Machine learning lab 1: basics of classification

October 21, 2020

Objectives: exploratory data analysis, run a classification algorithm, do cross-validation, evaluate performance of classification algorithm.

1 Basics of classification on a toy dataset

1) Download and open the heights/weight dataset. What are the different columns? Are there any missing values? Make a scatter plot of the dataset with height on the x-axis, weight on the y-axis, and gender as a color.

2) Before doing any ML on any data, we have to separate it into training and testing datasets. Why? Is there any alternative method to the basic train/test split? Apply a train/test split on the dataset.

3) We want an algorithm predicting gender from height and weight. Apply a logistic regression. What is the accuracy? Plot the decision boundary on the previous scatter plot.

4) Logistic regression does not only give binarized class decisions but also probabilities. Can you retrieve them?

5) Accuracy is not a good way to measure the performance of a classification algorithm. Why? Which measures are better (look at scikit-learn user guide, section 3.3)? Plot a ROC curve of the classifier.

6) (on board probably) Plot a decision boundary of the classifier on top of the scatter plot of question 1.

2 Project: EEG data classification

For this project you will need to produce a jupyter notebook that can be run alone, with answers and comments in text cells. To download the dataset, you will need to create a kaggle account and download it from `https://www.kaggle.com/birdy654/eeg-brainwave-dataset-feeling-emotions?select=emotions.csv`.

Caution: you will need to understand the concepts of multi-class classification, PCA and cross-validation to do it.

1) *Loading*. Open the dataset with pandas. What are the classes? How many features are there? What is the distribution of classes among the dataset?

2) *PCA*. Apply a PCA to the data. Plot the cumulative explained variance ratio for the first 20 components. Do a scatter plot of the data on the plane spanned by the first 2 components with class as color.

3) *Baseline algorithm.* Train a logistic regression and compute its performance (choose at least 2 appropriate measures). Tune the penalty type by doing a cross-validation.

4) *Baseline on PCA basis.* Train a logistic regression only with the 2 first PC components. What is the performance? Plot the decision boundaries on the scatter plot of question 2.

5) *Other algorithms* Choose at least 3 algorithms, tune some of their hyperparameters via cross-validation, and compare their performances. One can in particular look at SVMs or random trees. Looking at the algorithms of the library XGBoost is also encouraged.