Εργασία εξαμήνου
(συμμετοχή στον τελικό βαθμό: ~~50~~100%)

Ομάδες των 2-3 ατόμων

Στόχος της εργασίας είναι η εξοικείωση με τη διαδικασία εξόρυξης γνώσης (ΕΓ) από σύνολα δεδομένων. Για τους σκοπούς της εργασίας θα χρησιμοποιήσετε τα εργαλεία που παρουσιάστηκαν στο εργαστηριακό κομμάτι του μαθήματος. Συγκεκριμένα, για την οργάνωση των δεδομένων και την εκτέλεση αλγορίθμων εξόρυξης γνώσης θα χρησιμοποιήσετε τα εργαλεία PostgreSQL, R Project και WEKA.

Θα εργαστείτε με ένα σύνολο δεδομένων από το UCI ML Repository, συγκεκριμένα το: **Wine Quality Data Set** (http://archive.ics.uci.edu/ml/datasets/Wine+Quality). Το συγκεκριμένο σύνολο δεδομένων αφορά στην Πορτογαλική ποικιλία οίνου "vinho verde", και συγκεκριμένα στις δύο παραλλαγές της, τον ερυθρό και τον λευκό τύπο. Σκοπός είναι η προτυποποίηση της ποιότητας του κρασιού με βάση μια σειρά φυσικοχημικών δοκιμών. Αναλυτική περιγραφή του dataset θα βρείτε στο παραπάνω link, από όπου μπορείτε και να το κατεβάσετε. Τα γνωρίσματα (attributes) #1 - #11 αποτελούν τις μετρήσεις εισόδου και το γνώρισμα (attribute) #12 το χαρακτηριστικό εξόδου (για πρόβλεψη).

Έστω ότι θέλουμε να αναλύσουμε το παραπάνω σύνολο δεδομένων. Αρχικά αντιμετωπίζουμε το πρόβλημα της προπαρασκευής δεδομένων (data preprocessing) ώστε να κρατήσουμε όσα μας ενδιαφέρουν για τους σκοπούς της ανάλυσης. Κατόπιν εκτελούμε διάφορες λειτουργίες ΕΓ (classification, clustering, feature selection, association rules, outlier detection κλπ.) έχοντας προκαθορίσει το σκοπό της ανάλυσής μας. Η παραπάνω διαδικασία ενδέχεται να επαναληφθεί αρκετές φορές μέχρι το αποτέλεσμα της ανάλυσης να είναι ικανοποιητικό και η γνώση μας για τα δεδομένα να είναι επαρκής, άρα και άμεσα χρήσιμη (actionable) από τους υπεύθυνους αποφάσεων των εκάστοτε εφαρμογών.

Παρακάτω παραθέτουμε τις επιμέρους εργασίες – βήματα που πρέπει να πραγματοποιήσετε για την ανάλυση του συνόλου δεδομένων που έχετε επιλέξει:

1ο βήμα (κατασκευή ΒΔ – υλοποίηση σε PostgreSQL): Αφού κάνετε την απαραίτητη επεξεργασία μεταφοράς δεδομένων από το αρχείο, θα εισάγετε τα δεδομένα σε μια κατάλληλα σχεδιασμένη ΒΔ.

2ο βήμα (προπαρασκευή δεδομένων – υλοποίηση σε PostgreSQL με βοήθεια από άλλα εργαλεία στατιστικής επεξεργασίας, π.χ. Excel, R): Από το παραπάνω dataset θα επιλέξετε τα δεδομένα που θα χρησιμοποιήσετε για αναλυτική επεξεργασία, και θα προχωρήσετε στην όποια προπαρασκευαστική εργασία (επιλογή, καθαρισμό,

μετασχηματισμό, δειγματοληψία κλπ.) θεωρείτε απαραίτητη ώστε: α) να «καθαρίσετε» τα δεδομένα (από ελλιπείς ή εσφαλμένες - μη λογικές - τιμές), β) να κανονικοποιήσετε – διακριτοποιήσετε τα δεδομένα (π.χ. για αντιμετώπιση των συνεχών πεδίων τιμών), γ) να μειώσετε τον όγκο των δεδομένων (μείωση διαστάσεων, μείωση πλήθους εγγραφών). Ειδικά για τη μείωση του πλήθους των εγγραφών, επειδή το πλήθος εξαρχής δεν είναι μεγάλο, απλά θα πειραματιστείτε με τις διάφορες τεχνικές αλλά τελικά θα χρησιμοποιήσετε το αρχικό πλήθος για τα περαιτέρω βήματα.

3ο βήμα (classification/regression – υλοποίηση σε R): Σκεφτείτε κάποια σενάρια κατηγοριοποίησης και πρόβλεψης (όχι μόνο αυτό για το οποίο συνήθως χρησιμοποιείται το dataset) που θα μπορούσαν να ενδιαφέρουν τους αναλυτές της εκάστοτε εφαρμογής. Μην χρησιμοποιήσετε μόνο τεχνικές που διδαχθήκατε στη θεωρία, αλλά και άλλες που παρέχονται από το εργαλείο με το οποίο θα δουλέψετε. Συγκρίνετε τις διάφορες προσεγγίσεις, χωρίς να χρησιμοποιήσετε την τεχνική αυτόματης σύγκρισης των αλγορίθμων που χρησιμοποιεί προκαθορισμένη παραμετροποίηση των αλγορίθμων. Σκοπός είναι να μεγιστοποιήσετε την απόδοση κάθε αλγορίθμου ξεχωριστά «παίζοντας» με την προπαρασκευή του dataset και τις παραμέτρους του. Περιγράψτε τη διαδικασία και εξηγήστε τα αποτελέσματα που προκύπτουν. Κατά το ελάχιστο, θα πρέπει να υλοποιηθούν και να δοκιμαστούν πειραματικά ένας γραμμικός και ένας μη γραμμικός ταξινομητής, συγκεκριμένα linear classifier και k-nearest neighbor. Επιπλέον, θα πρέπει να γίνει διερεύνηση για το βέλτιστο υποσύνολο από τα χαρακτηριστικά εισόδου (#1 - #11) για την πρόβλεψη του χαρακτηριστικού-στόχου (#12), μετά από διακριτοποίηση σε κλάσεις (classification) και ως έχει (regression). Επίσης, θα πρέπει να παραχθούν τα αντίστοιχα διαγράμματα (π.χ. scatter plots) και τα confusion matrices, για την καλύτερη δυνατή παρουσίαση των αποτελεσμάτων. Η αξιολόγηση των αποτελεσμάτων να προέλθει μέσα από τη διαδικασία του cross validation.

4ο βήμα (clustering – υλοποίηση σε R και WEKA): Θα επαναλάβετε τη διαδικασία, αυτή τη φορά για συσταδοποίηση (clustering). Αυτό σημαίνει ότι όλες οι διαδικασίες που περιγράφονται στο προηγούμενο βήμα για την ταξινόμηση με επίβλεψη (classification) θα επαναληφθούν ξανά, αφού πρώτα έχει παραλειφθεί εντελώς το χαρακτηριστικό-στόχος (#12), ώστε να πραγματοποιηθεί ταξινόμηση χωρίς επίβλεψη. Στο τελικό στάδιο, οι πραγματικές τιμές του χαρακτηριστικού-στόχου (#12) μπορεί να συσχετιστούν με τις συστάδες που σχηματίζονται. Με τον τρόπο αυτό θα ελεγχθούν η ακρίβεια και τα σφάλματα της διαδικασίας, συγκρίνοντας την «κατά πλειοψηφία» τιμή κατηγορίας κάθε συστάδας και τις πραγματικές τιμές του χαρακτηριστικού-στόχου για κάθε μέλος της (cluster labeling).

5ο βήμα (association rule mining – υλοποίηση σε WEKA): Θα επαναλάβετε τη διαδικασία, αυτή τη φορά για ανάλυση συσχετίσεων.

6ο βήμα (η επίδραση του δείγματος στην ποιότητα του αποτελέσματος): Εφαρμόστε τους παραπάνω αλγορίθμους (βήματα 3-5), αυτή τη φορά όχι στο πλήρες σύνολο δεδομένων αλλά σε δείγμα αυτού που έχει προέλθει μέσω κάποιας τεχνικής δειγματοληψίας και συγκρίνετε την ποιότητα του αποτελέσματος που προκύπτει σε σχέση με το αρχικό τρέξιμο στο πλήρες σύνολο. Η σύγκριση να είναι σε μορφή γραφικής παράστασης όπου στον άξονα -x θα υπάρχουν τα διαφορετικά ποσοστά του

δείγματος σε σχέση με το πλήρες σύνολο (10%, 20%, …., 100%) και στον άξονα -y η απόδοση του εκάστοτε αλγορίθμου.

Παραδοτέο εργασίας, προθεσμία και τόπος παράδοσης

Η εργασία χωρίζεται σε 2 παραδοτέα: το 1ο παραδοτέο περιλαμβάνει τα βήματα 1-2 ενώ το 2ο παραδοτέο περιλαμβάνει τα βήματα 3-~~5~~6. Κάθε παραδοτέο θα αποτελείται από την τεχνική αναφορά (report) με αναλυτική περιγραφή των προσεγγίσεων που ακολουθήσατε σε καθένα από τα βήματα (σχήμα ΒΔ, screenshots κλπ., αιτιολόγηση των αποφάσεων που πήρατε, ερμηνεία των αποτελεσμάτων που προέκυψαν, κοκ.)

Στο εξώφυλλο θα υπάρχουν τα στοιχεία:

Μάθημα: «Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (6ο εξ.)»

Ομάδα εργασίας: (ΑΜ, ονοματεπώνυμο)

Η εκτυπωμένη εργασία θα πρέπει να έχει παραδοθεί στη θυρίδα του κ. Θεοδωρίδη (έξω από το γραφείο της Γραμματείας στον 5ο όροφο) μέχρι **7/5/2018** (το 1ο παραδοτέο) και **30/6/2018** (το 2ο παραδοτέο). Σε περίπτωση που η εργασία παραδοθεί στην εξεταστική Σεπτεμβρίου (μέχρι την ημ/νία εξέτασης του μαθήματος) θα υπάρχει penalty -10%.

Επιπλέον της εκτυπωμένης εργασίας, στις ίδιες προθεσμίες θα αποστείλετε και σε ηλεκτρονική μορφή την εργασία, δηλ. την τεχνική αναφορά (pdf), μια παρουσίαση των επιμέρους βημάτων με κατάλληλο σχολιασμό (pdf) καθώς και όλα τα συνοδευτικά αρχεία (PostgreSQL, R, WEKA files), στους εργαστηριακούς βοηθούς του μαθήματος (*). Κάθε email θα έχει ως τίτλο "Εργασία DWDM 2018-2019 - <ΑΜ μελών ομάδας>" και θα περιέχει τα ζητούμενα σε ένα zip αρχείο.

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία σχετικά με την άσκηση μπορείτε να απευθύνεστε στον Δρ. Χάρη Γεωργίου (εργ. 205, hgeorgiou@unipi.gr) για θέματα που αφορούν το R Project, και τον κ. Γιάννη Κοντούλη (εργ. 205, ikontoulis@unipi.gr) για θέματα που αφορούν το WEKA.

Ζητήματα δεοντολογίας

Είναι προφανές ότι η βαθμολογία πρέπει να αντικατοπτρίζει το επίπεδο της γνώσης που αποκόμισε ο φοιτητής μέσα από το μάθημα και κατάφερε να μεταφέρει αυτή τη γνώση στην άσκηση. Για να εξασφαλιστεί όσο είναι δυνατό η παραπάνω αρχή, (α) σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται, (β) σε περίπτωση αμφιβολίας για το κατά πόσο η ομάδα που αναγράφεται ήταν εκείνη που ανέπτυξε την εργασία, ενδέχεται να της ζητηθεί να την παρουσιάσει για τυχόν διευκρινίσεις, ~~(γ) σε περίπτωση μεγάλης απόκλισης του βαθμού της εργασίας από το βαθμό της γραπτής εξέτασης, ο πρώτος δεν θα λαμβάνεται υπόψη για τον τελικό βαθμό του φοιτητή~~.

(*) Εργαστηριακοί βοηθοί: Χάρης Γεωργίου[1], Γιάννης Κοντούλης[2]

(αίθ. 205, [1]hgeorgiou@unipi.gr, [2]ikontoulis@unipi.gr)