

ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΠΑΡΟΥΣΙΑΣΗ ΤΕΛΙΚΗΣ ΕΡΓΑΣΙΑΣ

ΠΡΟΕΠΙΣΚΟΠΗΣΗ ΔΕΔΟΜΕΝΩΝ

- Από τα ιστογράμματα των συχνοτήτων που αντιστοιχούν σε κάθε μεταβλητή βλέπουμε αν κάποια attributes μπορούν να αφαιρεθούν. Συγκεκριμένα, αφαιρούμε το attribute 38 το οποίο έχει σε όλα τα instances την τιμή μηδέν.
- Παρόμοια συμπεριφορά παρουσιάζουν και τα attributes 34 και 36 καθώς έχουν μόλις 15 και 1 αντίστοιχα instances διαφορετικά του μηδενός.
- Για να αποφασίσουμε αν θα παραμείνουν τα 34 και 36 θα εφαρμόσουμε μέθοδο παλινδρόμησης, συγκεκριμένα την Linear Regression. Η μέθοδος θα εφαρμοστεί σε ένα train set από το οποίο θα λείπουν τα 34,36,38 και σε ένα δεύτερο train set από το οποίο θα λείπει μόνο το 38 και στη συνέχεια θα εξετάσουμε τα σφάλματα για τα δύο sets.

Για το train set από το οποίο έχουμε αφαιρέσει μόνο το attribute 38 έχουμε από την Linear Regression τα εξής αποτελέσματα:

Correlation coefficient	0.743
Mean absolute error	0.2056
Root mean squared error	0.2847
Relative absolute error	56.8375 %
Root relative squared error	66.9218 %
Total Number of Instances	2528

Και από το train set από το οποίο έχουμε αφαιρέσει τα attributes 34,36, 38 έχουμε:

Correlation coefficient	0.7423
Mean absolute error	0.2059
Root mean squared error	0.285
Relative absolute error	56.918 %
Root relative squared error	66.9906 %
Total Number of Instances	2528

- Τα σφάλματα παλινδρόμησης διαφέρουν ελάχιστα, οπότε τελικά επιλέγουμε να διαγράψουμε τα attributes 34, 36 και 38.
- Αποθηκεύουμε το train set ως train_2.arff
- Ανοίγουμε το quizextended.arff , αφαιρούμε τα attributes 34,36, 38 και το αποθηκεύουμε ως quizextended_2.arff.
- Από διάφορους αλγόριθμους που δοκιμάσαμε για το train_2.set με supplied test set το quizextended.arff, επιλέγουμε τον μεταμαθησιακό αλγόριθμο Bagging και ορίζουμε ως εμφωλευμένο classifier τον Regression By Discretization.

- Στη συνέχεια ανοίγουμε το test set και αφαιρούμε και από αυτό τα attributes 34, 36 και 38. Αποθηκεύουμε το test set ως testextended2.arff.
- Στο τελευταίο βήμα της διαδικασίας μας, τρέχουμε το train set (το train_2.arff) με supplied test set το testextended2.arff.