ΜΑΣ452 Γραμμικά Μοντέλα ΙΙ - ΜΑΣ653 Γενικευμένα Γραμμικά Μοντέλα

Διδάσκων: Σέργιος Αγαπίου

Εργασία 2, Μάρτιος 2019

Ημερομηνία Παράδοσης: 16/04/2019

Η εργασία είναι ομαδική και ισχύουν οι ομάδες της εργασίας 1 (για τους προπτυχιακούς). Σκοπός είναι από τη μία να πάρετε μια γεύση από διάφορα προβλήματα κατηγοριοποίησης και από την άλλη να κάνετε μια σύγκριση στην απόδοση των δύο μεθόδων κατηγοριοποίησης που διδαχθήκατε στο μάθημα (γραμμική παλινδρόμηση με δείκτρια εξαρτημένη μεταβλητή και λογιστική παλινδρόμηση), με μια πιο σύγχρονη μέθοδο (Random Trees και Random Forest). Σε κάθε ομάδα αντιστοιχεί ένα σύνολο δεδομένων που παρατίθεται πιο κάτω. Θα πρέπει να κατεβάσετε τα δεδομένα από τον σύνδεσμο και να διαβάσετε και να κατανοήσετε πιο είναι το πρόβλημα κατηγοριοποίησης. Να χωρίσετε τυχαία τα δεδομένα σε training και test, όπου τα test θα πρέπει να αποτελούν περίπου το 1/3 του συνόλου των δεδομένων. Τέλος, να κάνετε προσαρμογή των τριών μεθόδων στα training δεδομένα και να υπολογίσετε το ποσοστό λάθος κατηγοριοποιήσεων κάθε μεθόδου όταν κάνετε πρόβλεψη στα test δεδομένα.

Βασικές πληροφορίες για τη μέθοδο Random Forest καθώς και την εφαρμογή της στην R, υπάρχουν στα video στον πιο κάτω σύνδεσμο:
https://www.youtube.com/playlist?list=PL5-da3qGB5IB23TLuA8ZgVGC8hV8ZAdGh
(αν και καλό θα ήταν να δείτε όλα τα video για να πάρετε μια ιδέα από αυτή την πολύ καλή σε απόδοση μέθοδο, για την εφαρμογή της στα πλαίσια της εργασίας αρκεί να δείτε το τελευταίο video μόνο, τα πρώτα 3 λεπτά κυρίως, αγνοήστε το out of back error).

Για τις άλλες δύο μεθόδους μπορείτε να χρησιμοποιήσετε το εργαστήριο που θα γίνει στο μάθημα. Για τη λογιστική παλινδρόμηση μπορείτε επιπλέον να παρακολουθήσετε τα video 1,2 και 9 στον πιο κάτω σύνδεσμο:
https://www.youtube.com/playlist?list=PL5-da3qGB5IC4vaDba5ClatUmFppXLAhE

Να παραδόσετε αρχείο script της R, που περιέχει κώδικα που εκτελεί τα πιο πάνω και σχόλια με τα συμπεράσματά σας.

Σημείωση: τα δεδομένα πιο κάτω είναι λιγότερο επεξεργασμένα από τα δεδομένα με τα οποία δουλέψαμε μέχρι τώρα. Πιθανόν τα δεδομένα κάποιων ομάδων να χρειαστούν επεξεργασία, όπως πχ να αγνοήσετε κάποιες μεταβλητές ή να υπάρχουν τιμές κάποιων επεξηγηματικών μεταβλητών που λείπουν (σε τέτοια περίπτωση μπορείτε πχ να αγνοήσετε τις συγκεκριμένες παρατηρήσεις αν είναι λίγες, ή να θέσετε τις τιμές μιας μεταβλητής που λείπουν να είναι ίσες με το μέσο όρο των τιμών της μεταβλητής στις υπόλοιπες παρατηρήσεις).

Δεδομένα:

Καζίκα, Ραουνά, Χριστοδουλίδου: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Πέτρου, Χαραλάμπους, Σάββα: https://www.kaggle.com/uciml/mushroom-classification

Λοΐζου, Μιχαήλ, Σόλωνος, Αβραμοπούλου: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Νικολάου, Δαμιανού, Εγκωμίτη: https://archive.ics.uci.edu/ml/datasets/Spambase

Αυγουστή, Κουτσογιάννης: https://www.kaggle.com/primaryobjects/voicegender

Μιχαήλ, Μουγής, Χριστοφή, Παπαναστασίου: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Κοκκινόφτα, Μούγιος, Τράχηλου (θα παραδώσετε 3 ξεχωριστές εργασίες): https://www.kaggle.com/nsharan/h-1b-visa (να αγνοήσετε τις παρατηρήσεις που αντιστοιχούν σε withdrawn και certified-withdrawn case_status καθώς και τις επεξηγηματικές μεταβλητές lon και lat. Αν καθυστερεί πολύ να τρέξει στον υπολογιστή σας δουλέψετε πχ με το 1/10 των δεδομένων)