ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Απαλλακτική Εργασία 2023-2024 (Ομάδες των 1-3 ατόμων)
Ημερομηνία παράδοσης: θα ανακοινωθεί

Σκοπός της εργασίας είναι η εξοικείωση με ένα πραγματικό σύνολο δεδομένων και η εφαρμογή τεχνικών Αναλυτικής Δεδομένων & Μηχανικής Μάθησης πάνω σε αυτό. Θα επιλέξετε ένα από τα δύο παρακάτω σύνολα δεδομένων:

- US Accidents Dataset (https://www.kaggle.com/sobhanmoosavi/us-accidents): Αποτελείται από περίπου 2,8 εκατομμύρια εγγραφές, και περιέχει πληροφορίες σχετικά με αυτοκινητιστικά ατυχήματα στις Η.Π.Α κατά το χρονικό διάστημα Φεβ. 2016 – Δεκ. 2021 (ανανεώνεται σε ετήσια βάση). Για τους σκοπούς της εργασίας θα επιλέξετε (με τυχαία δειγματοληψία) 100.000 εγγραφές.
- MovieLens small Dataset (https://grouplens.org/datasets/movielens/latest/): Αποτελείται από περίπου 100.000 εγγραφές, και περιέχει πληροφορίες σχετικά με κριτικές χρηστών σε κινηματογραφικές ταινίες (τελευταία ανανέωση: 09/2018).

Βήμα 1: Προπαρασκευή δεδομένων (Data proprocessing)

Αφού κατεβάσετε το dataset που επιλέξατε, προχωρήστε σε όποια προπαρασκευαστική εργασία (επιλογή, οπτικοποίηση, καθαρισμό, μετασχηματισμό, δειγματοληψία, κλπ.) θεωρείτε απαραίτητη ώστε: α) να «καθαρίσετε» τα δεδομένα από ελλιπείς ή εσφαλμένες τιμές, εάν υπάρχουν (π.χ., συμπλήρωση κενών πεδίων, απαλοιφή ακραίων τιμών), β) να κανονικοποιήσετε – διακριτοποιήσετε τα δεδομένα (π.χ. για αντιμετώπιση των συνεχών πεδίων τιμών), γ) να μειώσετε τον όγκο των δεδομένων (π.χ. μείωση διαστάσεων). Επίσης θα πρέπει να κάνετε μια απλή στατιστική ανάλυση, σε μορφή ιστογραμμάτων, box plots κλπ., των πιο βασικών (κατά τη γνώμη σας) χαρακτηριστικών του dataset.

Βήμα 2: Συσταδοποίηση (Clustering)

Έχοντας εξοικειωθεί με το dataset, το επόμενο βήμα της πειραματικής σας διαδικασίας είναι η χρήση τεχνικών συσταδοποίησης, προκειμένου να ανακαλύψετε ιδιότητες του dataset και πρότυπα που δεν είναι προφανή με μια απλή στατιστική ανάλυση. Σε αυτό το στάδιο, σημαντικό ρόλο παίζει η μοντελοποίηση του προβλήματος (τι ακριβώς ψάχνετε να εντοπίσετε). Διαδικαστικά, αφού επιλέξετε (α) τα χαρακτηριστικά του dataset τα οποία θα αποφασίσετε να εξετάσετε και (β) μια κατάλληλη μετρική απόστασης/ομοιότητας, χρησιμοποιήστε μέσω του

εργαλείου Scikit-Learn δύο διαφορετικές τεχνικές συσταδοποίησης (K-means, DBSCAN), συζητήστε τα αποτελέσματα και την επίπτωση των παραμέτρων των μεθόδων σε αυτά, και συγκρίνετέ τα ως προς την ποιότητα/αποτελεσματικότητα της συσταδοποίησης (π.χ. scatter plots, clustering metrics).

Βήμα 3: Ταξινόμηση (Classification/Regression)

Τελευταίο βήμα της πειραματικής σας διαδικασίας είναι η χρήση μοντέλων ταξινόμησης με στόχο την ανάθεση ενός αντικειμένου σε προκαθορισμένες κατηγορίες (κλάσεις). Όπως πριν, και σε αυτό το στάδιο, σημαντικό ρόλο παίζει η μοντελοποίηση του προβλήματος (τι ακριβώς ψάχνετε να εντοπίσετε). Διαδικαστικά, αφού μετασχηματίσετε κατάλληλα το dataset στη μορφή (<Feature(s)>, <Label(s)>), δημιουργήστε χρησιμοποιώντας ScikitLearn/Tensorflow/Keras, δύο ταξινομητές (π.χ., LS-SVM, Neural Networks) και (όπως και στο προηγούμενο βήμα) συγκρίνετε τις επιδόσεις τους (π.χ., υπολογίζοντας confusion matrix, ROC-AUC curve).

Βήμα 4: Σύνοψη

Λαμβάνοντας υπόψη τα αποτελέσματα των προηγούμενων βημάτων, καταγράψτε τις παρατηρήσεις σας και 2-3 βασικά συμπεράσματα (“take-home messages”) αναφορικά με το dataset με κατάλληλη απεικόνιση/οπτικοποίηση – με άλλα λόγια, “πείτε μια ιστορία” με τα δεδομένα σας (“data story telling”).

Το τελικό παραδοτέο θα αποτελείται από ένα αρχείο zip, το οποίο θα υποβληθεί ηλεκτρονικά στην ενότητα «Εργασίες» του μαθήματος στο e-class και θα περιέχει τα εξής:

- Τεχνική αναφορά (report) με αναλυτική περιγραφή των προσεγγίσεων που ακολουθήσατε σε καθένα από τα βήματα (π.χ. παράμετροι αλγορίθμων, προπαρασκευή δεδομένων, κλπ.) και ερμηνεία των αποτελεσμάτων που προέκυψαν. Στην τεκμηρίωση πρέπει να αναγράφονται τα στοιχεία των φοιτητών της ομάδας.
- Τα αρχεία πηγαίου κώδικα (source code) και τυχόν συμπληρωματικά αρχεία (που είναι απαραίτητα για την εκτέλεση του κώδικα), καθώς και τα αποτελέσματα που παρήχθησαν (π.χ. plots).

Ζητήματα δεοντολογίας

(α) σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται, (β) σε περίπτωση αμφιβολίας για το κατά πόσο η ομάδα που αναγράφεται ήταν εκείνη που ανέπτυξε την εργασία, ενδέχεται να της ζητηθεί να την παρουσιάσει για τυχόν διευκρινίσεις.