Εργασία 1. Επιβλεπόμενη Μάθηση: Ταξινόμηση

Μελέτη datasets του αποθετηρίου UCI και της πλατφόρμας Kaggle

Εισαγωγή

Στόχος της εργασίας είναι η μελέτη και βελτιστοποίηση ταξινομητών σε σύνολα δεδομένων. Κάθε ομάδα του εργαστηρίου θα μελετήσει δύο datasets, ένα από το αποθετήριο UCI και ένα από την πλατφόρμα Kaggle.

Η εκπαίδευση και βελτιστοποίηση των ταξινομητών στο UCI dataset θα γίνει αποκλειστικά με τις συναρτήσεις του skicit-learn ενώ στο Kaggle dataset θα πρέπει να χρησιμοποιήσετε μια βιβλιοθήκη βελτιστοποίησης (όποια θέλετε).

Υπάρχουν 14 διαφορετικά UCI datasets (U01-U14) και 14 διαφορετικά Kaggle datasets (K01-K14). Σε κάθε ομάδα αντιστοιχεί ένας μοναδικός συνδυασμός U-K datasets. Μπορείτε να βρείτε ποιοι κωδικοί datasets αντιστοιχούν στον αριθμό που έχει η ομάδα σας στο helios στον πίνακα Teams - Datasets.

Για να βρείτε ποιο dataset αντιστοιχεί στον κωδικό UCI συμβουλευτείτε τον πίνακα UCI Datasets.

Για να βρείτε ποιο dataset αντιστοιχεί στον κωδικό Kaggle συμβουλευτείτε τον πίνακα Kaggle Datasets.

Μέρος 1. UCI dataset (40%)

Εισαγωγή και επισκόπηση

Εισάγετε το dataset από το αρχείο text στο notebook σας. Στη συνέχεια, σε κελιά markdown γράψτε τις βασικές πληροφορίες ως προς αυτό:

- Σύντομη παρουσίαση του dataset δηλαδή ποιο είναι το πρόβλημα που περιγράφει.
- Χρειάστηκε να κάνετε μετατροπές στα αρχεία plain text για την εισαγωγή του; αν ναι, ποιες είναι αυτές;
- Δώστε το πλήθος δειγμάτων και χαρακτηριστικών, και το είδος όλων των χαρακτηριστικών. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
- Υπάρχουν επικεφαλίδες; Αρίθμηση γραμμών;
- Ποιες είναι οι ετικέτες των κλάσεων και σε ποια κολόνα βρίσκονται;
- Υπάρχουν απουσιάζουσες τιμές; Πόσα είναι τα δείγματα με απουσιάζουσες τιμές και ποιο το ποσοστό τους επί του συνόλου;
- Ποιος είναι το πλήθος 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 αν το dataset είναι ισορροπημένο ή όχι.

Προετοιμασία

- Διαχωρίστε το σύνολο δεδομένων σε σύνολο εκπαίδευσης (train set) και σύνολο (test set) με 30% των δειγμάτων στο test set. Αν το datasets σας είναι από την περιγραφή του ήδη χωρισμένο σε train και test, εφόσον τα ποσοστά είναι κοντά στο 70-30 μπορείτε να τα χρησιμοποιήσετε ως έχουν. Αν δεν είναι, ενοποιήστε train και test και προχωρήστε στο ίδιο split.
- Αν υπάρχουν απουσιάζουσες τιμές διαχειριστείτε τες και αιτιολογήστε.
- Διαχειριστείτε τυχόν κατηγορικά ή/και μη διατεταγμένα χαρακτηριστικά και αιτιολογήστε.

Ταξινόμηση

Ταξινομητές

Στο UCI θα μελετήσουμε τους ταξινομητές

- dummy,
- Gaussian Naive Bayes (GNB),
- KNeirestNeighbors (kNN), και
- Logistic Regression (LR).

Μετρικές

Η βελτιστοποίηση και η παρουσίαση των αποτελεσμάτων θα πρέπει κάθε φορά να γίνει ξεχωριστά για δύο μετρικές:

- ορθότητα (accuracy), και
- F1-score (macro σε προβλήματα multiclass).

Σχήμα διασταυρούμενης επικύρωσης

Για όλα τα πειράματα θα χρησιμοποιήσετε 10-fold cross-validation.

Επίδοση out-of-the-box

Αρχικά θα δούμε πως συμπεριφέρονται οι ταξινομητές χωρίς καμία βελτιστοποίηση (out-of-the-box) και με όλες τις παραμέτρους σε default τιμές.

Εκπαιδεύστε όλους τους εκτιμητές με ένα απλό fit σε ολόκληρο το training set και υπολογίστε την επίδοσή τους στο test set για τις δύο μετρικές.

Παρουσιάστε συνοπτικά και συγκριτικά την επιδοσή τους:

1. σε πίνακα markdown, και
2. σε bar plot σύγκρισης,

και σχολιάστε την επίδοσή τους.

Βελτιστοποίηση

Για όλους τους ταξινομητές βελτιστοποιήστε την επίδοσή τους μέσω των διαδικασιών

- προεπεξεργασίας,
- ορισμού pipelines, και
- εύρεσης βέλτιστων υπερμαραμέτρων με αναζήτηση πλέγματος με διασταυρούμενη επικύρωση

Για το καλύτερο μοντέλο κάθε ταξινομητή, εκπαιδεύστε το στο σύνολο του train set και εκτιμήστε την επίδοσή του στο test set. Επιπρόσθετα, για τα βέλτιστα μοντέλα, καταγράψτε τους χρόνους train και test.

Αποτελέσματα και συμπεράσματα

Παρουσιάστε συνοπτικά και συγκριτικά την επιδοσή τους:

1. σε πίνακα markdown όπου εκτός των δύο μετρικών θα περιλαμβάνεται η μεταβολή τους σε σχέση με το out-of-the-box καθώς και οι δύο χρόνοι, και
2. σε bar plot σύγκρισης που θα περιλαμβάνει και την μεταβολή (χωρίς τους χρόνους).

Σχολιάστε συνολικά την επίδοσή τους καθώς και τη μεταβολή από την επίδοση out-of-the-box.

Για τον καλύτερο και τον χειρότερο ταξινομητή (εξαιρουμένων των dummy) ως προς την ορθότητα εκτυπώστε τους πίνακες σύγχυσης με γραφικό τρόπο (πχ seaborn) και σχολιάστε.

Ποιον ταξινομητή προτείνετε τελικά για το συγκεκριμένο πρόβλημα και γιατί; Μπορείτε να δώσετε κάποια ερμηνεία για την καλή επίδοσή του στο πρόβλημα, απόλυτα ή/και σε σχέση με τους υπόλοιπους (εκτός των dummy);

Μέρος 2. Kaggle dataset (60%)

Συνολικός στόχος

Στο δεύτερο μέρος της εργασίας καλείστε να μελετήσετε ένα dataset από το Kaggle. Σε γενικές γραμμές, τα επιλεγμένα dataset του Kaggle είναι μεγαλύτερα ως πολύ μεγαλύτερα από αυτά του UCI.

Θα διαπιστώσετε ότι σας δίνεται μεγαλύτερη ελευθερία επιλογών για το πως θα βρείτε τα βέλτιστα μοντέλα για το dataset σας. Ο συνολικός στόχος σε αυτό το μέρος είναι τριπλός:

- Να βελτιστοποιήσετε τους ταξινομητές για την επίτευξη της καλύτερης δυνατής επίδοσης με την σωστή μεθοδολογία σε όλες τις επιλογές.
- Να μπορείτε να περιγράψετε με σύντομο και ουσιαστικό τρόπο τις επιλογές σας κατά τον πειραματισμό.
- Να παρουσιάσετε με πλήρη και εύγλωττο τρόπο τα συμπεράσματά σας.

Εισαγωγή του dataset

Kaggle

Εφόσον πρόκειτα για datasets του Kaggle το πιο απλό είναι να δουλέψετε στο Kaggle και είναι ο προτεινόμενος τρόπος.

Αρκεί από το "Code" να δημιουργήσετε ένα καινούριο notebook, να κάνετε "Add data", να αναζητήσετε το dataset με το όνομά του, και να το κάνετε "Add".

Στη συνέχεια απλώς τρέξετε το πρώτο έτοιμο κελί και θα σας εμφανίσει το path για όλα τα αρχεία του dataset.

Colab

Για να δουλέψετε στο Colab θα πρέπει να εισάγετε τα δεδομένα από το Kaggle. Γιαυτό χρειάζεστε ένα API key από το Kaggle. Επιπρόσθετα, το Colab έχει το μειονέκτημα ότι, σε αντίθεση με το Kaggle, μετά από 30 λεπτά χωρίς δραστηριότητα διαγράφει όλα τα δεδομένα του φακέλου "/content" (η τοποθεσία του vm). Υπάρχει ωστόσο η δυνατότητα να κάνετε mount το Google Drive σας και να έχετε persistancy.

Ακολουθήστε τον οδηγό "Downloading Kaggle datasets directly into Google Colab" που δείχνει βήμα-βήμα πως να επιτύχετε τα προαναφερθέντα.

Ταξινομητές

Στην περίπτωση του Kaggle dataset θα δουλέψετε με δύο ταξινομητές

- Mylti-Layer Perceptron (MLP), και
- Support Vector Machines (SVM).

Επισκόπηση

Χρησιμοποιήστε τις περιγραφές του Kaggle και τα ίδια τα δεδομένα για να κατανοήσετε το dataset και το task. Βεβαιωθείτε ότι έχετε εικόνα για όλα τα αρχεία του dataset, αν έχει περισσότερα.

Δώστε σε κελιά markdown τις βασικές πληροφορίες για το dataset, όπως κάνατε στο UCI dataset. Μπορείτε αν θέλετε να συμπεριλάβετε και οποιαδήποτε άλλη παρατήρηση κρίνετε σημαντική.

Μετρικές

Επιλέξτε την ή τις μετρικές με τις οποίες θα δουλέψετε και αιτιολογήστε την επιλογή σας.

Train-test split και σχήμα CV

Αν χρησιμοποιήσετε crossvalidation μέσω της βιβλιοθήκης σας, επιλέξτε εσείς το ποσοστό train-test και το σχήμα cross-validation. Αιτιολογήστε την επιλογή σας. Σε περίπτωση που δεν χρησιμοποιείται crossvalidation ή χρησιμοποιείται κάποια παραλλαγή, περιγράψτε τη διαδικασία και αιτιολογήστε τυχόν επιλογές σας.

Επίδοση out-of-the-box

Αρχικά θα δούμε πώς συμπεριφέρονται οι ταξινομητές χωρίς καμία βελτιστοποίηση (out-of-the-box) και με όλες τις παραμέτρους σε default τιμές.

Εκπαιδεύστε όλους τους εκτιμητές με ένα απλό fit σε ολόκληρο το training set και υπολογίστε την επίδοσή τους στο test set για τις δύο μετρικές.

Παρουσιάστε συνοπτικά και συγκριτικά την επιδοσή τους:

1. σε πίνακα markdown, και
2. σε bar plot σύγκρισης,

και σχολιάστε την επίδοσή τους συμπεριλαμβάνοντας και τους dummy ως baseline.

Βελτιστοποίηση

Για τους δύο ταξινομητές βελτιστοποιήστε την επίδοσή τους μέσω των διαδικασιών

- προεπεξεργασίας,

- ορισμού pipelines, και
- εύρεσης βέλτιστων υπερμαραμέτρων με αναζήτηση πλέγματος με διασταυρούμενη επικύρωση

Για την επιλογή μοντέλου θα χρησιμοποιήσετε συναρτήσεις από το sklearn και απαραιτήτως μια βιβλιοθήκη βελτιστοποίησης της επιλογής σας (Ray, Optuna, άλλη).

Δουλέψτε με κάθε ταξινομητή ξεχωριστά με στόχο τη βελτιστοποίησή του. Καθώς πειραματίζεστε, σημειώνετε τις επιλογές και τα συμπεράσματά σας, καθώς και όποια άλλα στοιχεία κρίνετε σημαντικά (πχ pipelines, διαδοχικά εύρη υπερπαραμέτρων, κλπ) ώστε να μπορείτε στη συνέχεια να περιγράψετε τη διαδικασία.

Τεκμηρίωση της διαδικασίας

Δώστε μια συνοπτική και μεστή περιγραφή όλης τις διαδικασίας που ακολουθήσατε για να καταλήξετε από το out-of-the-box στο βέλτιστο μοντέλο για κάθε ταξινομητή. Πρέπει να μπορούμε να καταλάβουμε σε κάθε βήμα ή στάδιο τί και γιατί το κάνατε, πάντα εν συντομία και ποιοτικά.

Παρουσίαση αποτελεσμάτων

Παρουσιάστε αναλυτικά την τελική αξιολόγηση της επίδοσης των δύο ταξινομητών μεμονωμένα και συγκριτικά.

Χρησιμοποιήστε για την παρουσίαση και τις επεξηγήσεις όλα τα εργαλεία που είναι διαθέσιμα όπως πίνακες, γραφήματα και φυσικά σχόλια.

Εστιάστε στις πιο σημαντικές παρατηρήσεις σας και αναλύστε τες διεξοδικά. Μην συμπεριλάβετε κάτι αν αυτό που δείχνει το έχουμε ήδη δει ποιοτικά και απλά επαναλαμβάνεται με μικρές ποσοτικές διαφορές.

Συμπεράσματα

Εξηγήστε μας ποιο είναι το τελικό μοντέλο ταξινομητή που προτείνετε για το dataset σας και γιατί. Εδώ μπορείτε επίσης να μιλήσετε και για επιμέρους επιλογές αν με βάση διαφορετικά κριτήρια υπερτερεί το ένα μοντέλο ή το άλλο.