Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων

Μεταπτυχιακό μάθημα: «Εξόρυξη Δεδομένων»

Εργασία 2

(Παράδοση: 7-6-2024)

Άσκηση 2.1. Ταξινόμηση με τη μέθοδο Bagging και με random forest

Για το dataset που χρησιμοποιήσατε στην 1η σειρά ασκήσεων να κατασκευάσετε συστήματα ταξινόμησης εφαρμόζοντας i) τη μέθοδο **Baging με decision-tree ως βασικό ταξινομητή**, καθώς και ii) τη μέθοδο **Random Forest.** Και στις δύο μεθόδους ο ελάχιστος αριθμός παραδειγμάτων σε φύλλο να τεθεί ίσος με 5. Να μελετήσετε πώς μεταβάλλεται η γενικευτική ικανότητα αυξάνοντας τον αριθμό των ταξινομητών που συμμετέχουν στο ensemble ως εξής: 25, 50, 75, 100. Για την μέτρηση της γενικευτικής ικανότητας να το out-of-bag-error. Να συγκρίνετε τις επιδόσεις των δύο μεθόδων σε σχέση με τους ταξινομητές που εξετάσατε στην 1η σειρά ασκήσεων στο ίδιο dataset.

Άσκηση 2.2. Ομαδοποίηση

Να χρησιμοποιήσετε τις μεθόδους ομαδοποίησης: 1) **kmeans** 2) **agglomerative clustering** και 3) **spectral clustering** με RBF kernel.

Στον κατάλογο "clustering" υπάρχουν διδιάστατα συνθετικά σύνολα παραδειγμάτων. Μπορείτε να κάνετε plot τα παραδείγματα για να διαπιστώσετε οπτικά τον πραγματικό αριθμό των ομάδων. Αφού βρείτε μια λύση ομαδοποίησης την παρουσιάζετε κάνοντας plot τα παραδείγματα βάζοντας ίδιο χρώμα για τα παραδείγματα της ίδιας ομάδας και διαφορετικά χρώματα για διαφορετικές ομάδες.

Για όλα σύνολα παραδειγμάτων **να χρησιμοποιήσετε τον πραγματικό αριθμό ομάδων**, όπως τον διαπιστώνετε από την οπτικοποίηση των δεδομένων. Να τυπώσετε την καλύτερη δυνατή λύση ομαδοποίησης που θα βρείτε για καθεμιά από τις παρακάτω 6 περιπτώσεις:

i) k-means
ii) agglomerative clustering (single link, average link),
iii) spectral clustering (RBF kernel για διάφορες τιμές του sigma: ενδεικτικά: 0.1, 0.5, 1).

Για το σύνολο παραδειγμάτων 'gaussian_rings' να βρείτε κάποια τιμή του sigma για την οποία η μέθοδος spectral να δίνει τη σωστή λύση ομαδοποίησης.

Να διατυπώσετε παρατηρήσεις σχετικά με την συγκριτική επίδοση των μεθόδων στα σύνολα παραδειγμάτων.

Στη συνέχεια για **τα σύνολα παραδειγμάτων που δεν περιέχουν δακτυλίους**, να δοκιμάσετε να εκτιμήσετε τον πραγματικό αριθμό ομάδων χρησιμοποιώντας ως αλγόριθμο ομαδοποίησης τον kmeans και ως κριτήριο αξιολόγησης το silhouette. Να διατυπώσετε παρατηρήσεις επί των αποτελεσμάτων που προκύπτουν.