Homework 3

COM SCI X 450.4 - Machine Learning

Thursday 16th July, 2020

1 Introduction

This homework will be divided into 3 parts. In the first one, you will complete the code to implement a Decision Tree algorithm. You will then implement one of the ensemble techniques described in class and in Chapter 7 of the book. The second part will ask you to use the algorithms you implemented to make predictions on a publicly-available dataset. The last part will require comparing your results with the results from the models you implemented to the implementations from Scikit-Learn. Please time how long you took to finish this homework, as it will be a question at the end of the assignment.

2 Part I - Model Implementation (10 points + 10 bonus points)

For this section, you will implement components of the Decision Tree algorithm. You will complete the code in the notebook that will be shared. Then, you will implement one of the Ensemble techniques we have discussed in class. You can decide which one (Bagging, Pasting, Boosting, or Stacking).

Extra Points: Implementing the Random Forest algorithm.**(5 extra points)**

Extra Points: Implement two ensemble techniques (Bagging, Pasting, Boosting, or Stacking) **(5 extra points)**. Your Random Forest implementation from the previous exercise will count here.

3 Making Predictions and Evaluating Results (10 Points + 5 bonus points)

After you have implemented your models, we will be evaluating their results on a famous datasets, the Wine Classification Dataset. You can load it using Scikit-learn through the provided link. You will need to split the results into train and test.

You will create a new Jupyter Notebook to run you analysis or use the one provided with the Decision Tree code. Write your answers to the questions in a separate document. You will submit it with your code and notebook.

3.1 Wine Classification

Using the implementation of the Decision Tree algorithm that was provided and the ensemble techniques you implemented, answer the following questions:

- What is the best result you can get?
- How many splits did your Decision Tree algorithm need?
- (Bonus if you have implemented the ensemble) Which one gets better results, your ensemble or your decision tree?

4 Using Sklearn and comparing results (75 points + 15 bonus)

Now, select the Decision Tree Classifier and Random Forest models from sklearn. Train them using the same training set and make predictions using the same test set that you have used for your own models.

- Which one gets better results?
- Which one is faster during training? And during inference time?
- Using the methods from Scikit-Learn, play around with the parameters. What is the best result you can get? What feature was the most important?
- How did your implementation compare in results to the Sklearn methods?
- (Bonus) Use two other ensemble techniques from Sklearn or other Python libraries like XGBoost. Compare the results and report which one performs better.

5 How long did it take? (5 Points)

Please add to your written document how many hours you took to finish this assignment.

6 Deliverable

You will submit your assignment before Sunday, July 19th, before 11:59pm. You should submit a zipped file with:

- The Python source files where you implemented your algorithms
- The Python Notebook with the code you wrote
- A separate PDF document with the answers to the questions.

Start early, especially if you are not familiar with Python! If you require help, please contact instructor. Office hours can be scheduled if necessary as long as the instructor is contacted with 48 hour notice.