Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2019_20

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 *winequality-red.csv* στο οποίο περιέχονται οίνοι και μετρήσεις που τους χαρακτηρίζουν. Ακόμα περιέχεται και μια εκτίμηση της ποιότητάς τους από κάποιον γευσιγνώστη την οποία και θα πρέπει να μαντέψετε χρησιμοποιώντας την οικογένεια αλγόριθμων κατηγοριοποίησης SVM (Support Vector Machines).

Α. Χωρίστε το dataset σε training-test με αναλογία 75%-25% και να μετρήστε την απόδοσή του μοντέλου σας χρησιμοποιώντας τις μετρικές f1 score, precision και recall. Προσπαθήστε να βελτιώσετε τα αποτελέσματά σας πειραματιζόμενοι με τις παραμέτρους εισόδου.

Β. Σε αυτό το ερώτημα σας ζητείται να αφαιρέσετε το 33% των τιμών του της στήλης ph του training dataset και να προσπαθήσετε να χειριστείτε τις ελλιπείς τιμές με τις ακόλουθες μεθόδους:

1. Αφαιρέστε τη στήλη
2. Συμπληρώστε τις τιμές με το μέσο όρο των στοιχείων της στήλης
3. Συμπληρώστε τις τιμές χρησιμοποιώντας Logistic Regression
4. Εφαρμόστε K-means και συμπληρώστε τις τιμές που λείπουν με τον αριθμητικό μέσο όρο της συστάδας στην οποία ανήκει το δείγμα.

Στα νέα μητρώα που προκύπτουν εκπαιδεύστε ένα SVM με τις καλύτερες παραμέτρους που βρήκατε στο υποερώτημα Α και παραθέστε τα ευρήματα σας σχετικά με το πόσο επηρεάστηκε η ποιότητα της κατηγοριοποίησης.

Υπόδειξη

Για την εκπαίδευση των μοντέλων που σας ζητούνται μπορείτε να χρησιμοποιήσετε την βιβλιοθήκη της python με όνομα scikit-learn.

Ερώτημα 2

Σε αυτό το ερώτημα σας δίνεται το αρχείο *onion-or-not.csv* το οποίο περιέχει δύο στήλες. Η πρώτη στήλη περιλαμβάνει τίτλους από ψευδείς ειδήσεις ενώ η δεύτερη στήλη μας πληροφορεί αν αυτές δημοσιεύθηκαν στο γνωστό χιουμοριστικό website theonion.com ή όχι. Σκοπός σας είναι να προσπαθήσετε να μαντέψετε την πληροφορία της δεύτερης στήλης χρησιμοποιώντας ένα νευρωνικό δίκτυο. Για να μετασχηματίσετε τους τίτλους των ταινιών έτσι ώστε να δημιουργήσετε το μητρώο το οποίο θα δώσετε ως είσοδο στο υπό εκπαίδευση μοντέλο θα πρέπει να ακολουθήσετε την παρακάτω διαδικασία:

1. Θα χωρίσετε τους τίτλους σε λέξεις, δημιουργώντας ένα διάνυσμα λέξεων.
2. Από τις λέξεις θα αφαιρέσετε τις καταλήξεις τους, κρατώντας μόνο το θέμα τους (stemming).
3. Θα αφαιρέσετε από την συλλογή σας εκείνες τις λέξεις που είναι αρκετά κοινές και δεν προσφέρουν πληροφορία (stopwords removal).
4. Στις εναπομείνασες λέξεις θα αναθέσετε ως βάρος την τιμή tf-idf.
5. Θα συνδυάσετε τα διανύσματα σας για να παραχθεί το τελικό μητρώο.

Μετά τη δημιουργία του 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.

Υπόδειξη

Για τις ενέργειες επεξεργασίας φυσικής γλώσσας, μπορείτε να χρησιμοποιήσετε το εργαλείο της python με όνομα NLTK (Natural Language Toolkit)

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 email στις ηλεκτρονικές διευθύνσεις mpompotas@ceid.upatras.gr, gsmani@upatras.gr και makri@ceid.upatras.gr, vasilis@ceid.upatras.gr (το στέλνετε υποχρεωτικά και στους τέσσερις). Το email αυτό πρέπει να ακολουθεί τους εξής κανόνες:
    - Το **θέμα** του πρέπει να είναι της μορφής **dm2020_AM1[_AM2]**
    - Το **σώμα** του πρέπει να περιέχει τα στοιχεία (**ΑΜ, ονοματεπώνυμο και email**) του φοιτητή ή των φοιτητών που παραδίδουν την άσκηση
    - Τα παραδοτέα της άσκησης θα πρέπει να περιέχονται σε ένα συνημμένο αρχείο με όνομα της μορφής **dm2020_AM1[_AM2].zip**
5. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.