INF264 - Exercise 6

October 2019

Instructions

- Deadline: 18/10/2019 , 23:59
- Submission place: https://mitt.uib.no/courses/19532/assignments
- Format: Your answers are to be returned in a single pdf report. You can also return scanned pages for your calculations. For results, your answers must include any values and plots that are requested in the Notebook.
- There will not be an exercise session on Tuesday 15/10/2019.

1 Bootstrap Aggregation Algorithm

A bootstrap data sample is a sample of a dataset with replacement. Bootstrap aggregating is a method in machine learning that can be used in some of the high-variance machine learning algorithms such as decision tree in order to prevent overfitting.

The dataset we will use in this tutorial is the Sonar dataset (click the name or download from the course repository). This is a dataset that describes sonar chirp returns bouncing off different surfaces. The 60 input variables are the strength of the returns at different angles. It is a binary classification problem that requires a model to differentiate rocks from metal cylinders. There are 208 observations.

Download the dataset and complete the following tasks.

- Implement the subsample method for choosing random samples of size ratio from the dataset.
- Implement the bagging algorithm for the number of classifiers given and train these classifiers on separate subsamples of training data and predict the test data by the rule of majority vote.
- Select the optimal number of trees (by cross-validation) and estimate its accuracy on unseen data.

2 Missing Values

In many of the real world machine learning problems, there are a lot of datapoints that have missing attributes. There are a few approaches when it comes to dealing with these values. Load the dataset "pima-indians-diabetes.csv". This dataset is known for having a lot of missing values.

- Count the number of missing values which are denoted by 0 in column indices (1,2,3,4,5).
- replace the zero values in these columns with numpy.NAN value to mark the missing values.
- remove the rows containing missing values and split the data into training and validation sets (1 to 3 ratio) and and try fitting a Multi layer perceptron classifier to the training data. What accuracy do you get on the validation set?
- load the dataset again and this time replace the NAN values with mean value of each column in a new training set and repeat the training. What accuracy do you get now on the previous validation set?

3 Imbalanced Data

There will be many datasets for a classification problem where the number of training samples are much higher for one class compared to another. These datasets are known as "Imbalanced Datasets". There are several methods of handling impalanced data.

Download the "creditcard.csv" from here and check the counts and percentages of two classes of dataset (normal and fraud). Prepare the dataset by splitting the train and validation sets.

- Train a Majority class classifier that always predicts the label of the most common class on the train set. (sklearn calls this method DummyClassifier) What is the accuracy on the validation set? Why is the accuracy high in spite of the simple approach? (hint: explain with the accuracy formula)
- Train Logistic regression on the train set and compute the accuracy, precision, recall, and F1 and plot the confusion matrix on the validation set.
- Use upsampling approach to balance the dataset and repeat the last step on the validation set and compute the values.
- Use downsampling on the train set and compute the measurements.
- Another way of dealing with imbalanced dataset is ensemble learning. Fit a random forest model on the dataset and compare the evaluation measurements of validation set with previous methods.