CS5101 Machine Learning

August-December 2021

Programming Assignment - 6

Topic: Ensemble Learning – Decision Tree, Random Forests

Follow the instructions carefully before attempting:

1) You are allowed to use all inbuilt libraries in today's assignment. Also you can choose either decision tree, random forest classifier or decision tree, random forest regression to solve the problem (you need to implement either classifier or regressor not both).

2) You must submit your code in a python .ipynb notebook with naming format as follows: Firstname_Lastname_assignment6.ipynb

3) For each question, create a separate text block containing the question followed by a code block containing the solution.

4) Your code must be properly commented explaining each step clearly.

5) If any of the above instructions are not followed, penalty will be there for the same.

6) Your code and answers will be checked for plagiarism and if found plagiarised, zero marks will be provided for assignment 6.

Problem Statement-1:

You are provided with a protein dataset. Learn a Decision Tree regressor/classifier and Random Forest (RF) regressor/classifier on the dataset separately and report your results with observation as mentioned below. You should optimize hyperparameters available for both Decision tree and RF regressor/classifier should report best results only.

- Your code should input train and test data from each of the corresponding files and learn both Decision Tree and RF models. X train has five feature values for each data point and Y train has actual y values. Same is the case with test data also.

  Regression

  If you are training decision tree and random forest regression on given dataset:

- Report following outputs in the python notebook itself with proper headings mentioning regressor name:

  1)Best hyper parameter values learned

  2)mse values for train data and test data for each of the regressor

  Classification

  If you are training decision tree and random forest classification on given dataset:

- Assume thresholds for y values as 6 and 12, i.e. all values below and equal to 6 belongs to class 1, all values above 6 and below and equal to 12 belongs to class 2, and all values above 12 belongs to class 3. Modify y labels accordingly and perform classification task.

- Report following outputs in the python notebook itself with proper headings mentioning classifier name :

  1)Best hyper parameter values learned

2)accuracy values and confusion matrix for train data and test data for each of the classifier.

Problem Statement-2:

- Generate a random n-class classification problem(Hint: may use make_classification method from sklearn.datasets) and implement AdaBoostClassifier on this custom dataset.

Evaluation Scheme:

- 1-mark: Decision Tree and Random Forest regressor/classifier (code)
- 2-marks: mse values/ accuracy value and confusion matrix (one for each ie., decision tree and Random forest)
- 1-mark: reporting best hyperparameter values obtained for each regressor/classifier
- 1-mark: for problem statement-2