Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2020-21

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

Γλώσσα Υλοποίησης

Ως γλώσσα υλοποίησης της άσκησης ορίζεται η python. Είστε ελεύθεροι να χρησιμοποιήσετε όποια βιβλιοθήκη επιθυμείτε αρκεί να την συμπεριλάβετε στην αναφορά σας.

Ερώτημα 1

Σας δίνετε το αρχείο *healthcare-dataset-stroke-data.csv* στο οποίο περιέχονται πληροφορίες ασθενών. Ακόμα περιέχεται και η πληροφορία ύπαρξης εγκεφαλικού επεισοδίου ή μη, την οποία και θα πρέπει να μαντέψετε.

Α. Να πραγματοποιηθεί ανάλυση του dataset και γραφική αναπαράσταση αυτής.

Β. Στο ερώτημα αυτό σας ζητείται να εντοπίσετε και να προσπαθήσετε να χειριστείτε τις ελλιπείς τιμές (missing values) με τις ακόλουθες μεθόδους:

1. Αφαίρεση στήλης
2. Συμπληρώστε τις τιμές με το μέσο όρο των στοιχείων της στήλης
3. Συμπληρώστε τις τιμές χρησιμοποιώντας Linear Regression
4. Εφαρμόστε k-Nearest Neighbors για να συμπληρώσετε τις τιμές.

Γ. Για τα νέα μητρώα που προκύπτουν στο υποερώτημα **Β**, να προβλέψετε αν ένας ασθενής είναι επιρρεπής ή όχι να πάθει εγκεφαλικό χρησιμοποιώντας Random Forest χωρίζοντας το dataset σε training-test με αναλογία 75%-25% και να μετρήσετε την απόδοσή του μοντέλου σας χρησιμοποιώντας τις μετρικές f1 score, precision και recall. Παραθέστε τα ευρήματα σας σχετικά με το πόσο επηρεάστηκε η ποιότητα της κατηγοριοποίησης. Στη συνέχεια, προσπαθήστε να βελτιώσετε τα αποτελέσματά σας πειραματιζόμενοι με τις παραμέτρους εισόδου.

Ερώτημα 2

Σε αυτό το ερώτημα σας δίνεται το αρχείο *spam_or_not_spam.csv* το οποίο περιέχει δύο στήλες. Η πρώτη στήλη περιλαμβάνει το κείμενο από διάφορα emails ενώ η δεύτερη στήλη μας πληροφορεί αν αυτά ήταν spam ή όχι: **τιμή 1 για spam, 0 αλλιώς**. Σκοπός σας είναι να προσπαθήσετε να μαντέψετε την πληροφορία της δεύτερης στήλης χρησιμοποιώντας ένα νευρωνικό δίκτυο. Για να μετασχηματίσετε το σύνολο δεδομένων που σας δόθηκε έτσι ώστε να μπορέσετε να το εισάγετε στο νευρωνικό σας δίκτυο θα πρέπει να μετατρέψετε το κείμενο των email σε διανύσματα αξιοποιώντας την τεχνική των Word Embeddings. Μπορείτε να χρησιμοποιήσετε είτε κάποιο προεκπαιδευμένο μοντέλο (π.χ. Wrod2Vev ή GloVe) είτε να εκπαιδεύσετε το δικό σας. Μετά τη δημιουργία του τελικού μητρώου, καλείστε να το χωρίσετε σε training-test dataset με αναλογία 75%-25%. Στη συνέχεια, θα πρέπει να εκπαιδεύσετε ένα νευρωνικό δίκτυο (όποιου τύπου επιθυμείτε εσείς) και να μετρήσετε την απόδοσή του χρησιμοποιώντας τις μετρικές f1 score, precision και recall.

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.