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

Ημερομηνία εκφώνησης άσκησης: 02/11/20

Περιγραφή της άσκησης

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 12 S και 12 B datasets προς μελέτη και καμία από τις 70+ 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 (οι βάρδιες βρίσκονται σε περισσότερα φύλλα). 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 eclass του μαθήματος **ένα jupyter notebook** (ipynb) **και ένα αρχείο .py** με τον κώδικα Python του notebook **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.

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

Παραδοτέα

Το notebook θα αποτελείται από τα ακόλουθα sections:

Στοιχεία ομάδας

Αριθμός ομάδας, ονοματεπώνυμά και ΑΜ σας. Παρακαλούμε μην το ξεχάσετε.

Μικρό dataset (S)

Βασικές πληροφορίες

1. Σύντομη παρουσίαση του dataset (τι περιγράφει).
2. Αριθμός δειγμάτων και χαρακτηριστικών, είδος χαρακτηριστικών. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
3. Υπάρχουν επικεφαλίδες; Αρίθμηση γραμμών;
4. Ποιες είναι οι ετικέτες των κλάσεων και σε ποια κολόνα βρίσκονται;
5. Χρειάστηκε να κάνετε μετατροπές στα αρχεία text και ποιες?
6. Υπάρχουν απουσιάζουσες τιμές; Πόσα είναι τα δείγματα με απουσιάζουσες τιμές και ποιο το ποσοστό τους επί του συνόλου;
7. Ποιος είναι ο αριθμός των κλάσεων και τα 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.
8. Διαχωρίστε σε train και test set. Εάν υπάρχουν απουσιάζουσες τιμές και μη διατεταγμένα χαρακτηριστικά διαχειριστείτε τα και αιτιολογήστε τις επιλογές σας.

Ταξινόμηση

- Οι ταξινομητές που θα εξετάσετε στο μικρό dataset είναι οι: dummy, Gaussian Naive Bayes, kNN.

- Θα χρησιμοποιήστε 20% για test set και σχήμα 10-fold για cross-validation.
- Θα χρησιμοποιήσετε 2 διαφορετικές μετρικές απόδοσης στο cross-validation για την επιλογή του μοντέλου: α) f1_micro και β) f1_macro.

Baseline classification

1. Διαχειριστείτε τυχόν απουσιάζουσες τιμές. 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-micro average και f1-macro average.
2. Για κάθε averaged metric, εκτυπώστε bar plot συγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.
3. Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης.

Βελτιστοποίηση ταξινομητών

1. 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-micro average και f1-macro average.
2. Για το τελικό fit του κάθε ταξινομητή στο σύνολο του training set και για το predict στο test set εκτυπώστε πίνακες με τους χρόνους εκτέλεσης.
3. Για κάθε averaged metric, εκτυπώστε bar plot σύγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.
4. Τυπώστε πίνακα με τη μεταβολή της επίδοσης των ταξινομητών πριν και μετά τη βελτιστοποίησή τους.
5. Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης, τη μεταβολή της απόδοσης και τους χρόνους εκτέλεσης.

Υπερπαράμετροι προς βελτιστοποίηση

kNN: n_neighbors.

Μεγάλο dataset (B)

Βασικές πληροφορίες

1. Σύντομη παρουσίαση του dataset (τι περιγράφει).
2. Αριθμός δειγμάτων και χαρακτηριστικών, είδος χαρακτηριστικών. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
3. Υπάρχουν επικεφαλίδες; Αρίθμηση γραμμών;
4. Ποιες είναι οι ετικέτες των κλάσεων και σε ποια κολόνα βρίσκονται;
5. Χρειάστηκε να κάνετε μετατροπές στα αρχεία text και ποιες?
6. Υπάρχουν απουσιάζουσες τιμές; Πόσα είναι τα δείγματα με απουσιάζουσες τιμές και ποιο το ποσοστό τους επί του συνόλου;
7. Ποιος είναι ο αριθμός των κλάσεων και τα 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.

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

Ταξινόμηση

- Οι ταξινομητές που θα εξετάσετε στο μεγάλο dataset είναι οι: dummy, Gaussian Naive Bayes, kNN, Multi-Layer Perceptron (MLP), Support Vector Machines (SVM)
- Θα χρησιμοποιήστε 30% για test set και σχήμα 5-fold για cross-validation.
- Θα χρησιμοποιήσετε 2 διαφορετικές μετρικές απόδοσης στο cross-validation για την επιλογή του μοντέλου: α) f1_micro και β) f1_macro.

Baseline classification

1. Διαχειριστείτε τυχόν απουσιάζουσες τιμές. 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-micro average και f1-macro average.
2. Για κάθε averaged metric, εκτυπώστε bar plot συγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.
3. Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης.

Βελτιστοποίηση ταξινομητών

1. 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-micro average και f1-macro average.
2. Για το τελικό fit του κάθε ταξινομητή στο σύνολο του training set και για το predict στο test set εκτυπώστε πίνακες με τους χρόνους εκτέλεσης.
3. Για κάθε averaged metric, εκτυπώστε bar plot σύγκρισης με τις τιμές του συγκεκριμένου f1 για όλους τους classifiers.
4. Τυπώστε πίνακα με τη μεταβολή της επίδοσης των ταξινομητών πριν και μετά τη βελτιστοποίησή τους.
5. Σχολιάστε τα αποτελέσματα των plots και των τιμών precision, recall, f1 των πινάκων σύγχυσης, τη μεταβολή της απόδοσης και τους χρόνους εκτέλεσης.

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.
SVM: Για linear kernel θα χρησιμοποιήσετε τη LinearSVC με βασικές υπερπαραμέτρους για βελτιστοποίηση: loss, tol, C. Για πυρήνες poly και rbf θα χρησιμοποιήσετε την SVC με βασικές υπερπαραμέτρους για βελτιστοποίηση C, degree (για πυρήνα poly), gamma, tol.

Μπορείτε να πειραματιστείτε και με τις υπόλοιπες υπερπαραμέτρους των ταξινομητών.

Multiclass datasets: οι kNN και MLP είναι εγγενώς multiclass classifiers, ενώ τα SVM δεν είναι. Ωστόσο, το SVC υλοποιεί μια στρατηγική multiclass one-vs-one ενώ και το LinearSVC έχει μια παραλλαγή multiclass SVM.

Μη ισορροπημένα datasets: σημειώστε ότι στα SVM αντί να αντιμετωπίσετε το πρόβλημα στο στάδιο της προεπεξεργασίας μπορείτε να χρησιμοποιήσετε την παράμετρο class_weight κατά την εκπαίδευση (χωρίς να γνωρίζουμε εξαρχής ποια από τις δύο προσεγγίσεις δίνει τα καλύτερα αποτελέσματα).

Συμβουλές

Ξεκινήστε με μια επανάληψη όλων των notebooks του μαθήματος. Όλα τα ζητούμενα της άσκησης καλύπτονται από όσα έχουμε κάνει στα notebooks αυτά.

Μελετάτε πάντοτε και πλήρως το documentation των συναρτήσεων που χρησιμοποιείτε. Στα notebooks του μαθήματος καλύπτουμε ένα μικρό μέρος των δυνατοτήτων της βιβλιοθήκης.

Μελετήστε καλά την περιγραφή του dataset στη σελίδα του στο UCI και δείτε προσεκτικά τα txt αρχεία των δεδομένων (ανοίξτε τα σε έναν text editor όπως ο Notepad++ για Windows).

Υπάρχουν ευκολότερα και δυσκολότερα datasets, δεν είναι αναγκαστικό να έχετε πάντα υψηλές τιμές του f1 σε απόλυτους αριθμούς. Για παράδειγμα σε ένα dataset μπορεί ένα f1 0.9 από ένα ταξινομητή να θεωρείται χαμηλό και σε ένα άλλο ένα f1 0.6 από κάποιον ταξινομητή να θεωρείται κορυφαίο. Το ζητούμενο είναι κάθε φορά να παίρνετε το βέλτιστο από τον κάθε ταξινομητή για το dataset που μελετάτε. Μπορείτε να συμβουλευτείτε τη δημοσιευμένη έρευνα για τις αποδόσεις των ταξινομητών στο dataset που δουλεύετε (ψάξτε το Google Scholar με το όνομα του dataset σας).

Συμβουλευτείτε **απαραίτητα το FAQ** για την αρχιτεκτονική (pipeline) του εκτιμητή, το grid search και διάφορα θέματα επεξεργασίας των datasets και ερωτήσεις dataset-specific.

Ορίστε συναρτήσεις για πράξεις που κάνετε συχνά και επαναπαραγοντοποιήστε όπου χρειάζεται. Θα αξιολογηθεί και η ποιότητα του κώδικα.

Θα αξιολογηθούν επίσης η οργάνωση του notebook, η χρήση του markdown και η παρουσίαση (plots, πίνακες κλπ) των αποτελεσμάτων.

Πριν παραδώσετε βεβαιωθείτε ότι έχετε συμπεριλάβει όλα τα παραδοτέα στο notebook. Εάν απουσιάζει κάποιο παραδοτέο χάνετε αυτομάτως τα μόρια που του αντιστοιχούν που είναι κρίμα.

Δεν θέλουμε τα data files εντός του zip. Θα πρέπει όμως στο notebook **να έχετε τρέξει όλα τα κελιά και να είναι ορατή η έξοδος τους** όταν τα κάνετε download για να τα ανεβάσετε στο mycourses. Μην ξεχάσετε το αρχείο κώδικα Python (.py) μέσα στο zip ώστε να μπορούμε να συγκρίνουμε αυτόματα τον κώδικα όλων των ομάδων.

Τα filenames των αρχείων .ipynb, .py και του zip πρέπει να περιλαμβάνουν τον αριθμό της ομάδας σας.

Για απορίες, πρώτα **συμβουλευτείτε το FAQ** το οποίο περιέχει λήμματα ειδικά για την Άσκηση 1. Αν εξακολουθείτε να έχετε απορία σε κάποιο θέμα μπορείτε να απευθύνεστε στην περιοχή συζητήσεων του μαθήματος στο Eclass (Εξαμηνιαία Εργασία - Εργαστήριο). Το FAQ θα ανανεώνεται με απαντήσεις στις ερωτήσεις που θα θέσετε στην περιοχή συζητήσεων. Εάν αφήσετε την άσκηση για την τελευταία στιγμή και σας προκύψουν απορίες, πιθανότατα δεν θα προλάβουμε να τις απαντήσουμε οπότε είναι καλύτερο να αρχίσετε να την δουλεύετε νωρίς.

Τέλος, προσοχή στους χρόνους εκτέλεσης. Ένα μεγάλο gridsearch μαζί με ένα μη-παραμετρικό ταξινομητή, μπορεί να χρειαστεί πολύ χρόνο γιια να ολοκληρωθεί.

Ημερομηνία παράδοσης

Τετάρτη 9 Δεκεμβρίου 2020 στις 11.59 το βράδυ (μεσάνυχτα).

Το σύστημα υποβολής δεν κλείνει αλλά εκπρόθεσμη υποβολή σημαίνει μείωση του βαθμού αναλογική με το χρόνο εκτός προθεσμίας (plz μην μας ρωτάτε αναλογικά με τι συντελεστή κλπ, προσπαθήστε να είστε συνεπείς στην προθεσμία της άσκησης).

Τέλος, παρακαλούμε *θερμά* όχι αποστολές απαντήσεων σε email.