Εισαγωγή στη Μηχανική Μάθηση με χρήση του Weka

21/5/2015

Η παρούσα εργασία έχει ως σκοπό την εξοικείωσή σας με μεθόδους Μηχανικής Μάθησης, μέσω του διαγωνισμού "Titanic: Machine Learning From Disaster" που φιλοξενείται στην πλατφορμα kaggle. Πληροφορίες θα βρείτε εδώ http://www.kaggle.com/c/titanic-gettingStarted. Το θέμα του διαγωνισμού είναι το ναυάγιο του Τιτανικού και στόχος είναι να κατασκευαστεί ένα μοντέλο το οποίο θα προβλέπει με βάση τα χαρακτηριστικά του κάθε επιβάτη αν ο συγκεκριμένος επιβάτης επέζησε ή όχι.

Περιγραφή

Το σύνολο εκπαίδευσης του διαγωνισμού αποτελείται από 891 επιβάτες (παραδείγματα), καθένας από τους οποίους περιγράφεται από 10 ανεξάρτητες μεταβλητές όπως το όνομα, η ηλικία, το φύλο, το κόστος του εισιτηρίου κτλ. και μία εξαρτημένη μεταβλητή (μεταβλητή στόχος) η οποία είναι δυαδική (Survived={0,1}, 0=όχι και 1=ναι).

Για τη συγκεκριμένη εργασία, το σύνολο εκπαίδευσης του διαγωνισμού (training.csv) έχει μετασχηματιστεί στην μορφή arff και είναι έτοιμο για επεξεργασία από το Weka. Στα πλαίσια της εργασίας καλείστε να εισάγετε το σύνολο δεδομένων στο Weka και να απαντήσετε στα παρακάτω:

Ερωτήματα:

1) Προσδιορίστε ποιες από τις μεταβλητές είναι συνεχείς και ποιες κατηγορικές.
2) Εντοπίστε τη μεταβλητή με το μεγαλύτερο αριθμό ελλιπών τιμών.
3) Ποια είναι η ακρίβεια πρόβλεψης του αλγορίθμου 1-κοντινότερου γείτονα και του J48 (με default παραμέτρους) στο σύνολο εκπαίδευσης;
4) Ποιο είναι το ποσοστό των παραδειγμάτων που ταξινομείται σωστά σε κάθε περίπτωση του ερωτήματος 3;
5) Επαναλάβετε την παραπάνω διαδικασία χρησιμοποιώντας 5-fold cross-validation και

αναφέρετε τα αποτελέσματα. Σχολιάστε το αποτέλεσμα.

6) Χρησιμοποιείστε το φίλτρο ReplaceMissingValues για αντικατάσταση των ελλιπών τιμών και επαναλάβετε το ερώτημα 5. Βελτιώθηκε η ακρίβεια;
7) Δοκιμάστε και άλλους αλγορίθμους ταξινόμησης καταγράφοντας την ακρίβειά τους με 10-fold cross-validation. Σχολιάστε το αποτέλεσμα.
8) *Προαιρετικό ερώτημα*: Χρησιμοποιήστε το μοντέλο σας για πρόβλεψη της μεταβλητής survived στο test-set και να υποβάλετε τις προβλέψεις σας στο kaggle (μετά από κατάλληλη επεξεργασία ώστε η μορφή τους να είναι αποδεκτή).

Υποβολή εργασίας

Θα υποβάλετε στο σύστημα ΠΗΛΕΑΣ ένα συμπιεσμένο αρχείο με τίτλο το ΑΕΜ και το ονοματεπώνυμό σας (με λατινικούς χαρακτήρες), το οποίο θα περιλαμβάνει ένα PDF αρχείο με την τεκμηρίωση της εργασίας και (το username σας στο kaggle εάν προχωρήσατε στην υλοποίηση του 8 ζητήματος). Η τεκμηρίωση θα πρέπει να περιλαμβάνει όλα τα βήματα που ακολουθήσατε, απαντώντας αναλυτικά σε κάθε ένα από τα ζητήματα.