Αναγνώριση Προτύπων & Νευρωνικά Δίκτυα
Εργασία

Ακ. Έτος 2021 – 2022

Σκοπός της εργασίας είναι η ανάπτυξη μεθόδων για την ταξινόμηση της ποιότητας λευκών κρασιών. Κάθε γεγονός (διαφορετικό κρασί) χαρακτηρίζεται από 11 βιοχημικές μεταβλητές και μία βαθμολογία (1–10) κατόπιν γευσιγνωσίας. Περισσότερες πληροφορίες θα βρείτε εδώ: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. Το αρχείο των δεδομένων (αρχείο csv) βρίσκεται εδώ: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv. Θεωρήστε ότι κάθε κρασί χαρακτηρίζεται ως «καλό» ή «κακό».

1. Επιλέξτε το κατώφλι βαθμολογίας που διαχωρίζει τις δύο κλάσεις (π.χ. μεγαλύτερη ή ίση του 5) ώστε να υπάρχει περίπου ίσος αριθμός γεγονότων (κρασιών) σε κάθε κλάση.
2. Να χωρίσετε με τυχαίο τρόπο τα δεδομένα σε σύνολο εκπαίδευσης (75%) και σύνολο αξιολόγησης (25%) με ίσο αριθμό γεγονότων σε κάθε κλάση.
3. Να απεικονίσετε για το σύνολο εκπαίδευσης τις κατανομές των 11 μεταβλητών (ένα διάγραμμα για κάθε μεταβλητή στο οποίο να φαίνεται ξεχωριστά η κατανομή για τα γεγονότα κάθε κλάσης).
4. Να απεικονίσετε για το σύνολο εκπαίδευσης τις συσχετίσεις όλων των μεταβλητών (ξεχωριστά για τις δύο κλάσεις).
5. Να κάνετε ανάλυση PCA στο σύνολο εκπαίδευσης, αφού πρώτα κάνετε τυποποίηση των μεταβλητών. Δείξτε την κατανομή των ιδιοτιμών του πίνακα διασποράς. Σχολιάστε το αποτέλεσμα.
6. Να υλοποιήσετε έναν γραμμικό ταξινομητή ελαχίστων τετραγώνων με κατάληλη βελτιστοποίηση των παραμέτρων του.
7. Να υλοποιήσετε ένα νευρωνικό δίκτυο (με ένα ή δύο κρυφά στρώματα και κατάλληλο αριθμό νευρώνων) που να διαχωρίζει τις δύο κλάσεις με κατάληλη βελτιστοποίηση των παραμέτρων του. Δείξτε την καμπύλη εκπαίδευσης και σχολιάστε το φαινόμενο της υπερεκπαίδευσης.
8. Να υλοποιήσετε ένα ενδυναμωμένο δέντρο απόφασης (boosted decision tree) που να διαχωρίζει τις δύο κλάσεις. Επιλέξτε κατάλληλο αριθμό δέντρων (μέγιστου βάθους 3).
9. Να συγκρίνετε την απόδοση των ταξινομητών με κατάλληλη επιλογή μετρικής και να σχολιάσετε τα αποτελέσματα.
10. Για κάθε ταξινομητή, να κατατάξετε τις μεταβλητές σύμφωνα με την επίδρασή τους σε αυτόν. Σχολιάστε τις ομοιότητες και τις διαφορές μεταξύ των διαφορετικών ταξινομητών.