

Τρίτη Σειρά Ασκήσεων

Αυτό είναι το πρώτο μέρος της τρίτης σειράς ασκήσεων. Η ημερομηνία παράδοσης της άσκησης είναι **14 Ιανουαρίου** στην αρχή του μαθήματος. Η παράδοση των ασκήσεων θα γίνει είτε μέσω του turn-in, είτε μέσω email. Λεπτομέρειες για το turn-in στη σελίδα του μαθήματος.

Ερώτηση 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 Υλικό του μαθήματος. Μπορείτε να βρείτε λεπτομέρειες για το κάθε dataset στο UCI repository. Θα πειραματιστείτε με τρεις διαφορετικούς αλγόριθμους 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 .

Ερώτηση 2

Στην ερώτηση αυτή θα πειραματιστείτε με την κατηγοριοποίηση των tips. Μέχρι τώρα έχετε δουλέψει με tips για εστιατόρια. Τώρα θα δουλέψουμε με tips από διάφορες κατηγορίες. Σας δίνεται ένα νέο αρχείο "category_tips.txt" το οποίο μπορείτε να κατεβάσετε από τη σελίδα Υλικό του μαθήματος. Το αρχείο περιέχει τα tips με τα οποία δουλέψατε μέχρι σήμερα, και επιπλέον tips για κέντρα νυχτερινής διασκέδασης, και για καταστήματα. Κάθε γραμμή στο αρχείο περιέχει δύο πεδία χωρισμένα με tab. Το πρώτο πεδίο είναι η κατηγορία της επιχείρησης: είναι είτε "food", είτε "nightlife", είτε "shops"; το δεύτερο πεδίο είναι το κείμενο του tip.

Ο στόχος σας είναι να δημιουργήσετε ένα classifier ο οποίος δοθέντος ενός tip θα προβλέπει την κατηγορία του. Σαν features θα χρησιμοποιήσετε τις λέξεις στα tips, με βάρος το tf-idf score (παρόμοια με την Άσκηση 2). Θα πρέπει να αφαιρέσετε τα stop words, και μπορείτε να κάνετε ότι άλλο feature selection θέλετε. Θα χρησιμοποιήσετε ένα Logistic Regression classifier, χρησιμοποιώντας μια υλοποίηση από ένα υπάρχον πακέτο λογισμικού. Μπορείτε να χρησιμοποιήσετε είτε το WEKA (χρησιμοποιώντας μια αραιή αναπαράσταση των δεδομένων), είτε το πακέτο LibLinear (υπάρχει link στο πακέτο LibLinear στη σελίδα Υλικό του μαθήματος).

Θα χρησιμοποιήσετε ένα τυχαίο δείγμα 80% των δεδομένων για εκμάθηση και το υπόλοιπο 20% των δεδομένων για τεστ (μπορείτε να χρησιμοποιήσετε τον κώδικα για δειγματοληψία που δημιουργήσατε στην πρώτη σειρά ασκήσεων). Ο Logistic Regression classifier επιστρέφει μια πιθανότητα για κάθε tip να ανήκει σε μια κατηγορία, και θα αναθέτετε το κάθε tip στην κατηγορία με την μεγαλύτερη πιθανότητα. Υπολογίστε την ακρίβεια του classifier. Θεωρήστε επίσης την περίπτωση όπου αναθέτουμε το tip στην κατηγορία με την μεγαλύτερη πιθανότητα μόνο αν η πιθανότητα είναι πάνω από ένα κατώφλι θ . Για κάθε κλάση υπολογίστε την precision-recall καμπύλη για τιμές του κατωφλίου $\theta \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Γράψτε μια σύντομη αναφορά για την υλοποίησή σας, τα αποτελέσματά σας και ένα σχόλιο για την ποιότητα των αποτελεσμάτων. Υπάρχει κάποιο κατώφλι θ που φαίνεται να δουλεύει καλά? Παραδώστε τον κώδικα σας και τα αποτελέσματά σας.

Ερώτηση 3

Για αυτή την ερώτηση θα πρέπει να υλοποιήσετε ένα Naïve Bayes classifier για κείμενο, όπως τον περιγράψαμε στην τάξη. Θα πρέπει να κάνετε δικιά σας υλοποίηση χρησιμοποιώντας όποια γλώσσα σας βολεύει. Θα χρησιμοποιήσετε τα ίδια δεδομένα όπως στην Ερώτηση 2, και θα κάνετε την ίδια επεξεργασία, αλλά χωρίς να χρησιμοποιήσετε το tf-idf score αυτή τη φορά.

Υπολογίστε την ακρίβεια του classifier και συγκρίνετε την με την ακρίβεια του Logistic Regression classifier στην Ερώτηση 2. Προσθέστε στην αναφορά για την Ερώτηση 2 τις παρατηρήσεις σας για την ποιότητα του Naïve Bayes classifier καθώς και την σύγκριση με τον Logistic Regression classifier. Παραδώστε τον κώδικα και την αναφορά.

Σημείωση: Για την επεξεργασία δεδομένων χωρισμένων με tab είναι χρήσιμες οι Unix εντολές που είναι στη σελίδα Υλικό του μαθήματος.