Πανεπιστήμιο Πατρών
Τμήμα Οικονομικών Επιστημών
Πρόγραμμα Μεταπτυχιακών Σπουδών «Εφαρμοσμένη Οικονομική και Ανάλυση Δεδομένων»

Ακαδημαϊκό έτος 2023- 2024

2η Εργασία μαθήματος «Διαχείριση Μεγάλων Δεδομένων»

https://www.youtube.com/watch?v=oY2nVQNlUB8
-The man, the myth, the legend: Scott Sterling

https://www.youtube.com/watch?v=ogx_24aoRpg
https://www.youtube.com/watch?v=ZWWjQ9vexhE
https://www.youtube.com/watch?v=kyx1U9QjWVQ
-Peter Rosenthal: ο σημαντικότερος κριτικός κινηματογράφου. Ever

Και τώρα, κάτι πιο σοβαρό:

https://www.youtube.com/watch?v=CofZ7xjGyI8
-Mike Hill, Jurassic World, Jurassic Values

https://www.youtube.com/watch?v=CHPjVgYDL6Y
-Mike Hill, Spielberg's Subtext

Εισαγωγή

Σκοπός της εργασίας είναι να αποκτήσετε μία εξοικείωση με τη χρήση αλγορίθμων στην περιοχή της κατηγοριοποίησης. Η υλοποίηση των αλγορίθμων θα πρέπει να γίνει με εργαλείο R ή/και την Python, όπου αυτό ζητείται.

Θέμα 1

Απαντήστε, ως ομάδα, στις ερωτήσεις που υπάρχουν στην παρακάτω άσκηση διαθέσιμη στο eclass:

https://eclass.upatras.gr/modules/exercise/exercise_submit.php?course=ECON1332&exerciseId=6983

Θέμα 2

Μαζί με την εκφώνηση της εργασίας, δίνονται τρεις (5) δημοσιεύσεις (αρχεία paper1.pdf, paper2.pdf, paper3.pdf, paper4.pdf, paper5.pdf). Μελετήστε τις δημοσιεύσεις αυτές και για κάθε

μία από αυτές, γράψτε μία περίληψη. Η περίληψη που θα γράψετε, θα πρέπει οπωσδήποτε να απαντά στα ακόλουθα ερωτήματα:

1) Ποιος είναι ο στόχος της εργασίας;
2) Παρουσιάζει η προσέγγιση που υιοθετείται κάτι καινοτόμο; Αν ναι τί;
3) Ποια μοντέλα/αλγόριθμοι (μηχανικής μάθησης και μη) έχουν χρησιμοποιηθεί και γιατί;
4) Ποια σύνολα δεδομένων έχουν χρησιμοποιηθεί και ποιες μεταβλητές περιείχαν τα δεδομένα αυτά;
5) Με ποιον τρόπο/μέθοδο γίνεται η αξιολόγηση των μοντέλων/αλγορίθμων που χρησιμοποιήθηκαν;
6) Ποια είναι τα κύρια ευρήματα κάθε δημοσίευσης; Σε ποια συμπεράσματα καταλήγει κάθε έρευνα;

Στην περίληψή σας μπορείτε ελεύθερα να αναφέρετε οτιδήποτε άλλο κρίνετε σκόπιμο και άξιο αναφοράς.

Θέμα 3

Απαντήστε στις δύο παρακάτω ερωτήσεις:

I. Στην ενότητα Lecture 4: Classification (https://eclass.upatras.gr/modules/units/?course=ECON1332&id=9968) μπορείτε να βρείτε το αρχείο Python-NaiveBayes-SentimentAnalysis.rar που περιέχει κώδικα σε Python ο οποίος κάνει ανάλυση συναισθήματος (sentiment analysis) πάνω σε κριτικές ταινιών χρησιμοποιώντας τον αλγόριθμο κατηγοριοποίησης Naïve Bayes. Το συμπιεσμένο αρχείο περιέχει και δύο αρχεία μορφής .csv που περιέχουν τις κριτικές χρηστών. Στα πλαίσια του θέματος αυτού, θα χρησιμοποιήσετε μόνο το αρχείο **IMDBDataset.csv** που θα βρείτε στο συμπιεσμένο αρχείο.

   Στο θέμα αυτό θα πρέπει να συγγράψετε πρόγραμμα **στη γλώσσα R**, που κάνει ανάλυση συναισθήματος στα δεδομένα του αρχείου IMDBDataset.csv με τον αλγόριθμο Naïve Bayes.

   Ειδικότερα, το πρόγραμμά σας που θα συγγράψετε σε R θα πρέπει να κάνει τα ακόλουθα:

   1) Θα κάνει χρήση των κριτικών στο αρχείο IMDBDataset.csv που υπάρχει στο παραπάνω συμπιεσμένο αρχείο και τις χρησιμοποιεί για την εκπαίδευση και την αξιολόγηση του κατηγοριοποιητή.
   2) Θα κάνει την ίδια ακριβώς προεπεξεργασία των δεδομένων (κριτικών) που κάνει και το πρόγραμμα Python. Πιο αναλυτικά, θα κάνει τα εξής:
      a. Θα αφαιρεί από όλες τις λέξεις μιας κριτικής εκείνους τους χαρακτήρες που δεν είναι γράμμα ή αριθμός. Προς τούτο, εγκαταστήστε τη βιβλιοθήκη stringr και κάντε χρήση της συνάρτησης str_replace_all που αυτή παρέχει. Καλέστε τη συνάρτηση str_replace_all με τα ακόλουθα ορίσματα, για να αντικατασταθούν όλοι οι ειδικοί χαρακτήρες μίας συμβολοσειράς: *str_replace_all(text, "[^[:alnum:]]", "")*
      b. Θα μετατρέπει όλα τα γράμματα τους κειμένου σε πεζά (μικρά)
      c. Θα αφαιρεί απ'όλες τις κριτικές τα stopwords. Για το κάνετε αυτό, εγκαταστήστε τη βιβλιοθήκη της R με όνομα tm (Text Mining) που έχει όλες τις απαραίτητες συναρτήσεις για την επεξεργασία κειμένου στην R. Κάντε χρήση του αγγλικού λεξικού stopwords. Ανατρέξτε στο εγχειρίδιο χρήσης της βιβλιοθήκης για να επιλέξετε την κατάλληλη συνάρτηση.
      d. Θα κάνει stemming όλων των λέξεων που υπάρχουν στις κριτικές. Προς τούτο, εγκαταστήστε τη βιβλιοθήκη της R Snowball (όνομα βιβλιοθήκης: SnowballC) και ανατρέξτε στο εγχειρίδιο χρήσης για το πως θα κάνετε stemming των όλων των λέξεων στην αγγλική γλώσσα.
   3) Αφού έχουν προεπεξεργαστεί τα δεδομένα του αρχείου με τον παραπάνω τρόπο, θα δημιουργείται το DocumentTermMatrix, με την κατάλληλη συνάρτηση από τη

βιβλιοθήκη tm. Για την βοήθειά σας, δείτε το άρθρο https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf προκειμένου να δείτε τί είναι το DocumentTermMatrix της βιβλιοθήκης tm και με ποιον τρόπο δημιουργείται.

4) Θα χρησιμοποιεί το 80% των κριτικών του αρχείου IMDBData.csv ως δεδομένα εκπαίδευσης και το υπόλοιπο 20% ως δεδομένα ελέγχου.
5) Θα δημιουργεί κατηγοριοποιητή βασισμένο στον Naive Bayes χρησιμοποιώντας τα δεδομένα εκπαίδευσης των κριτικών. Εγκαταστήστε τη βιβλιοθήκη της R με όνομα e1071 που παρέχει συνάρτηση που υλοποιεί τον αλγόριθμο Naïve Bayes. Ανατρέξτε στο εγχειρίδιο χρήσης της προκειμένου να δείτε ποια συνάρτηση είναι η κατάλληλη και τα ορίσματα που πρέπει να δεχθεί.
6) Για την αξιολόγηση του κατηγοριοποιητή, το πρόγραμμά σας θα πρέπει, μετά την εκπαίδευση και κατηγοριοποίηση του συνόλου ελέγχου, να εμφανίζει στην οθόνη μόνο την ακρίβεια πρόβλεψης (accuracy) στο σύνολο δεδομένων ελέγχου.

II. Με ποιον τρόπο θα χρησιμοποιήσετε τον κατηγοριοποιητή που έχετε εκπαιδεύσει στο ερώτημα I) του θέματος αυτού προκειμένου να απαντήσετε στο ακόλουθο ερώτημα:

"Αν τα σχόλια χρηστών στα κοινωνικά δίκτυα επηρεάζουν την τιμή ενός συγκεκριμένου προϊόντος τιμή ενός συγκεκριμένου προϊόντος που βρίσκεται στο supermarket."

Η απάντησή σας θα πρέπει να περιγράφει ένα σενάριο με λόγια, όπου φαίνεται πως θα χρησιμοποιούσατε τον κατηγοριοποιητή και θα φαίνονται ξεκάθαρα τα βήματα και τις ενέργειες που θα κάνατε εσείς προκειμένου να απαντήσετε στο ερώτημα αυτό. Η περιγραφή σας θα πρέπει να καλύπτει τα εξής:

1) Από ποιες πηγές θα συλλέγατε δεδομένα και τί μορφή θα είχαν αυτά
2) Πως θα επεξεργαζόσασταν τα δεδομένα αυτά και με ποιόν τρόπο
3) Ποιες στατιστικές μεθόδους ή αλγορίθμους μηχανικής μάθησης θα χρησιμοποιούσατε σε κάθε βήμα και για ποιον λόγο.

Στην περιγραφή σας, μην σας προβληματίζουν τα τεχνικά ζητήματα (δεν χρειάζεται να αναφέρετε συγκεκριμένα εργαλεία). Μπορείτε να αναφέρετε οποιαδήποτε άλλη πτυχή κρίνετε εσείς σκόπιμη. Για να πάρετε μία καλύτερη ιδέα για το πως θα περιγράψετε ένα τέτοιο σενάριο, μπορείτε να δείτε την ενότητα με τίτλο "How does it work?" από το ακόλουθο άρθρο: https://hbr.org/2015/11/a-refresher-on-regression-analysis .

Θέμα 4

Από την σελίδα https://archive.ics.uci.edu/ml/datasets/Mushroom κατεβάστε το σύνολο δεδομένων Mushroom dataset, το οποίο περιέχει τα χαρακτηριστικά διαφόρων ειδών μανιταριών όπου αναφέρεται για το εάν αυτά είναι εδώδιμα ή δηλητηριώδη. Διαβάστε προσεκτικά τις πληροφορίες της παραπάνω σελίδας και ιδιαίτερα την ενότητα "Attribute Information" που αναφέρει πως πρέπει να ερμηνευτούν οι τιμές που υπάρχουν στο σύνολο δεδομένων Mushroom dataset.
Ζητούνται τα εξής:

1) Συγγράψτε πρόγραμμα σε R και Python, το οποίο χτίζει ένα δέντρο απόφασης (decision tree), από τα δεδομένα του Mushroom dataset και το οποίο να προβλέπει αν ένα μανιτάρι είναι εδώδιμο ή δηλητηριώδες. Για το κτίσιμο του δέντρου στο περιβάλλον της R, κάντε χρήση του πακέτου *rpart*. Το σχετικό εγχειρίδιο για το πακέτο *rpart* της R δίνεται μαζί με την εκφώνηση της εργασίας. Επειδή το πακέτο αυτό δεν είναι προεγκατεστημένο στην R, θα πρέπει να εγκατασταθεί και να χρησιμοποιηθεί με τη χρήση της εντολής *library()* της R.
Για την εκπαίδευση και τον έλεγχο του μοντέλου σας θα πρέπει να δημιουργήσετε δύο τυχαία δείγματα μεγέθους ίσου με το 80 και 20% του μεγέθους του αρχικού συνόλου δεδομένων αντιστοίχως (δηλαδή το 80% του αρχικού συνόλου να χρησιμοποιηθεί για εκπαίδευση και το υπόλοιπο 20% για έλεγχο). Ο κώδικάς σας για το χτίσιμο του δέντρου απόφασης θα πρέπει να περιέχει και τις εντολές για τον διαχωρισμό αυτό. Ο

κώδικάς σας θα πρέπει να οπτικοποιεί το δέντρο απόφασης που έχει δημιουργηθεί και να εμφανίζει τις ετικέτες στους κόμβους. Επίσης, τα προγράμματά σας θα πρέπει να εμφανίζουν στην οθόνη τον πίνακα σύγχυσης καθώς και την ακρίβεια του μοντέλου (accuracy) Στην απάντησή σας συμπεριλάβετε το κώδικα σε R και Python που έχετε δημιουργήσει.

2) Για τις 30 πρώτες εγγραφές του αρχείου Mushroom Data Set (και μόνο γι'αυτές), υπολογίστε χειρωνακτικά, χρησιμοποιώντας τους κατάλληλους τύπους, το κέρδος εντροπίας (Entropy gain) του γνωρίσματος "habitat", εάν το γνώρισμα κατηγοριοποίησης είναι εκείνο το γνώρισμα που αναφέρει εάν το μανιτάρι είναι εδώδιμο ή δηλητηριώδες.
3) Συγγράψτε πρόγραμμα μόνο σε Python, το οποίο θα υλοποιεί κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes για το εάν το μανιτάρι είναι εδώδιμο ή δηλητηριώδες. Ακολουθήστε τις οδηγίες για τη δημιουργία του συνόλου εκπαίδευσης και ελέγχου που υπάρχουν στο υποερώτημα 1). Το πρόγραμμά σας θα πρέπει να οπτικοποιεί το δέντρο απόφασης που έχει δημιουργηθεί και να εμφανίζει τις ετικέτες στους κόμβους. Επιπλέον, θα πρέπει να εμφανίζεται ο πίνακας σύγχυσης καθώς και η ακρίβεια (accuracy) πρόβλεψης του μοντέλου (δέντρου απόφασης) που έχει προκύψει.

Ομάδες εργασίας

Η εργασία θα εκπονηθεί ομαδικά από τις ίδιες ομάδες που εκπόνησαν και την εργασία 1.

Χρόνος και Τρόπος Παράδοση της εργασίας

Κάθε ομάδα θα πρέπει να παραδώσει μία αναφορά σε αρχείο μορφής .pdf, γραμμένη σε LaTeχ, η οποία περιέχει τον κώδικα σε R και σε python, τις γραφικές παραστάσεις και τις απαντήσεις σας στα θέματα της εργασίας. Επιπλέον, ο κώδικας R και python που θα δημιουργήσετε για όλα τα θέματα, θα πρέπει να σταλεί και σε μορφή κειμένου, ώστε να μπορεί να εκτελείται από την R και το περιβάλλον της python. **Τα προγράμματα σε R και python που θα συγγράψετε για να απαντήσετε στα ερωτήματα της εργασίας, θα πρέπει οπωσδήποτε να περιέχουν και σχόλια που θα βοηθούν στην κατανόησή του.**
Θα ενημερωθείτε για τον τρόπο παράδοσης της εργασίας κατά τη διάρκεια των διαλέξεων.

Η καταληκτική ημερομηνία παράδοσης της 2ης εργασίας είναι η Πέμπτη 23 Νοεμβρίου 2023. Η παράδοση θα γίνει ανεβάζοντας τα αρχεία απαντήσεών σας στον χώρο της ομάδας σας στο eclass.

Ερωτήσεις/Απορίες

Για οποιαδήποτε ερώτηση ή απορία σχετικά με την εργασία μπορείτε να στείλετε email στη διεύθυνση tzagara@upatras.gr . Απορίες μπορούν επίσης (**και συστήνεται!**) να συζητηθούν κατά τη διάρκεια του μαθήματος.

Καλή επιτυχία!