Πανεπιστήμιο Ιωαννίνων,
Τμήμα Πληροφορικής

Εξόρυξη Δεδομένων – ΠΛΕ 059

Τρίτη Σειρά Ασκήσεων

Η προτεινόμενη ημερομηνία παράδοσης της άσκησης είναι **Τρίτη, 29 Μαΐου**, αλλά μπορείτε να παραδώσετε την άσκηση χωρίς ποινή μέχρι την Παρασκευή 8 Ιουνίου. Την Τετάρτη 30 Μαΐου θα δοθεί η τέταρτη και τελευταία άσκηση η οποία θα έχει ημερομηνία παράδοσης επίσης Παρασκευή 8 Ιουνίου, οπότε αν δεν έχετε παραδώσετε την τρίτη άσκηση μέχρι τότε θα πρέπει να τις κάνετε και τις δύο παράλληλα. Η παράδοση των ασκήσεων θα γίνει μέσω email.

Άσκηση 1

Στην άσκηση αυτή θα πειραματιστείτε με κατηγοριοποίηση στο WEKA. Θα πειραματιστείτε με τρία διαφορετικά datasets, το Iris dataset, το Mushroom dataset, και το Spambase dataset τα οποία μπορείτε να βρείτε στη σελίδα του μαθήματος. Μπορείτε να βρείτε λεπτομέρειες για το κάθε dataset στο UCI repository. Θα πειραματιστείτε με τρεις διαφορετικούς αλγορίθμους του WEKA: Naïve Bayes, Logistic Regression, και SimpleCart decision tree. Για κάθε αλγόριθμο και κάθε dataset, θα κάνετε 10-fold cross validation, και θα παραδώσετε το summary που δίνει στο τέλος το WEKA. Για το decision tree, θα παραδώσετε και το δέντρο το οποίο κατασκεύασε το WEKA.

Μαζί με τα αποτελέσματα θα παραδώσετε και μία αναφορά στην οποία θα συζητήσετε τα αποτελέσματα: πόσο δύσκολο είναι το πρόβλημα της κατηγοριοποίησης για κάθε dataset, και ποιος αλγόριθμος δουλεύει καλύτερα. Κοιτάζοντας τα παραγόμενα δέντρα σχολιάστε ποια attributes φαίνεται να έχουν μεγαλύτερη σημασία στην κατηγοριοποίηση .

Άσκηση 2

Στην άσκηση αυτή θα δημιουργήσετε ένα classifier για spam emails. Θα χρησιμοποιήσετε το SpanAssassin dataset το οποίο διατίθεται online (το link είναι στη σελίδα του μαθήματος). Για να εκπαιδεύσετε τον classifier, χρειάζεστε δεδομένα εκμάθησης τα οποία να ανήκουν στην θετική κλάση (spam) και στην αρνητική κλάση (non-spam), τα οποία θα πάρετε από το site του SpamAssasin, και θα τα συνδυάσετε σε ένα dataset (τουλάχιστον 1000 παραδείγματα). Το dataset αυτό θα το χωρίσετε σε training (80%) και testing (20%) υποσύνολα.

Κάθε παράδειγμα είναι το κείμενο από ένα email σε ξεχωριστό αρχείο. Από το κείμενο θα δημιουργήσετε features για τον classifier σας. Τα features μπορεί να είναι οτιδήποτε πιστεύετε ότι βοηθάει να διαχωριστούν τα spam emails από τα non-spam emails. Π.χ., όλες οι λέξεις που εμφανίζονται στα emails, ή οι συχνές λέξεις, ή οι λέξεις που εμφανίζονται στα subjects, ή το μέγεθος του μηνύματος, ή το domain του αποστολέα, κλπ, καθώς και συνδυασμός των παραπάνω . Ο classifier σας θα πρέπει να έχει τουλάχιστον 10 features (αν είναι οι λέξεις θα είναι πολύ παραπάνω). Το κάθε email γίνεται πλέον ένα διάνυσμα στο χώρο των features.

Αφού αποφασίσετε τα features θα τα εξάγετε από τα κείμενα και θα δημιουργήσετε έτσι την είσοδο για training και testing. Για τον αλγόριθμο κατηγοριοποίησης μπορείτε είτε να υλοποιήσετε το Naïve Bayes Classifier (καλύτερα να έχετε κατηγορικά features), είτε να χρησιμοποιήσετε το LibLinear πακέτο (link στη σελίδα του μαθήματος) το οποίο υλοποιεί τον Logistic Regression classifier και τον SVM classifier (θα πρέπει να μετατρέψετε την είσοδο σε αυτό που ζητάει το πακέτο). Τρέξτε μετά τον classifier που δημιουργήσατε πάνω στο test set.

Παραδώσετε μια αναφορά όπου θα περιγράφετε όλη την παραπάνω διαδικασία: Το πώς διαλέξατε τα δεδομένα, πως τα σπάσατε σε training και test σύνολα, τα features που δημιουργήσατε, τον αλγόριθμο που χρησιμοποιήσατε και τα αποτελέσματα του classifier στο test set. Για τα αποτελέσματα δώστε την ακρίβεια του classifier, καθώς και το confusion matrix. Παραδώσετε τον κώδικα σας καθώς και τα αρχεία εισόδου και εξόδου.

Bonus: Κατεβάσετε 50 spam, και 50 κανονικά emails από το προσωπικό σας email και τρέξτε τον αλγόριθμο σας πάνω σε αυτά τα emails. Αναφέρετε την ακρίβεια του αλγορίθμου και το confusion table. Παραδώσετε τα emails εισόδου.