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

Ακαδημαϊκό έτος 2022- 2023

3η 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, συσταδοποίησης και ανάλυσης συσχετίσεων. Η υλοποίηση των αλγορίθμων θα πρέπει να γίνει με εργαλείο R ή/και την Python, όπου αυτό ζητείται.

Θέμα 1

Μαζί με την εκφώνηση της εργασίας, δίνονται τρεις (3) δημοσιεύσεις (αρχεία ***3rdExercise-Paper1.pdf, 3rdExercise-Paper2.pdf, 3rdExercise-Paper3.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 2

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 Ι) του θέματος αυτού προκειμένου να απαντήσετε στο ακόλουθο ερώτημα:

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

Η απάντησή σας θα πρέπει να περιγράφει ένα σενάριο με λόγια, όπου 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) Ποια στατιστική μέθοδο ή αλγόριθμο μηχανικής μάθησης θα χρησιμοποιούσατε σε κάθε βήμα και για ποιον λόγο.

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 3

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.

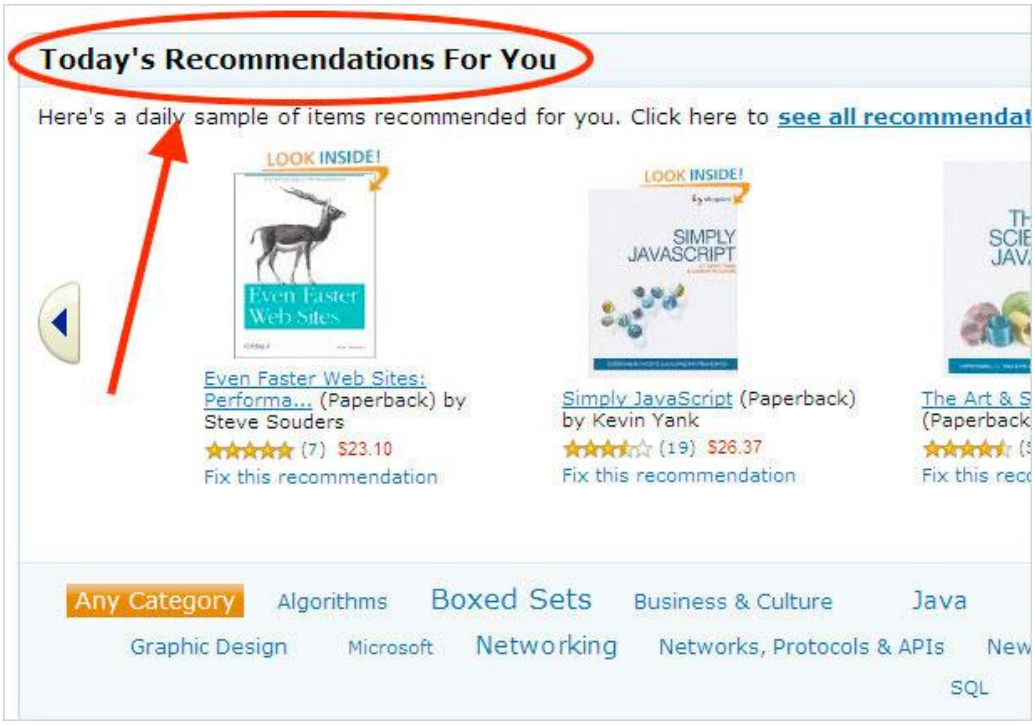
3) Συγγράψτε πρόγραμμα μόνο σε Python, το οποίο θα υλοποιεί κατηγοριοποίηση με τον αλγόριθμο Naïve Bayes για το εάν το μανιτάρι είναι εδώδιμο ή δηλητηριώδες. Ακολουθήστε τις οδηγίες για τη δημιουργία του συνόλου εκπαίδευσης και ελέγχου που υπάρχουν στο υποερώτημα 1). Το πρόγραμμά σας θα πρέπει να οπτικοποιεί 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.

Θέμα 4

Στόχος του θέματος αυτού είναι η εξοικείωση με θέματα συσταδοποίησης (clustering). Ειδικότερα θα κάνετε χρήση των αλγορίθμων K-means (και ειδικότερα μία συγκεκριμένη εκδοχή του για κατηγορικά δεδομένα) και Hierarchical clustering για την αντιμετώπιση των ζητημάτων που αναφέρονται παρακάτω.

Ζητούνται τα εξής:

I. **CAVEAT: Μην τρομάξετε με την έκταση της εκφώνησης του ερωτήματος αυτού. Αναφέρει αναλυτικά τα βήματα που θα πρέπει να εκτελέσετε .**
Στο ερώτημα αυτό σας ζητείται να φτιάξετε ένα σύστημα που ανήκει στην κατηγορία των προτάσεων/συστάσεων για ταινίες (movie recommender system) με χρήση της γλώσσας R και Python.
Τα συστήματα προτάσεων ή συστάσεων (recommender systems - https://en.wikipedia.org/wiki/Recommender_system) είναι συστήματα, τα οποία προτείνουν ή συστήνουν σε χρήστες νέα προϊόντα βάσει είτε των προτιμήσεών τους είτε βάσει προϊόντων που έχουν αγοράσει στο παρελθόν και τους άρεσαν. Ο στόχος τέτοιων συστημάτων προτάσεων/συστάσεων είναι να παρέχουν εξατομικευμένες υπηρεσίες στους χρήστες. Όλα τα σύγχρονα συστήματα ηλεκτρονικού εμπορίου όπως Amazon και eBay παρέχουν τέτοιου είδους συστήματα προτάσεων/συστάσεων.



Εικόνα 1: Παράδειγμα υπηρεσίας προτάσεων/συστάσεων βιβλίων στο Amazon. Τα αποτελέσματα που βλέπετε στην παραπάνω εικόνα, έχουν προκύψει από ένα σύστημα προτάσεων (recommender system).

Στα πλαίσια του ερωτήματος αυτού, σας ζητείται να συγγράψετε πρόγραμμα σε R και σε python, το οποίο προτείνει νέες ταινίες σε έναν χρήστη βάσει των προτιμήσεων ταινιών του συγκεκριμένου χρήστη. Για τον σκοπό αυτόν, σας δίνονται μαζί με την εκφώνηση δύο αρχεία δεδομένων: *movies.csv* και *ratings.csv*.

Το αρχείο *movies.csv* είναι αρχείο τύπου csv (comma separated values) το οποίο περιέχει τους τίτλους 9125 ταινιών μαζί με τις κατηγορίες στην οποία ανήκει κάθε ταινία (αν είναι περιπέτεια, δράμα, τρόμου, Film-Noir κλπ). Κάθε γραμμή του αρχείου movies.csv είναι της μορφής

<movieId>,<title>,Action,Adventure,Animation,Children,Comedy,Crime,Documentary,
Drama,Fantasy,Film-Noir,Horror,IMAX,Musical,Mystery,Romance,Sci-
Fi,Thriller,War,Western, (no genres listed)

η οποία ερμηνεύεται ως εξής: η τανία με κωδικό <movieId> έχει τίτλο <title>[1] και η οποία ανήκει σε μία ή παραπάνω από τις ακόλουθες κατηγορίες: Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, IMAX, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, (no genres listed) . Οι μεταβλητές που δηλώνουν κατηγορίες καλούνται μεταβλητές κατηγοριών και είναι δυαδικές μεταβλητές (dummy variables) λαμβάνοντας μόνο τις τιμές 0 ή 1 προσδιορίζοντας συνολικά τις κατηγορίες στις οποίες η ταινία ανήκει. Η ύπαρξη άσσου (1) σε μία μεταβλητή κατηγορίας σημαίνει ότι ανήκει στην κατηγορία αυτή. Η τιμή 0 σημαίνει ότι δεν ανήκει στην κατηγορία αυτή. Μία ταινία μπορεί να ανήκει σε περισσότερες από μία κατηγορία. Για παράδειγμα η γραμμή του αρχείου movies.csv

2,Jumanji (1995),0,1,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0

[1] Μαζί με τον τίτλο της κάθε ταινίας εμφανίζεται –εντός παρενθέσεων- και το έτος πρώτης προβολής της.

σημαίνει ότι η ταίνια με κωδικό 2 έχει τίτλο “Jumanji (1995)” και ανήκει στις κατηγορίες Adventure, Children και Fantasy αφού οι αντίστοιχες μεταβλητές έχουν τιμή 1 ενώ π.χ. δεν ανήκει στην κατηγορία Western αφού στην αντίστοιχη μεταβλητή υπάρχει η τιμή 0. Αν υπάρχει τιμή 1 στη μεταβλητή “(no genres listed)” αυτό σημαίνει ότι η κατηγορία της αντίστοιχης ταινίας είναι είτε άγνωστη είτε δεν μπορεί να προσδιοριστεί από τις υπόλοιπες μεταβλητές κατηγοριών. Μία ταινία μπορεί να ανήκει σε πολλές κατηγορίες όπως για παράδειγμα η ταινία Jumanji (1995) παραπάνω. Κάθε ταινία έχει μοναδικό κωδικό (movieId) που φαίνεται στο αρχείο (movieId).

Το αρχείο *ratings.csv* περιέχει αξιολογήσεις ταινιών από 671 διακριτούς χρήστες. Συνολικά περιέχει 100004 αξιολογήσεις από τους 671 χρήστες. Κάθε γραμμή του αρχείου έχει τη μορφή

<userId>, <movieId>, <rating>,<timestamp>

η οποία ερμηνεύεται ως εξής: *ο χρήστης με κωδικό <userId> αξιολόγησε με βαθμό <rating> την ταινία με κωδικό <movieId> και η αξιολόγηση έγινε στις <timestamp>*. Η τιμή <timestamp> καθορίζει την χρονική στιμγή που έγινε η αξιολόγηση και αναφέρεται στην ημερομηνία και ώρα. Ειδικότερα, η τιμή <timestamp> είναι ένας ακέραιος αριθμός που εκφράζει πόσα δευτερόλεπτα πέρασαν από τα μεσάνυχτα της 1ης Ιανουαρίου 1970 (η οποία είναι γνωστή και με το όνομα “the Epoch”) μέχρι τη στιγμή που έγινε η αξιλόγηση. Για παράδειγμα η παρακάτω γραμμή του αρχείου ratings.csv:

3, 296, 4.5, 1298862418

ερμηνεύεται ως εξής: ο χρήστης με κωδικό 3 αξιολόγησε την ταινία με κωδικό 296 με βαθμολογία 4.5 και ότι η αξιολόγηση έγινε τη χρονική στιγμή 1298862418 (δηλαδή η αξιλόγηση έγινε 1298862418 δευτερόλεπτα μετά τα μεσάνυχτα της 1ης Ιανουαρίου 1970). Οι κωδικοί ταινιών του αρχείου ratings.csv αναφέρονται στο αρχείο movies.csv και έτσι μπορούμε να δούμε ότι η ταινία με κωδικό 296 της παραπάνω γραμμής του αρχείου ratings.csv αναφέρεται στην ταινία “Pulp Fiction (1994)” όπως προκύπτει από το αρχείο movies.csv. Δηλαδή ο χρήστης με κωδικό 3 βαθμολόγησε με 4.5 την ταινία “Pulp Fiction”. Η αξιολόγηση (rating) γίνεται σε κλίμακα από 1 έως και 5, με 1 να είναι η χαμηλότερη βαθμολογία και 5 η υψηλότερη. Μπορούν επίσης να δοθούν ως βαθμολογία «μισά» δηλαδή 2.5, 3.5 κλπ. Ένας χρήστης μπορεί να βαθμολογήσει παραπάνω από μία ταινία, αλλά ένας συγκεκριμένος χρήστης μπορεί να βαθμολογήσει μία συγκεκριμένη ταινία μία μόνο φορά. Στα πλαίσια της εργασία αυτής, μπορείτε να αγνοήσετε την χρονική στιγμή της βαθμολόγησης <timestamp> του αρχείου ratings.csv.

Ο κώδικάς σας που θα συγγράψετε σε R και σε python θα πρέπει να επεξεργάζεται καταλλήλως τα αρχεία movies.csv και ratings.csv και για έναν συγκεκριμένο κωδικό χρήστη (που θα τον δίνετε εσείς στο πρόγραμμά σας), να εμφανίζει στην οθόνη ταινίες, που ταιριάζουν στις προτιμήσεις του συγκεκριμένου χρήστη και δεν τις έχει δει.

Αν και υπάρχουν διάφορες προσεγγίσεις για να φτιαχτεί ένα τέτοιο σύστημα προτάσεων/συστάσεων για το συγκεκριμένο πρόβλημα ταινιών, το σύστημα προτάσεων ταινιών που θα φτιάξετε θα βασιστεί στην εξής προσέγγιση:

Προτείνει στον χρήστη με κωδικό X να δει εκείνες τις ταινίες T που δεν έχει δει και οι οποίες μοιάζουν αρκετά με ταινίες τις οποίες τις έχει δει και του άρεσαν πολύ[2].

[2] Το 2009 η Netflix έτρεξε διαγωνισμό για την πρόβλεψη του κατά πόσο ένας χρήστης θα του άρεσε μία ταινία, γνωστό ώς Netflix Prize. Ειδικότερα, ο στόχος ήταν να βρεθεί αλγόριθμος ο οποίος μπορεί να βελτιώσει την πρόβλεψη αξιολόγησης άγνωστων ταινιών βάσει προηγούμενων αξιολογήσεων χρηστών κατά τουλάχιστον 10% σε σχέση με τον αλγόριθμο που ήδη είχε η Netflix. Το έπαθλο ήταν 1.000.000

Το σύστημα προτάσεων/συστάσεων που θα υλοποιήσετε σε R και python θα κάνει χρήστη του αλγορίθμου συσταδοποίησης K-means.

Παρακάτω περιγράφεται με λόγια ένας τέτοιος αλγόριθμος και τον οποίο θα πρέπει να υλοποιήσετε τόσο με τη γλώσσα προγραμματισμού R όσο και με τη γλώσσα προγραμματισμού python:

1) Διαβάστε το αρχείο ταινιών movies.csv
2) Κάντε συσταδοποίηση των ταινιών που διαβάσατε από το αρχείο movies.csv βάσει των κατηγοριών στις οποίες αυτές ανήκουν με τον αλγόριθμο K-means. Ο στόχος της συσταδοποίησης με τον αλγόριθμο K-means είναι, ταινίες που ανήκουν στις ίδιες κατηγορίες (και κατά συνέπεια μοιάζουν μεταξύ τους) να μπουν στην ίδια συστάδα. Θα κάνετε τη συσταδοποίηση με τον αλγόριθμο K-means λαμβάβοντας μόνο υπόψιν τις μεταβλητές που υποδηλώνουν τις κατηγορίες της τανίας από τα δεδομένα του αρχείου movies.csv (δηλαδή τις ψευδομεταβλητές Action, Adventure, Animation, Children κλπ). Ωστόσο, επειδή οι μεταβλητές κατηγοριών των ταινιών δεν λαμβάνουν συνεχείς τιμές (λαμβάνουν δυαδικές τιμές 0 και 1 και κατά συνέπεια είναι αυτό που καλούμε εικονικές μεταβλητές ή ψευδομεταβλητές - dummy variables) δεν μπορεί να γίνει χρήση συναρτήσεων συσταδοποίησης που βασίζονται στην Ευκλείδεια απόσταση. Γι'αυτόν τον λόγο θα πρέπει τόσο στο περιβάλλον της R όσο και στο περιβάλλον python να βρείτε και να εγκαταστήσετε τις κατάλληλες βιβλιοθήκες, οι οποίες θα σας επιτρέψουν να τρέξετε τον αλγόριθμο K-means με εκείνη τη μέθοδο υπολογισμού αποστάσεων των δεδομένων κατάλληλη για τα δεδομένα του αρχείου ratings.csv. Σας παραπέμπουμε στις βιβλιοθήκες amap της R και Kmodes της python τις οποίες θα πρέπει να εγκαταστήσετε στο δικό σας υπολογιστή και σας παρέχουν τις κατάλληλες εκδοχές του αλγορίθμου K-means για κατηγορικά δεδομένα. Δώστε ιδιαίτερη έμφαση στα εγχειρίδια χρήσης των βιβλιοθηκών αυτών για το ποια συνάρτηση να χρηστιμοποιήσετε και με ποια ορίσματα να την εκτελέσετε.
   Για τον προσδιορισμό του ακριβούς πλήθους των συστάδων K τόσο στο πρόγραμμα R όσο και στο πρόγραμμα python, κάντε χρήση της μεθόδου του αγκώνα (Elbow method). Ειδικότερα, τρέξτε τον αλγόριθμο συσταδοποίησης K-means με όλες τις τιμές K (κέντρων) από 2 έως και 100. Για κάθε τιμή K που θα εκτελέσετε τον αλγόριθμο K-means (K=2, 3,4, 5, … ,100) κρατήστε την τιμή της αντικειμενικής συνάρτησης που σας λέει πόσο καλή ήταν η συσταδοποίηση για την συγκεκριμένη τιμή K (όπως Sum of Squared Error, Average dispersion κλπ – μπορείτε εσείς να επιλέξετε την αντικειμενική συνάρτηση). **Απεικονίστε γραφικά τις τιμές K μαζί με την τιμή της αντικειμενικής συνάρτησης που επιλέξατε και επιλέξτε εκείνη την τιμή K η όποια παρουσιάζει τη μεγαλύτερη μείωση της αντικειμενικής συνάρτησης και συνεχίζει με μη-σημαντικές μεταβολές**. Στο γράφημα αυτό θα εφαρμόσετε τη μέθοδο του αγκώνα για τον προσδιορισμό της τιμής K (κέντρων).
3) Μόλις καταλήξετε στην κατάλληλη τιμή K (κέντρων) με την οποία προκύπτει η καλύτερη τιμή της αντικειμενικής συνάρτησης, ξανατρέξτε τον αλγόριθμο K-means() με την επιλεγείσα τιμή K για να πάρετε τις τελικές συστάδες των δεδομένων.
4) Διαβάστε το αρχείο αξιολογήσεων ratings.csv και βρείτε για κάθε ταινία του αρχείου movies.csv, τον μέσο όρο αξιολόγησης που έδωσαν σε αυτήν οι χρήστες. Θεωρείστε ότι οι τιμές αξιολόγησης ταινιών είναι ποσοτικό μέγεθος και ότι μπορεί να υπολογιστεί ο αριθμητικός μέσος όρος[3]. Σε περίπτωση που δεν θέλετε να κάνετε χρήση του μέσου όρου, μπορείτε να υπολογίσετε την επικρατούσα τιμή (mode) για κάθε ταινία.

Ευρώ . Για περισσότερες πληροφορίες δείτε https://en.wikipedia.org/wiki/Netflix_Prize και http://www.netflixprize.com/

[3] Αν και η μεταβλητή ratings φαίνεται να είναι κατηγορική, μπορείτε να υπολογίσετε τον μέσο όρο, μιας και λαμβάνει και «μισές» τιμές π.χ. 2.5, 4.5 κλπ και κατά συνέπεια μπορεί να θεωρηθεί ότι συμπεριφέρεται σαν ποσοτική μεταβλητή.

5) Επιλέξτε έναν συγκεκριμένο χρήστη π.χ. τον χρήστη με κωδικό 198 (ή οποινδήποτε άλλο). Βρείτε ποιες ταινίες έχει αξιολογήσει και βρείτε για κάθε ταινία που ο χρήστης με κωδικό 198 έχει αξιολογήσει, σε ποια συστάδα ανήκει, όπως αυτή προέκυψε από το βήμα 3) παραπάνω. Δηλαδή, από τα δεδομένα του αρχείου ratings.csv, απομονώστε τις αξιολογήσεις του χρήστη 198 και εισάγετε μία νέα μεταβλητή με όνομα *clusterId*. Για κάθε ταινία που έχει αξιολογήσει ο χρήστης 198, θέστε ως τιμή στη μεταβλητή *clusterId* τη συστάδα στην οποία αυτή η ταινία ανήκει και όπως αυτή προέκυψε από το βήμα 3).
6) Βρείτε τον μέσο όρο βαθμολογίας (ή την επικρατούσα τιμή αν θέλετε) που έχει δώσει ο χρήστης με κωδικό 198 στις συστάδες στις οποίες ανήκουν οι ταινίες που έχει αξιολογήσει. Ειδικότερα, ομαδοποιείστε[4] τις ταινίες που έχει αξιολογήσει ο χρήστης 198 βάσει της συστάδας στην οποία ανήκει κάθε ταινία, και για κάθε ομάδα βρείτε τον μέσο όρο βαθμολογίας των ταινιών που αξιολόγησε ο χρήστης 198 και που ανήκουν στην ομάδα αυτή. Με τον τρόπο αυτό μπορείτε να πάρετε μία άποψη του χρήστη για κάθε συστάδα. Για παράδειγμα αν ο χρήστης 198 έχει αξιολογήσει τις εξής ταινίες, οι οποίες ανήκουν στις παρακάτω συστάδες (βλέπε clusterId), όπως αυτές προέκυψαν από το βήμα 3):

userId	**movieId**	**rating**	**clusterId**
198	345	2.5	45
198	153	1.5	16
198	76	4.5	45
198	236	4.5	16
198	58	3.5	16

στο βήμα αυτό θα πρέπει να προκύψει το εξής αποτέλεσμα:

clusterId	**Μ.Ο. βαθμολογίας**
16	(1.5+4.5+3.5) / 3 = 3.16
45	(2.5 + 4.5) / 2 = 3.5

Το παραπάνω αποτέλεσμα ερμηνεύεται ως εξής: ο μέσος όρος βαθμολογίας ταινιών του χρήστη 198 για τη συστάδα 16 είναι 3.16 ενώ για τη συστάδα 45 είναι 3.5.

7) Ακολούθως, αφαιρέστε εκείνες τις συστάδες του χρήστη 198, οι οποίες έχουν μέσο όρο αξιολόγησης ταινιών χαμηλό. Επειδή το «χαμηλό» είναι σχετική έννοια, ορίστε ως χαμηλή αξιολόγηση μία τιμή αξιολόγησης μικρότερη από 3.5. Στο παραπάνω παράδειγμα του χρήστη 198, θα πρέπει να αφαιρεθεί η συστάδα 16, η οποία έχει μέσο όρο αξιολόγησης ταινιών μικρότερη από 3.5 από τον χρήστη. Η συστάδα 45 δεν θα πρέπει να αφαιρεθεί, μιας και έχει αξιολόγηση μεγαλύτερη ή ίση από 3.5.
8) Αν για τον χρήστη δεν υπάρχουν συστάδες με μέσο όρο αξιολόγησης ταινιών ίση ή μεγαλύτερη από 3.5 (γιατί π.χ. όλοι οι μέσοι όροι συστάδων είναι μικρότεροι από 3.5), τότε δεν μπορούν να γίνουν προτάσεις/συστάσεις ταινιών για τον χρήστη και θα πρέπει να εμφανίζεται το μήνυμα: “Sorry, no recommendations for you!”.
9) Αν υπάρχουν συστάδες με μέσο όρο αξιολόγησης μεγαλύτερη ή ίση με 3.5 για τον χρήστη 198, τότε για κάθε τέτοια συστάδα βρείτε τις 2 ταινίες με την υψηλότερη βαθμολογία εντός της συστάδας αυτής και τις οποίες δεν έχει δει ο χρήστης 198. Εμφανίστε τον τίτλο των ταινιών αυτών και αυτές οι ταινίες είναι οι συστάσεις/προτάσεις ταινιών για τον χρήστη 198 που μοιάζουν με τις προτιμήσεις του. Ειδικότερα, εμφανίστε το μήνυμα “You may also like the following movies” και από κάτω εμφανίστε τους τίτλους των ταινιών που προτείνονται στον χρήστη.

[4] Προσοχή! “Ομαδοποιείστε” όχι “συσταδοποιείστε”! Με τον όρο ομαδοποίηση εννοούμε να βάλετε στην ίδια ομάδα τις ταινίες που ανήκουν στην ίδια συστάδα.

ΣΗΜΕΙΩΣΗ: Για το πρόγραμμα που θα συγγράψετε σε R, ενδεχομένως να σας φανούν χρήσιμες οι εξής συναρτήσεις της R: subset(), match(), aggregate() και order(). Ανατρέξτε στα σχετικά εγχειρίδια για να δείτε τι ακριβώς κάνουν και πως λειτουργούν οι συναρτήσεις αυτές.

Συνοψίζοντας τα ζητούμενα του θέματος αυτού, θα πρέπει να παραδώσετε τα εξής:

a) **Κώδικα γραμμένο σε R**, ο οποίος υλοποιεί τον παραπάνω αλγόριθμο συστάσεων ταινιών σε χρήστη. Ο κώδικας σε R θα πρέπει να εμφανίζει τη γραφική παράσταση που αναφέρετε στο σημείο 2) και να εμφανίζει προτάσεις/συστάσεις ταινιών για έναν συγκεκριμένο χρήστη, που θα πρέπει να μπορεί να τον δίνει ο χρήστης του προγράμματος.
b) **Κώδικα γραμμένος σε Python**, ο οποίος υλοποιεί τον παραπάνω αλγόριθμο συστάσεων ταινιών σε χρήστη. Ο κώδικας σε python θα πρέπει να εμφανίζει τη γραφική παράσταση που αναφέρετε στο σημείο 2) και να εμφανίζει προτάσεις/συστάσεις ταινιών για έναν συγκεκριμένο χρήστη, που θα πρέπει να μπορεί να τον δίνει ο χρήστης του προγράμματος.

II. Συγγράψτε κώδικα σε R και Python, ο οποίος διαβάζει τα δεδομένα του αρχείου *europe.csv*, που δίνεται μαζί με την εκφώνηση της εργασίας και εκτελεί ιεραρχική συσταδοποίηση πάνω στο σύνολο δεδομένων του αρχείου europe.csv. Στο περιβάλλον της R κάντε χρήση των συναρτήσεων *dist()* και *hclust()* της R, ενώ στην python κάντε χρήση των κατάλληλων συναρτήσεων της βιβλιοθήκης scikit-learn που μπορείτε να βρείτε εδώ: http://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering και θα πρέπει να έχετε εγκαταστήσει στο περιβάλλον εργασίας σας.
Τόσο το πρόγραμμα σε R όσο και το πρόγραμμα σε python που θα φτιάξετε θα πρέπει να εμφανίζει ως έξοδος, το δενδρόγραμμα κλάσεων με ετικέτες το όνομα της κάθε χώρας.

III. Μελετήστε τη δημοσίευση που υπάρχει στο αρχείο **3rdExercise-Paper4.pdf** και δώστε μία συνοπτική περίληψη. Η περίληψη θα πρέπει οπωσδήποτε να αναφέρει τις τεχνικές ανάλυσης δεδομένων που παρουσιάζονται καθώς και τον λόγο που έχουν χρησιμοποιηθεί αυτές. Μπορείτε να αναφέρετε οποιοδήποτε άλλο στοιχείο κρίνετε εσείς σκόπιμο.

Θέμα 5

Στόχος του θέματος αυτού είναι η εξοικείωση με θέματα ανάλυσης συσχετίσεων (association rules) με τη χρήση του πακέτου *arules* του εργαλείου R[5]. Επειδή το πακέτο *arules* δεν είναι προεγκατεστημένο στο περιβάλλον της R, θα πρέπει να εγκατασταθεί και να χρησιμοποιηθεί στο σύστημά σας. Το εγχειρίδιο βοήθειας του πακέτου arules μπορείτε να το βρείτε εδώ: https://cran.r-project.org/web/packages/arules/arules.pdf
Για τις ανάγκες του θέματος αυτού, κατεβάστε το σύνολο δεδομένων “Fertility Data Set” από το UCI Machine Learning Repository, ακολουθώντας τον εξής σύνδεσμο: https://archive.ics.uci.edu/ml/datasets/Fertility . Η σελίδα περιέχει πληροφορίες σχετικά με την ερμηνεία των τιμών που υπάρχουν στο σύνολο δεδομένων. Επιπλέον στοιχεία για την ερμηνεία του συνόλου δεδομένων μπορούν να βρεθούν στην εξής δημοσίευση: http://cs229.stanford.edu/proj2014/Axel%20Guyon,Florence%20Koskas,Yoann%20Buratti,SemenFertilityPrediction.pdf

Το σύνολο δεδομένων “Fertility Data Set” περιέχει στοιχεία για την αξιολόγηση της γονιμότητας ανδρών μαζί με στοιχεία του τρόπου ζωής τους (lifestyle). Τα ερωτήματα που

[5] Στο θέμα αυτό, δεν θα χρειαστεί να υλοποιήσετε τον αλγόριθμο σε Python. Θα συγγράψετε το πρόγραμμά σας μόνο στη γλώσσα R.

επιχειρεί μελετήσει το θέμα αυτό είναι να ανακαλύψει παράγοντες του τρόπου ζωής που συνεμφανίζονται με την αλλαγή της γονιμότητας ανδρών.

Ειδικότερα, ζητούνται τα εξής:

I. Αφαιρέστε από το σύνολο δεδομένων “Fertility Data Set” τις στήλες που σχετίζονται με τις μεταβλητές “Age at the time of analysis” και “Number of hours spent sitting per day ene-16”. Το σύνολο δεδομένων “Fertility Data Set” χωρίς τα δύο αυτά γνωρίσματα θα αναφέρεται ως τροποποιημένο σύνολο δεδομένων “Fertility Data Set”. Ακολούθως, συγγράψτε κώδικα σε R, ο οποίος εφαρμόζει τον αλγόριθμο Apriori πάνω σε όλα τα γνωρίσματα που απομένουν του τροποποποιημένου συνόλου δεδομένων με τις προκαθορισμένες (default) τιμές. Ποιο είναι το πλήθος των κανόνων που επιστρέφονται και τί δηλώνουν οι κανόνες που επιστρέφονται ως αποτέλεσμα από τον συγκεκριμένο αλγόριθμο;

II. Για το τροποποποιημένο σύνολο δεδομένων “Fertility Data Set”, συγγράψτε κώδικα σε R ο οποίος εφαρμόζει τον αλγόριθμο Apriori για ελάχιστη υποστήριξη (support threshold) ίση με 0.02, εμπιστοσύνη (confidence) ίση με 1. Επιπλέον δώστε ως περιορισμό στο δεξί μέλος των κανόνων να υπάρχει μόνο το Diagnosis=altered. Πόσοι και ποιοι κανόνες επιστρέφονται;

III. Ένας κανόνας Υ θεωρείται περιττός, όταν υπάρχει κανόνας Χ, ο οποίος έχει μεγαλύτερο ή ίσο lift από τον Υ και επιπλέον ο Υ είναι υπερκανόνας του Χ. Ένας κανόνας έχει την γενική μορφή LHS => RHS. Ο κανόνας Υ είναι υπερκανόνας του Χ αν LHS(Y) ⊃ LHS(X) και RHS(Y) == RHS(X). Για παράδειγμα, ο Α,Β => Γ είναι υπερκανόνας του Α=>Γ. Το lift για κάθε κανόνα δίνεται από τον αλγόριθμο Apriori.
Ταξινομήστε τους κανόνες που προέκυψαν από το ερώτημα II. ως προς το lift και στη συνέχεια αφαιρέστε τους περιττούς κανόνες. Δώστε το σύνολο των κανόνων που απομένουν, μετά την αφαίρεση των περιττών κανόνων.

Ομάδες εργασίας

Η εργασία θα εκπονηθεί ομαδικά και θα είναι οι ίδιες ομάδες που εκπόνησαν τις εργασίες 1 και 2.

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 εργασία θα πρέπει να παραδοθεί πριν την τελική εξέταση του μαθήματος της εξεταστικής περιόδου Φεβρουαρίου 2023. Ειδικότερα, η καταληκτική ημερομηνία παράδοσης της 3ης εργασίας είναι μία ημέρα πριν την τελική εξέταση του μαθήματος της εξεταστικής περιόδου Φεβρουαρίου 2023.

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.

Βαρύτητα της εργασίας

Η εργασία είναι υποχρεωτική και συνεισφέρει το 10% του τελικού τους βαθμού.

Καλή επιτυχία!