3η Εργασία - Αναγνώριση Προτύπων

Διδάσκων: Επικ. Καθ. Παναγιώτης Πετραντωνάκης (ppetrant@ece.auth.gr)
Βοηθός διδασκαλίας: Υπ. Διδ. Στέφανος Παπαδόπουλος (stefpapad@iti.gr)

9 Δεκεμβρίου 2022

Σε αυτή την εργασία θα εργασθείτε πάνω στην εφαρμογή Δεντρων (Trees) για την ταξινόμηση δεδομένων καθώς και την επέκταση τους σε Random Forest ταξινομητές χρησιμοποιώντας την τεχνική Bootstrap.

Μέρος Α (0.5)

Εργάζεστε ως βοηθός έρευνας στο Εργαστήριο Ανθοκομίας του Τμήματος Γεωπονίας, ΑΠΘ, με ειδίκευση στην ανάλυση δεδομένων. Ένα τμήμα έρευνας στο εργαστήριο αφορά την αυτοματοποιημένη αναγνώριση διαφορετικών ειδών από ένα συγκεκριμένο φυτό, της Ίριδας. Τρία συγκεκριμένα είδη η Iris setosa, η Iris versicolor, και η Iris virginica παρουσιάζον διαφορές στο μήκος και πλάτος των σεπάλων και των πετάλων του ανθους τους. Από τη βιβλιοθήκη sklearn μπορείτε να κατεβάσετε μια βάση από 150 (50 για κάθε είδος) μετρήσεις τους μήκους και του πλάτους των σεπάλων και των πετάλων του άνθους κάθε είδους. Απομονώνοντας μόνο τα δύο πρώτα χαρακτηριστικά της βάσης, χρησιμοποιήστε τον έτοιμο αλγόριθμο DecisionTreeClassifier από τη βιβλιοθήκη sklearn και ταξινομήστε το 50% τυχαίων δειγμάτων του συνόλου αφού πρώτα έχετε εκπαιδεύσει τον αλγόριθμο με το υπόλοιπο 50%.

1. Τι ποσοστό σωστής ταξινόμησης λαμβάνετε; Ποιο βάθος δέντρου σας δίνει το καλύτερο ποσοστό;
2. Απεικονήστε τα όρια απόφασης του ταξινομητή για το καλύτερο αποτέλεσμα (*Βοήθεια: χρησιμοποιήστε τη συνάρτηση contourf*)

Μέρος Β (0.5)

Δημιουργήστε ένα Random Forest ταξινομητή 100 δέντρων με την τεχνική Bootstrap. Πιο συγκεκριμένα, το 50% των δειγμάτων που χρησιμοποιήσατε για εκπαίδευση στο προηγούμενο μέρος (σύνολο Α) χρησιμοποιήστε το τώρα για την δημιουργία 100 νέων συνόλων εκπαίδευσης ένα για κάθε δέντρο όπου κάθε φορά θα χρησιμοποιήται το $\gamma = 50\%$ του συνόλου Α. Το σύνολο που ταξινομήσατε στο προηγούμενο μέρος χρησιμοποιήστε το και εδώ για αξιολόγηση του αλγορίθμου. Όλα τα δέντρα να έχουν το ίδιο μέγιστο βάθος.

1. Τι ποσοστό σωστής ταξινόμησης λαμβάνετε; Ποιο βάθος δέντρου σας δίνει το καλύτερο ποσοστό;
2. Απεικονήστε τα όρια απόφασης του ταξινομητή για το καλύτερο αποτέλεσμα. Τι παρατηρείτε σε σχέση με τον απλό ταξινομητή του Μέρους Α;
3. Πώς πιστεύετε ότι επηρεάζει το ποσοστό γ την απόδοση του αλγορίθμου; Δώστε παραδείγματα.

Οδηγίες

1. Η Υλοποίηση της εργασίας θα γίνει σε Python. Επιλέξτε ένα notebook (π.χ., Jupyter, Collab) και γράψτε τον κώδικα όσο και τα σχόλιά σας.
2. Για την παράδοση θα ανεβάσετε ΕΝΑ αρχείο με όνομα: surname1_AEM_surname2_AEM_Assignment3.ipynb (σε περίπτωση που κάνετε την έργασία μόνοι, βαλτε μονο το επώνυμο και το ΑΕΜ σας. Αν είστε ομάδα δύο ατόμων, ΜΟΝΟ ένας κατεθέτει την εργασία) με όλες τις απαντήσεις. Στο αρχείο αυτό θα αναγράφονται τα στοιχεία σας (ονοματεπώνυμο, ΑΕΜ) σε ένα textbox στην αρχή. ΠΡΟΣΟΧΗ: Οι ομάδες πρέπει να είναι οι ίδιες με τις εργασίες 1 και 2!
3. Κάθε ένα από τα ερωτήματα θα απαντηθεί (κώδικας) σε ξεχωριστό κελί. Και ο κώδικας σε κάθε κελί θα συνοδεύεται από σύντομα σχόλια.

4. Τελική ημερομηνία υποβολής: Παρασκευή 13 Ιανουαρίου, 2023, 23:59.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!