Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων

Μεταπτυχιακό μάθημα: «Εξόρυξη Δεδομένων»

Εργασία 2

(Παράδοση 6-6-2018)

Άσκηση 1. Ταξινόμηση με τη μέθοδο Bagging και με random forest

Για το dataset [*.mat] που χρησιμοποιήσατε στην $1^{η}$ σειρά ασκήσεων να κατασκευάσετε συστήματα ταξινόμησης εφαρμόζοντας τη μέθοδο **Bagging** (matlab), καθώς και τη μέθοδο **Random Forest** (matlab, Min_Leaf=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 για τη μέθοδο Bagging να χρησιμοποιήσετε cross-validation, ενώ για τη μέθοδο RandomForest το out-of-bag-error. Να συγκρίνετε τις επιδόσεις των δύο μεθόδων σε σχέση με τους ταξινομητές που εξετάσατε στην $1^{η}$ σειρά ασκήσεων στο ίδιο dataset.

Άσκηση 2. Ομαδοποίηση

Να χρησιμοποιήσετε τις μεθόδους ομαδοποίησης: 1) **kmeans** (matlab toolbox) 2) **agglomerative clustering** (matlab toolbox) και 3) **spectral clustering** με RBF kernel (αρχείο "spectral.m" στον κατάλογο "clustering").

Επίσης στον κατάλογο "clustering" υπάρχουν:
α) δισδιάστατα σύνολα παραδειγμάτων
β) η συνάρτηση plot_max10_clusters για να κάνετε plot τις λύσεις ομαδοποίησης που βρίσκετε (για δισδιάστατα σύνολα παραδειγμάτων και μέχρι 10 clusters).

Για όλα σύνολα παραδειγμάτων **να χρησιμοποιήσετε τον πραγματικό αριθμό ομάδων**. 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 για τιμές του sigma: 0.1, 0.5, 1.

Για το σύνολο παραδειγμάτων 'gaussian_rings' να βρείτε κάποια τιμή του sigma στο διάστημα [0.1, 0.4] 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.

Ο αλγόριθμος spectral δίνει πάντα την ίδια λύση; Να αιτιολογήσετε την απάντησή σας.

Στη συνέχεια για **τα σύνολα παραδειγμάτων που δεν περιέχουν δακτυλίους**, να δοκιμάσετε να εκτιμήσετε τον πραγματικό αριθμό ομάδων χρησιμοποιώντας τη συνάρτηση evalclusters της Matlab με αλγόριθμο ομαδοποίησης τον kmeans και κριτήριο αξιολόγησης το silhouette. Να διατυπώσετε παρατηρήσεις επί των αποτελεσμάτων που βρίσκετε.