ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

3 ASSIGNMENT BIG DATA

Dimitra Adami 1067738
Eleni Bonatsou 1067623
Filippos Mitsos 1019913

University of Patras
MSc Applied Economics and Data Analysis
February 2023

Contents

A Abstract **3**

B Question 1
Paper 1.2.3 **3**

C Question 4 III
Paper 4 **4**

D Question 5 **5**

A Abstract

The purpose of the assignment is to gain familiarity with the use of algorithms in area of categorization, clustering and correlation analysis. The implementation of algorithms should be done with an R tool and/or Python, where requested.

B Question 1 Paper 1.2.3

Ξεκινώντας με το πρώτο άρθρο το οποίο αναφέρεται στην υβριδική προσέγγιση φήμης με βάση τη συγχώνευση απόψεων και ανάλυση συναισθήματος. Η υβριδική προσέγγιση, διαχωρίζει κριτικές με βάση το συναίσθημα σε θετικές και αρνητικές και εφαρμόζοντας τα θεωρήματα Bayes και LSVM, στη συνέχεια γίνεται ομαδοποίηση σε κύρια σύνολα με βάση τις σημασιολογικές σχέσεις, έπειτα ο υπολογισμός της προσαρμογής και τέλος την τελική φήμη. Η δημιουργία φήμης με τη χρήση μέσου όρου αξιολογήσεων με βασικούς παράγοντες την αξιοπιστία του βαθμολογητή σε διάφορους διαδικτυακούς τόπους. Επίσης αυτό μπορεί να επιτευχθεί και χειροκίνητα μέσω tweets, για την ολοκλήρωση συσχέτισης συναισθημάτων και διάφορων γεγονότων που έχουν συμβεί. Τόσο οι κριτικές των συντακτών όσο και οι αξιολογήσεις των πελατών αυξάνουν την εμπιστοσύνη αποφάσεων. Νέα προσέγγιση αποτελεί η έκφραση σε φυσική γλώσσα. Η ανάλυση συναισθήματος γίνεται με φυσική γλώσσα, όπως τα λεξικά τα οποία αποτελούν τεχνική μηχανικής μάθησης. Μια κριτική είναι συνιστώμενη αν ο μέσος όρος των φράσεων είναι θετικός. Για να προσδιοριστεί η πολικότητα ενός κειμένου χρησιμοποιείται η μηχανή διανυσμάτων υποστήριξης (SVM) και ταξινομητή μέγιστης εντροπίας. Με 4 βήματα μπορούμε να δούμε πως φτάνουμε στην δημιουργία φήμης, ξεκινώντας με τη φάση της ταξινόμησης όπου είναι ο διαχωρισμός θετικών και αρνητικών σχολίων με βάση την πολικότητα του συναισθήματος και την αξιολόγηση σύμφωνα με το θεώρημα Bayes και τη γραμμική μηχανή διανυσμάτων υποστήριξης (LSVM). Για σύνολο δεδομένων χρησιμοποιούμε κριτικές ταινιών, 1000 θετικές και 1000 αρνητικές αντίστοιχα. Στην πρόβλεψη πιο ακριβής είναι ο Bayes. Στη συνέχεια, είναι η φάση σύντηξης και ομαδοποίησης όπου γίνεται με τον αλγόριθμο ο οποίος ξεχωρίζει τις θετικές από τις αρνητικές και τις ταξινομεί σε σύνολα με βάση τις σημασιολογικές σχέσεις. Επόμενο είναι, η προσαρμογή δημιουργίας φήμης, που υπολογίζονται οι προσαρμοσμένες τιμές για θετικές και αρνητικές τιμές με βάση τα στατιστικά στοιχεία των κύριων συνόλων γνώμης. Και τέλος είναι η δημιουργία φήμης, που υπολογίζεται προς την οντότητα-στόχο με βάση την προσαρμοσμένη τιμή που αναφέραμε και πιο πάνω για τις κριτικές χρησιμοποιώντας αριθμητικό μέσο όρο. Τα δεδομένα που συλλέξαμε για το παράδειγμα αυτό, τις κριτικές για τις ταινίες και τις αξιολογήσεις είναι από το IMDb. Τέλος το άρθρο αναφέρεται στην προσέγγιση δημιουργίας φήμης με 2 ταξινομήσεις, τον υπολογισμό της τιμής της φήμης με τον σταθμισμένο μέσο. Στη συνέχεια ήταν τα βήματα ταξινόμησης και τέλος ότι οι άχρηστες πληροφορίες – κριτικές φιλτράρονται για να μειωθεί ο χρόνος επεξεργασίας.

Συνεχίζουμε στο δεύτερο άρθρο το οποίο αναφέρεται στους προσδιοριστικούς παράγοντες τιμής Bitcoin. Τα bitcoin, εικονικά νομίσματα, αποτελούν την υψηλότερη αγοραία αξία και υψηλότερο όγκο συναλλαγών εικονικών νομισμάτων που κυκλοφορούν. Στο άρθρο αναφέρεται το GARCH, το οποίο είναι γενικευμένο αυτοπαλίνδρομο υπο συνθήκη ετεροσκεδαστικό μοντέλο. Το μοντέλο αυτό χρησιμοποιείται για την πρόβλεψη διακυμάνσεων, αλλα στο συγκεκριμένο άρθρο χρησιμοποιείται για την ερμηνεία το πώς επηρεάζονται τα bitcoin, τα οποία αποτελούν εξαρτημένη μεταβλητή. Η μηχανή διανύσματος υποστήριξης (SVM), είναι για επίβλεψη και χρησιμοποιείται για την πρόβλεψη. Αυτό το επιτυγχάνει με τον διαχωρισμό του συνόλου σε 2 κατηγορίες και στη συνέχεια βρίσκει το βέλτιστο όριο τη μεγιστοποίηση των περιθωρίων. Επίσης, το δέντρο απόφασης, αποτελεί μέσο εξόρυξης δεδομένων για την πρόβλεψη ταξινόμησης, με τη βοήθεια διακλαδώσεων ξεκινώντας από πάνω προς τα κάτω, καταλήγοντας στους πιο σημαντικούς παράγοντες οι οποίοι χρησιμοποιούνται για δείκτες απόφασης. Υπάρχουν 2 κατηγορίες, τα δέντρα ταξινόμησης ποιοτικών δεδομένων και το δέντρο παλινδρόμησης που αναλύει ποσοτικά δεδομένα. Τα δεδομένα για αυτή τη μελέτη συλλέχθηκαν από τις ιστοσελίδες, bitcoin.org, investing.com, FRED και World Gold Council, με χρονικές περιόδους από 19/07/10 – 31/12/18 με δείγμα 1922 παρατηρήσεις. Το SVM, διαχωρίζει τα δεδομένα σε 2 ομάδες. Η παράμετρος της πολυπλοκότητας ελέγχει το πόσο ευέλικτη είναι η διαδικασία για τον διαχωρισμό των ομάδων, ενώ στην ταξινόμηση με τα δέντρα απόφασης τα πειράματα ρυθμίζουν τις παραμέτρους, τον παράγοντα εμπιστοσύνης, ελάχιστο αριθμό περιπτώσεων διαχωρισμού, αλλά το μέγεθος και των 2 μεθόδων είναι 100. Τα δεδομένα χωρίζονται σε 2 σύνολα, το σύνολο εκπαίδευσης και το σύνολο δοκιμής, για 10 διασταυρώσεις και για κάθε αναδίπλωση παίρνουμε το μέσο αποτέλεσμα. Στο μοντέλο του GARCH, οι χρονοσειρές χρησιμοποιούν έλεγχο μοναδιαίας ρίζας για την επεξεργασία μη στάσιμων δεδομένων για την εμπειρική ανάλυση. Η μηδενική υπόθεση για μια ρίζα απορρίπτεται, αν τα δεδομένα είναι σταθερά συνεχίζουμε την ανάλυση. Για τον έλεγχο μεταβλητότητας χρησιμοποιείται ARCH-LM και έλεγχο πληροφόρησης AIC, όπου όσο μικρότερη η τιμή του ελέγχου τόσο καλύτερη προσαρμογή. Η μέθοδος SVM, χρησιμοποιεί ανεξάρτητες μεταβλητές με χρονική υστέρηση. Τα πειραματικά αποτελέσματα λαμβάνουν ακρίβεια 90%, για αυτό και είναι αποτελεσματική ταξινόμηση. Τέλος το άρθρο αναφέρεται για την ανάπτυξη bitcoin και πως συνδυάζει τις έρευνες για τον προσδιορισμό παραγόντων επίδρασης του bitcoin. Εφαρμόζονται, 3 μοντέλα GARCH, SVM και δέντρα απόφασης. Η SVM, έχει μεγαλύτερη ακρίβεια πρόβλεψης τιμής bitcoin, ελαχιστοποιεί το όριο σφάλματος γενίκευσης και επηρεάζεται λιγότερο από την υπερβολική προσαρμογή. Στο μοντέλο των δέντρων απόφασης το επιτόκιο της Fed έχει μεγαλύτερη επίδραση στην τιμή των bitcoin.

Τέλος το τρίτο άρθρο και τελευταίο, αναφέρεται στην πρόβλεψη ύφεσης με χρήση ταξινόμησης κατά Bayes. Συγκεκριμένα αναφέρονται 2 προσεγγίσεις κατά Bayes για πρόβλεψη, μοντέλα Markov και λογιστική παλινδρόμηση (LR). Συγκεκριμένα το θεώρημα κατά Bayes, πλούσιο σύνολο δεδομένων, δομή υστέρησης και αποτύπωση εμμονής φάσεων του οικονομικού κύκλου με τη χρήση Markov- switching. Τα σημεία καμπής του NBER χρησιμοποιούνται ως δεδομένα όπως και στην λογιστική προσέγγιση και ισοδυναμείτε με την προσέγγιση NB. Ο NB ενσωματώνει καλύτερα μεγάλο όγκο δεδομένων με πλούσια δομή υστέρησης. Η προσέγγιση κατά Bayes, χωρίζεται σε 2 τμήματα στο πρώτο γίνεται με τη χρήση συνόλου δεδομένων για το εάν θα υπάρξει ύφεση στο μέλλον με δυαδική απόκριση. Και το δεύτερο είναι για την πρόβλεψη και ταξινόμηση του οικονομικού κύκλου. Για την προσέγγιση της δυαδικής απόκρισης, χρησιμοποιούμε το πλαίσιο probitπου περιλαμβάνει μεταβλητές για την πρόβλεψη ύφεσης και το term-spread, όπου τονίζει τα σημεία ρήξης. Όταν προστεθεί το term premium και το επίπεδο επιτοκίου και τροποποιείται η εξαρτημένη μεταβλητή. Το μοντελό που χρησιμοποιεί τη θεωρία Bayes, για να μάθει αν παρατηρούμενα δεδομένα προέρχονται από μια κατηγορία. Η πιθανότητα χωρίς όρια για μια οικονομία σε ύφεση, είναι η αρχή για τη χρήση αλγόριθμου NB. Επίσης η συνεκτίμηση της εμμονής ύφεσης και επεκτάσεων βελτιώνουν το αποτέλεσμα. Τα υποδείγματα Markov χρησιμοποιούν το πότε συμβαίνουν οι μετατοπίσεις και τις πιθανότητες μετάβασης για εκτίμηση. Το υπόδειγμα εναλλαγής Markov, επιβάλουν την εμμονή για κατάσταση ύφεσης για την πρόβλεψη σημείων καμπής, διότι οι φάσεις του οικονομικού κύκλου διατηρούνται με τη πάροδο του χρόνου. Στο NB, δεν συμβαίνει αυτό διότι όλες οι περίοδοι αντιμετωπίζονται ανεξάρτητες μεταξύ τους. Ένας τρόπος προσέγγισης είναι τα μοντέλα probit και lobit, χρησιμοποιώντας συναρτήσεις σιγμοειδούς για τα παρατηρούμενα δεδομένα. Οι υποθέσεις ισοδυναμίας NB και LR δεν ισχύουν πάντα και τα όρια απόφασης είναι για να διαχωρίζουν τις κλάσεις. Τα δεδομένα που χρησιμοποιήσαμε είναι 135 μακροοικονομικές μεταβλητές, μηνιαία από το FRED-MD, ξεκινώντας από 01/1959 εως 06/2016. Οι ακατέργαστες τιμές μετασχηματίζονται, επίσης για δεδομένα πραγματικού χρόνου παίρνουμε από την Federal Reverse Bank. Τα σημεία καμπής προσδιορίζονται από τις ημερομηνίες ύφεσης (BCDC). Οι προβλέψεις γίνονται για το αν θα αρχίσει η ύφεση εντός κάποιου χρονικού ορίου. Η ενσωμάτωση υστερήσεων σε ένα σύνολο 135 μεταβλητών δεν είναι αρκετά βοηθητικό και δημιουργεί δυσκολίες. Η ποιότητα πρόβλεψης της λογιστικής παλινδρόμησης καταρρέει για μεγάλους ορίζοντες πρόβλεψης. Η δυαδική πρόβλεψη μπορεί να είναι πιο κατατοπιστική από μια απώλεια πραγματικής αξίας. Για την αντιμετώπιση σφαλμάτων κατασκευάζουμε συντελεστές στάθμισης σφαλμάτων εξαρτώμενοι από τον χρόνο. Με αυτόν τον τρόπο τιμωρούνται τα σφάλματα όταν συμβαίνουν στον επιχειρηματικό κύκλο. Στη μέση του κύκλου τιμωρούνται περισσότερο τα σφάλματα από ότι στις άκρες. Τέλος το υπόδειγμα NB υστερεί για την πρόβλεψη σημείων καμπής του οικονομικού κύκλου της LR και παρέχεται μια βαρύτητα στα σφάλματα πρόβλεψης, όπως η τιμωρία.

C Question 4 III Paper 4

Το άρθρο αυτό παρουσιάζει με τη βοήθεια παραδειγμάτων πως γίνεται η ανάλυση συστάδων των οικονομικών δεδομένων και με ποιους τρόπους. Αρχικά είναι η ανάλυση συστάδων, το οποίο αποτελεί εργαλείο πολυμεταβλητότητας διερευνητικής ανάλυσης δεδομένων. Στόχος του είναι ο εντοπισμός ομάδων παρόμοιων αντικειμένων, σύμφωνα με επιλεγμένες μεταβλητές. Βασικές προσεγγίσεις είναι η ιεραρχική ομαδοποίηση και η ομαδοποίηση k-means. Συσσωρευτική ιεραρχική ομαδοποίηση, με αντικείμενα που θεωρούνται μεμονωμένες ομάδες. Οι συστάδες συνδέονται σταδιακά μέχρι να γίνουν μια συστάδα. Η k-means είναι ορισμένο αριθμό συστάδων. Βασικός όρος στην ανάλυση είναι η ομοιότητα. Η ιεράρχηση ανάλυσης, χρησιμοποιεί ζεύγη και έχει κύριο πλεονέκτημα ότι η γραφική έξοδος γίνεται με δενδροδιάγραμα. Αν το αρχείο διαθέτει ονομαστικές μεταβλητές τότε χρειάζεται ειδικό μέτρο για την αξιολόγηση ομοιότητας. Όπως ο απλός συντελεστής ταύτισης ή αλλιώς μέτρο επικάλυψης. Για την αξιολόγηση χρησιμοποιούμε την i-ιοστή και j-ιοστή γραμμή του πίνακα. Η k-clustering, διαιρεί το σύνολο σε ορισμένο αριθμό συστάδων. Στην πρώτη περίπτωση θα έχει ως αποτέλεσμα έναν πίνακα με μονάδες και μηδενικά και το κέντρο του θα αποτελεί διάνυσμα, ενώ στη δεύτερη περίπτωση k-centroid και k-medoids επιλέγει το αντικείμενο. Το k-centroid, εφαρμόζεται σε μεγάλα δεδομένα διότι αλλιώς δημιουργούνται διαφορετικές αναθέσεις. Το k-clustering, είναι για τη διερευνητική ανάλυση δεδομένων και βρίσκει τη βέλτιστη λύση. Το k-medoids, υπολογίζει τα πλάτη με βάση την απόσταση των αντικειμένων της ίδιας συστάδας η και διαφορετικής. Τέλος έχουμε την ανάλυση συστάδων 2 βημάτων όπου ομαδοποιεί μεγάλα σύνολα δεδομένων με ποσοτικές και ποιοτικές μεταβλητές, ο αλγόριθμος BIRCH, ο οποίος τα χωρίζει σε υποσυστάδες με βάση τα χαρακτηριστικά ομαδοποιούνται και σε άλλες ομάδες.

D Question 5

Για το 5ο ερωτημα τησ εργασιασ χρησιμοποιήσαμε το FertilityDataSet από το UCIM Machine Learning Repository, το οποίο περιέχει πληροφορίες σχετικά με τον τρόπο ζωής και την ανδρική γονιμότητα.
Οι στήλες 2 και 8, διαγράφονται καθωσ αναφερονται στην ηλικία και τον αριθμό των ωρών που δαπανώνται.Εποχή, παιδικές ασθένειες, ατύχημα ή σοβαρό τραύμα, χειρουργική επέμβαση, υψηλός πυρετός, συχνότητα κατανάλωσης αλκοόλ, συνήθεια καπνίσματος και διάγνωση είναι οι επόμενες ιδιότητες που προσθέτουμε στα χαρακτηριστικά που αποτελούν τις παραμέτρους που επηρεάζουν τον τρόπο ζωής και το επίπεδο γονιμότητας των ανδρών. Οι μεταβλητές μετατρεπονται σε παράγοντες προκειμένου να εκτελεστεί ο αλγόριθμος Apriori.

Graphical representation question mark 5 :