Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2023-2024

Διδάσκοντες:
Καθηγητής Β. Μεγαλοοικονόμου,
Αναπληρωτής Καθηγητής Χ. Μακρής

Γλώσσα Υλοποίησης

Ως γλώσσα υλοποίησης της άσκησης ορίζεται η Python. Μπορείτε να χρησιμοποιήσετε όποια βιβλιοθήκη επιθυμείτε αρκεί να την συμπεριλάβετε στην αναφορά σας.

Σύνολο δεδομένων

Το σύνολο δεδομένων *Human Activity Recognition Trondheim (HARTH)* [1] που θα βρείτε στον σύνδεσμο https://archive.ics.uci.edu/dataset/779/harth περιέχει δεδομένα που συλλέχθηκαν από δύο επιταχυνσιόμετρα που φορούσαν 22 συμμετέχοντες στην έρευνα για περίπου 2 ώρες σε ένα περιβάλλον ελεύθερης κίνησης. Οι αισθητήρες ήταν τοποθετημένοι στο δεξί μηρό και το κάτω μέρος της πλάτης των συμμετεχόντων και οι εγγραφές του συνόλου δεδομένων έχουν χαρακτηριστεί βάσει της φυσικής δραστηριότητας τους κάθε χρονική στιγμή.

Ερώτημα 1

Πραγματοποιήστε μια πρώτη ανάλυση του συνόλου δεδομένων αλλά και κατάλληλες γραφικές παραστάσεις για αυτό έτσι ώστε να το κατανοήσετε καλύτερα. Πιο συγκεκριμένα, καλείστε να υπολογίστε τα βασικά συγκεντρωτικά στατιστικά μεγέθη για τις δοθέντες τιμές, να ανακαλύψετε αν η μορφή των γραφικών παραστάσεων ακολουθεί συγκεκριμένα μοτίβα

[1] Logacjov,Aleksej, Kongsvold,Atle, Bach,Kerstin, Bårdstu,Hilde Bremseth, and Mork,Paul Jarle. (2023). HARTH. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC90.

αλλά και να προσπαθήσετε να εντοπίσετε συσχετίσεις μεταξύ των διάφορων στηλών του συνόλου δεδομένων αλλά και μεταξύ των στατιστικών στοιχείων που υπολογίσατε.

Ερώτημα 2

Προσπαθήστε να εκπαιδεύσετε 3 ταξινομητές: έναν βασισμένο σε Neural Networks, έναν σε Random Forests και έναν σε Bayesian Networks, οι οποίοι να μαντεύουν το είδος της φυσικής δραστηριότητας των χρηστών κάθε χρονική στιγμή βάσει των μετρήσεων τόσο εκείνη την στιγμή όσο και τις προηγούμενες. Αξιολογήστε και συγκρίνετε τα μοντέλα σας χρησιμοποιώντας τις γνωστές μετρικές για την ταξινόμηση.

Ερώτημα 3

Επιχειρήστε να χωρίσετε τους συμμετέχοντες σε συστάδες με βάση τη δραστηριότητα τους το δίωρο των μετρήσεων. Μετασχηματίστε το σύνολο δεδομένων με κατάλληλο τρόπο ώστε αυτός ο διαχωρισμός να είναι εφικτός. Χρησιμοποιήστε τουλάχιστον 2 διαφορετικούς αλγορίθμους συσταδοποίησης και συγκρίνετε τα αποτελέσματα τους.

Παραδοτέα

1. Τα αρχεία κώδικα που υλοποιούν τα ζητούμενα των ασκήσεων.
2. Μια αναφορά σε μορφή pdf η οποία θα πρέπει να περιέχει τα ακόλουθα:
    - Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
    - Σύντομη περιγραφή της διαδικασία υλοποίησης.
    - Σχολιασμό των τελικών αποτελεσμάτων.

Διαδικαστικά

1. Η άσκηση μπορεί να υλοποιηθεί είτε **ατομικά** είτε σε **ομάδες των δύο**.
2. Η άσκηση μπορεί να υποβληθεί έως και **τρεις ημέρες πριν την ημερομηνία της γραπτής εξέτασης** του μαθήματος στις **23:59**.
3. Η άσκηση θα εξεταστεί προφορικά σε ημερομηνία που θα ανακοινωθεί στο τέλος του εξαμήνου.
4. Η υποβολή της άσκησης πρέπει να γίνει μέσω του eclass του μαθήματος.
5. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.