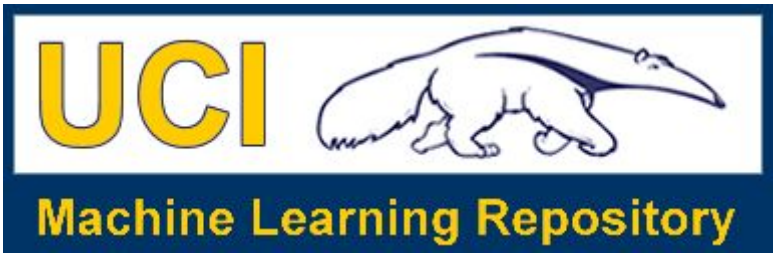
Άσκηση 1. Επιβλεπόμενη Μάθηση: Ταξινόμηση. Μελέτη datasets του UCI Machine Learning Repository



Κάθε ομάδα του εργαστηρίου των Νευρωνικών θα μελετήσει ως προς την ταξινόμηση 2 datasets από το UCI Machine Learning repository. Το ένα dataset είναι μικρό (Small) με λιγότερα από 1000 δείγματα και το άλλο μεγάλο (Big) με περισσότερα από 1000 δείγματα. Υπάρχουν 9 S και 9 B datasets οπότε καμία από τις 80 ομάδες του εργαστηρίου δεν έχει τον ίδιο συνδυασμό (S,B) datasets με άλλη.

Μπορείτε να βρείτε τα (S,B) που έχουν ανατεθεί στην ομάδα σας στον πίνακα Ομάδες - UCI Datasets. Για να δείτε ποια datasets του UCI αντιστοιχούν στους κωδικούς σας καθώς και με ποια αρχεία δεδομένων πρέπει να δουλέψετε, συμβουλευτείτε τον πίνακα UCI classification datasets.

Η κάθε ομάδα θα παραδώσει στο mycourses δύο jupyter notebooks (ipynb) σε ένα zip file. Θα συνδυάζετε κώδικα για την υλοποίηση και markdown για τις απαντήσεις, εξηγήσεις και εκτιμήσεις σας. Χρησιμοποιήστε το markdown formatting για να οργανώσετε το notebook σε sections.

Στο πρώτο (notebook 1) θα εκπαιδεύσετε και βελτιστοποιήσετε τους dummy classifiers και ένα MultiLayer Perceptron (MLP) στο μικρό dataset χωρίς τη χρήση pipelines και gridsearchcv. Θα χρησιμοποιήστε 20% για test set και σχήμα 10-fold για cross-validation.

Στο δεύτερο (notebook 2) θα εκπαιδεύσετε και βελτιστοποιήσετε τους ταξινομητές dummy classifiers, Gaussian Naive Bayes, kNN, MLP στο μεγάλο dataset με χρήση pipelines και gridsearchcv. Θα χρησιμοποιήστε 30% για test set και σχήμα 5-fold για cross-validation.

Με εξαίρεση τους dummy classifiers, για τους υπόλοιπους ταξινομητές πρέπει για τον καθένα ξεχωριστά σε κάθε dataset να βρείτε τη βέλτιστη αρχιτεκτονική μετασχηματιστών (στάδια προ-επεξεργασίας) και τις βέλτιστες υπερ-παραμέτρους (τόσο των μετασχηματιστών όσο και του ταξινομητή) μέσω cross-validation.

Θα χρησιμοποιήσετε 2 διαφορετικά metrics απόδοσης στο cross-validation για την επιλογή του μοντέλου, f1_macro και f1_weighted.

Παραδοτέα

Το κάθε notebook θα αποτελείται από τα ακόλουθα sections:

Στοιχεία ομάδας

Αριθμός ομάδας, ονοματεπώνυμά και ΑΜ σας

Εισαγωγή του dataset

Σύντομη παρουσίαση του dataset (τι περιγράφει).

Αριθμός δειγμάτων και χαρακτηριστικών, είδος χαρακτηριστικών. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
Υπάρχουν επικεφαλίδες; Αρίθμηση γραμμών;
Ποιες είναι οι ετικέτες των κλάσεων και σε ποια κολόνα βρίσκονται;

Εισαγωγή του dataset στο notebook. Χρειάστηκε να κάνετε μετατροπές στα αρχεία text?

Υπολογισμός και εκτύπωση εντός του notebook:
Υπάρχουν απουσιάζουσες τιμές; Πόσα είναι τα δείγματα με απουσιάζουσες τιμές και ποιο το ποσοστό τους επί του συνόλου; Παρόμοια, αριθμός κλάσεων και ποσοστά δειγμάτων τους επί του συνόλου.

Αν θεωρήσουμε ότι ένα dataset είναι μη ισορροπημένο αν μια οποιαδήποτε κλάση είναι 1.5 φορά πιο συχνή από κάποια άλλη (60%-40% σε binary datasets) εκτιμήστε την ισορροπία του dataset.

Διαχωρίστε σε train και test set. Εάν υπάρχουν απουσιάζουσες τιμές και μη διατεταγμένα χαρακτηριστικά διαχειριστείτε τα και αιτιολογήστε τις επιλογές σας.

Baseline classification

Εκπαιδεύστε στο train τους classifiers με default τιμές (απλή αρχικοποίηση). Κάντε εκτίμηση στο test set (μαζί με τους dummy) και τυπώστε για κάθε estimator: confusion matrix, f1-weighted average και f1-macro average.

Για κάθε averaged metric, εκτυπώστε bar plot συγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.

Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης.

Βελτιστοποίηση ταξινομητών

Για κάθε ταξινομητή βελτιστοποιήστε την απόδοσή του στο training set μέσω της διαδικασίας προεπεξεργασίας και εύρεσης βέλτιστων υπερπαραμέτρων (δεν έχουν όλοι οι ταξινομητές υπερπαραμέτρους). Κάντε εκτίμηση στο test set (μαζί με τους dummy) και τυπώστε για κάθε estimator: confusion matrix, f1-weighted average και f1-macro average. Για το τελικό (με τις βέλτιστες παραμέτρους) fit στο train set και για το predict στο test set εκτυπώστε πίνακες με τους χρόνους εκτέλεσης.

Για κάθε averaged metric, εκτυπώστε bar plot σύγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.

Τυπώστε πίνακα με τη μεταβολή της επίδοσης των ταξινομητών πριν και μετά τη βελτιστοποίησή τους.

Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης, τη μεταβολή της απόδοσης και τους χρόνους εκτέλεσης.

Βασικές υπερπαράμετροι προς βελτιστοποίηση

kNN: n_neighbors, metric, weights

MLP: hidden_layer_sizes (χρησιμοποιήστε μόνο ένα επίπεδο κρυμμένων νευρώνων), activation, solver, max_iter, learning_rate, alpha

ΠΑΡΑΔΟΣΗ στο mycourses Παρασκευή 15/12/2017 ΑΥΣΤΗΡΑ

Tips

Για απορίες, πρώτα συμβουλευθείτε το FAQ το οποίο περιέχει λήμματα ειδικά για την Άσκηση 1. Αν εξακολουθείτε να έχετε απορία σε κάποιο θέμα μπορείτε να απευθύνεστε στο nnlab@islab.ntua.gr (παρακαλούμε όχι ατομικά email). Το FAQ θα ανανεώνεται με απαντήσεις στις ερωτήσεις που θα στέλνετε στο nnlab..

Μελετήστε καλά την περιγραφή του dataset στη σελίδα του στο UCI και δείτε προσεκτικά τα txt αρχεία των δεδομένων.

Ορίστε συναρτήσεις για πράξεις που κάνετε συχνά.