχικών σημείων (σε χιλιόμετρα).

Τελικά θα πρέπει να αφαιρέσετε από το dataset όσες διαδρομές έχουν TotalDistance μικρότερο από 2km και όσες διαδρομές έχουν MaxDistance μεγαλύτερο από 2km.

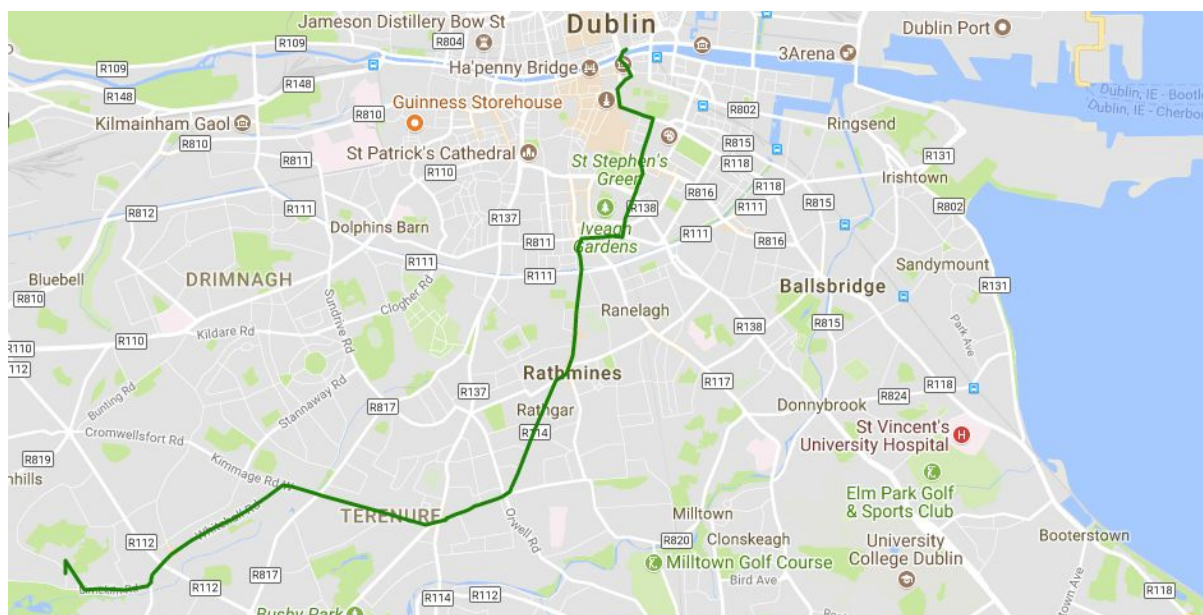
Αποθηκεύσετε τις διαδρομές που απέμειναν σε ένα νέο αρχείο με όνομα “tripsClean.csv”.

Καταγράψτε στην αναφορά σας:

- Πόσες διαδρομές υπήρχαν συνολικά στο dataset “trips.csv”.
- Πόσες διαδρομές διαγράφονται εφαρμόζονται κάθε φίλτρο ξεχωριστά.
- Πόσες διαδρομές παραμένουν τελικά στο “tripsClean.csv”.

(Γ) Οπτικοποίηση των Δεδομένων

Στο σημείο αυτό καλείστε να οπτικοποιήσετε 5 διαδρομές από διαφορετικές γραμμές λεωφορείων (journeyPatternID), όπως φαίνεται στην παρακάτω εικόνα. Για την οπτικοποίηση μπορείτε να χρησιμοποιήσετε τη βιβλιοθήκη της Python [gmaplot](#).



Ερώτημα 2

(Α-1) Εύρεση κοντινότερων γειτόνων

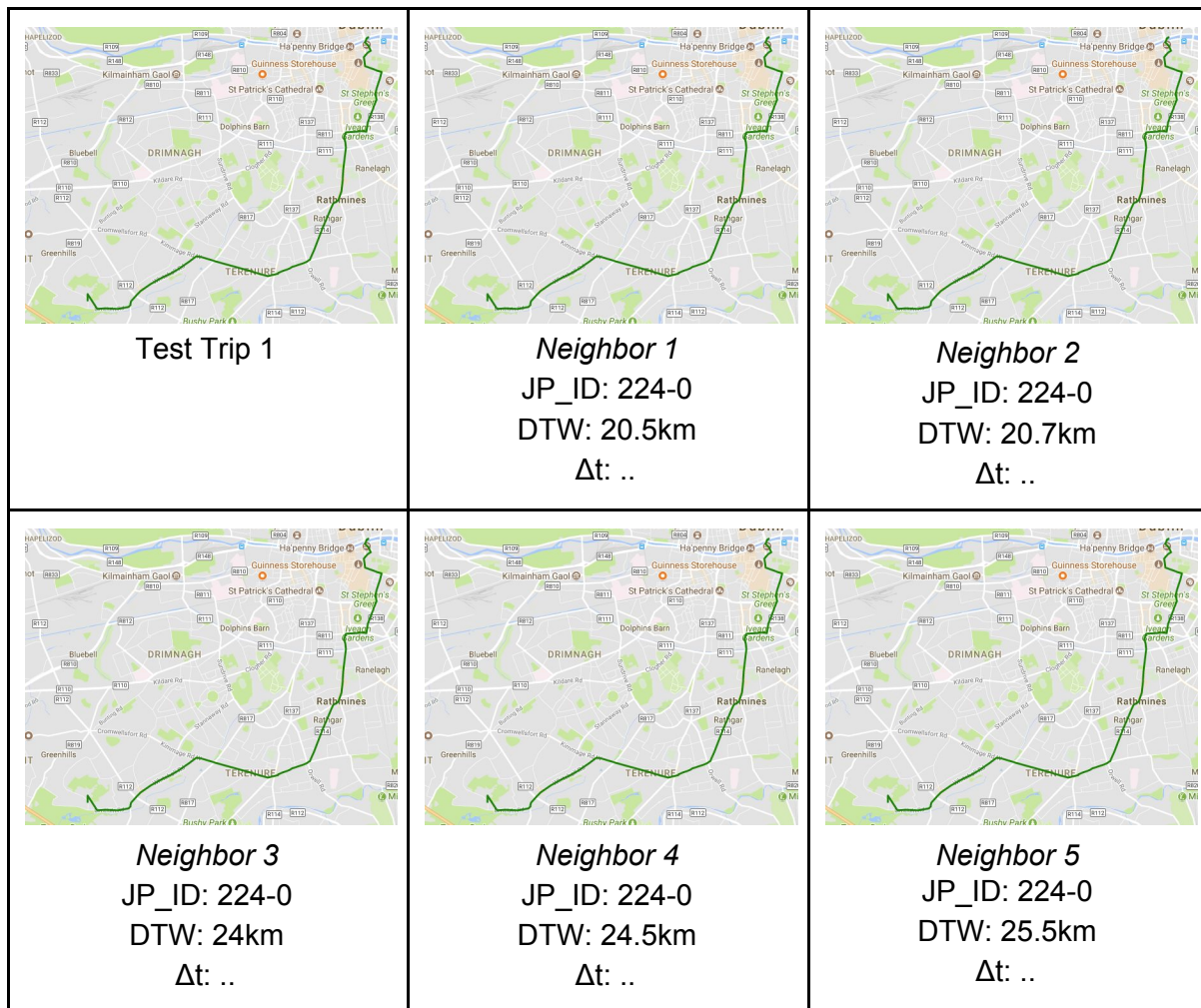
Στο σημείο αυτό καλείστε να εντοπίσετε τους κοντινότερους γείτονες χρησιμοποιώντας την τεχνική Dynamic Time Warping (DTW). Θα σας δοθεί το αρχείο “test_set_a1.csv” το οποίο θα περιέχει ένα σύνολο από διαδρομές. Για κάθε μια από τις διαδρομές αυτές θα πρέπει να βρείτε τους 5 κοντινότερους γείτονες από το dataset “tripsClean.csv”.

Οι γεωγραφικές αποστάσεις ανάμεσα σε δυο σημεία GPS θα πρέπει να υπολογιστούν με τον [τύπο Harversine](#) εκφρασμένες σε km.

Για κάθε μια από τις διαδρομές του αρχείου “test_set_a1.csv” θα πρέπει να παρουσιάσετε τα παρακάτω:

- Το JourneyPatternId για κάθε έναν από τους γείτονες που εντοπίστηκαν.
- Την DTW απόσταση με κάθε έναν από τους 5 γείτονες.
- Την οπτικοποίηση της διαδρομής που εξετάστηκε και επίσης την οπτικοποίηση των πέντε γειτόνων (6 εικόνες).
- Το χρόνο που απαιτήθηκε από το το πρόγραμμα σας για τον εντοπισμό των πλησιέστερων γειτόνων.

Χρησιμοποιήστε το παρακάτω format για την παρουσίαση των αποτελεσμάτων σας.



(A-2) Εύρεση κοντινότερων υποδιαδρομών

Στο σημείο αυτό θα σας δοθεί το αρχείο “test_set_a2.csv” το οποίο περιέχει ένα σύνολο από διαδρομές. Καλείστε για κάθε μια από αυτές να εντοπίσετε τα k κομμάτια των διαδρομών (του αρχείου “tripsClean.csv”) που είναι παρόμοια. Στο ερώτημα αυτό θα χρησιμοποιήσετε την τεχνική Longest Common Subsequence (LCSS). Διο σημεία θα γίνονται *match* εάν η απόσταση του δε ξεπερνάει τα 200m.

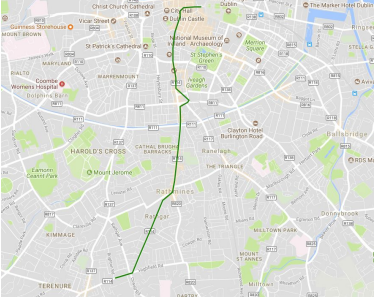
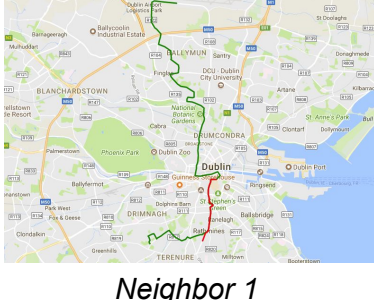
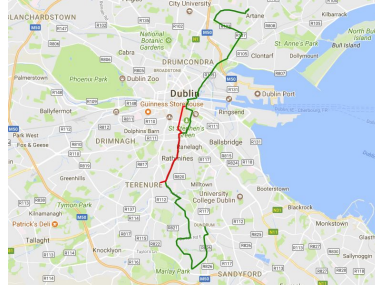
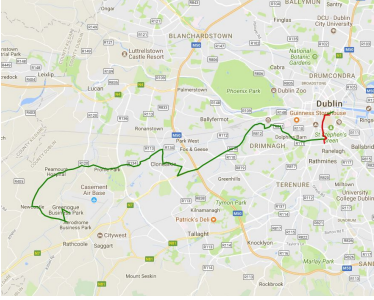
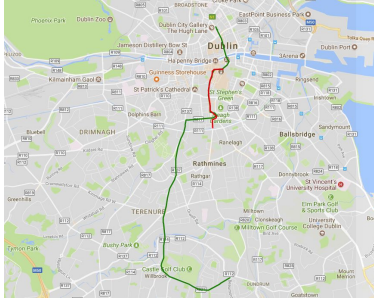
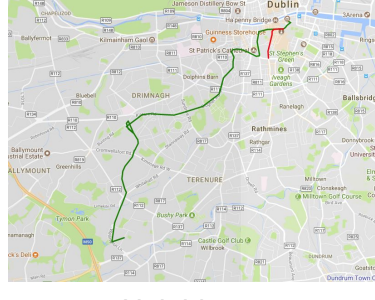
Οι γεωγραφικές αποστάσεις ανάμεσα σε δυο σημεία GPS θα πρέπει να υπολογιστούν με τον [τύπο Haversine](#) εκφρασμένες σε km.

Για κάθε μια από τις διαδρομές του αρχείου “test_set_a2.csv” θα πρέπει να παρουσιάσετε τα παρακάτω:

- Το JourneyPatternId για κάθε έναν από τους γείτονες που εντοπίστηκαν.
- Τον αριθμό των σημείων που έχουν γίνει *match* με κάθε έναν από τους 5 γείτονες.
- Την οπτικοποίηση της διαδρομής που δίνεται και επίσης την οπτικοποίηση των πέντε πλησιέστερων υποδιαδρομών που εντοπίστηκαν με κόκκινο χρώμα και με πράσινο χρώμα ολόκληρη τη διαδρομή του γείτονα (6 εικόνες).

- Το χρόνο που απαιτήθηκε από το το πρόγραμμα σας για τον εντοπισμό των πλησιέστερων γειτόνων.

Χρησιμοποιήστε το παρακάτω format για την παρουσίαση των αποτελεσμάτων σας.

 <p>Test Trip 1</p>	 <p>Neighbor 1 JP_ID: 224-0 #Matching Points: 25 Δt: ..</p>	 <p>Neighbor 2 JP_ID: 224-0 #Matching Points: 22 Δt: ..</p>
 <p>Neighbor 3 JP_ID: 224-0 #Matching Points: 16 Δt: ..</p>	 <p>Neighbor 4 JP_ID: 224-0 #Matching Points: 16 Δt: ..</p>	 <p>Neighbor 5 JP_ID: 224-0 #Matching Points: 7 Δt: ..</p>

(B) Εξαγωγή των Features για Κατηγοριοποίηση

Σε αυτό το ερώτημα θα πρέπει να εφαρμόσετε ένα δισδιάστατο Grid στις συντεταγμένες των διαδρομών, προκειμένου να τις αναπαραστήσετε σαν ένα σύνολο από κελιά. Για παράδειγμα, αν οι συντεταγμένες ενός σημείου “πέφτουν” στο κελί “C1,1” του Grid, τότε το σημείο θα αντικατασταθεί με το “C1,1”.

Για παράδειγμα η διαδρομή που ακολουθεί το λεωφορείο στην παρακάτω εικόνα μπορεί να αναπαρασταθεί από την παρακάτω ακολουθία κελιών του Grid:



"C0,2;C1,3;C1,4;C1,5;C2,6;C3,6;C4,6;C5,7;C5,6;C6,6"

(Γ) Κατηγοριοποίηση

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω 3 μεθόδους Classification για να εντοπίσετε το JourneyPatternId ενός trip:

- K-Nearest Neighbors.
- Logistic Regression
- Random Forest

Θα πρέπει να χρησιμοποιήσετε την αναπαράσταση των trips (η οποία βασίζεται στο Grid) που βρήκατε στο παραπάνω ερώτημα για να εκπαιδεύσετε τους κατηγοριοποιητές.

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου χρησιμοποιώντας 10-fold Cross Validation χρησιμοποιώντας τη μετρική Accuracy.

Beat the Benchmark

Τέλος θα πρέπει να πειραματιστείτε με όποια μέθοδο Classification θέλετε, κάνοντας οποιαδήποτε προ-επεξεργασία στα δεδομένα επιθυμείτε με στόχο να ξεπεράσετε όσο περισσότερο μπορείτε την απόδοση σας στο προηγούμενο ερώτημα. Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε.

Το καλύτερο μοντέλο σας θα πρέπει να χρησιμοποιηθεί για την πρόβλεψη των γραμμών των διαδρομών που περιέχονται στο αρχείο "test_set.csv". Τις προβλέψεις σας θα πρέπει να τις εισάγετε στο αρχείο "testSet_JourneyPatternIDs.csv". Το format του αρχείου

“testSet_JourneyPatternIDs.csv”, το οποίο θα περιέχει τις κατηγορίες των άρθρων που δίνονται στο Test set φαίνεται παρακάτω:

Test_Trip_ID	Predicted_JourneyPatternID
1	224-0
2	250-0
...	

Για το αρχείο “testSet_JourneyPatternIDs.csv” θα πρέπει να χρησιμοποιηθεί αυστηρά η παραπάνω μορφοποίηση διαχωρίζοντας τα δυο πεδία με τον χαρακτήρα TAB (‘\t’) και επίσης θα πρέπει στην πρώτη γραμμή να υπάρχουν οι δυο επικεφαλίδες (Test_Trip_ID και Predicted_JourneyPatternID) και ακολούθως οι προβλέψεις του μοντέλου σας στις επόμενες γραμμές διευκρινίζοντας το ID της διαδρομής από το test set και το αντίστοιχο JourneyPatternID.

Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα Ass1_όνοματεπώνυμο1_AM1_ονοματεπώνυμο2_AM2. Ο φάκελος θα περιέχει:

1. Ένα κείμενο με τον σχολιασμό στα πειράματα που κάνατε και στις μεθόδους που δοκιμάσατε σε μορφή PDF. Η αναφορά σας θα πρέπει να περιέχει και τα αποτελέσματα και δε θα πρέπει να ξεπερνάει τις 30 σελίδες.
2. Τα ζητούμενα αρχεία εξόδου.
3. Τους χρόνους εκτέλεσης για τα ερωτήματα 2-(A1), 2-(A2) και 2-(Γ).
4. Τα αρχεία κώδικα που γράψατε.

Το εκτενές κείμενο που θα παραδώσετε, θα περιέχει την περιγραφή των δοκιμών σας και οτιδήποτε σκεφτείτε για να δείξετε τι δοκιμές κάνατε, για ποιο λόγο έχουν τα συγκεκριμένα αποτελέσματα οι μέθοδοι που επιλέξατε, πως λειτουργούν αυτές οι μέθοδοι και σχολιασμό των αποτελεσμάτων σας. Όλες οι εργασίες θα αξιολογηθούν στη βάση της σωστής τεκμηρίωσης και στο βαθμό που υλοποιούν τα ζητούμενα της εργασίας.