Assignment 4

CS5824/ECE5424 – Fall 2020

Out: Oct 27, 2020
Due: Nov 6, 2020 (11:59pm)

For all questions, submit both PDF and .ipynb file with all of your code and output. Both Colab and Jupyter are allowed. Please submit both files separately on Canvas, no need to submit a zip file.

Late submissions incur a 0.5% penalty for every rounded up hour past the deadline for the first 24 hours and 0.6% penalty for every rounded up hour past the deadline for the second 24 hours. For example, an assignment submitted 5 hours and 15 min late will receive a penalty of ceiling(5.25) * 0.5% = 3%. A grade of zero will be given on the third late day.

Be sure to include your name and student number with your assignment.

1. **[50 points] Boosting:**

   (a) Extension to Assignment 1, Question 2. Instead of simple linear regression, implement boosting to linear regression. Use the simple linear regression without any penalty.

      i. **[20 points]** Does Boosting help improve the accuracy of linear regression? Give the detailed reasons on your assertions; both qualitatively and quantitatively. For the quantitative answer, prove the theory underlying your assertions mathematically.

   (b) Implement Adaptive Boosting classifier with logistic regression as the base learner. Use the same dataset as that used for Assignment 1, Question 1. Feel free to use the scikit-learn package for this question (1b).

      i. **[10 points]** Submit your code for AdaBoost using logistic regression as a base learner.

      ii. **[10 points]** Report the accuracy for all combinations of base learners and learning rates as follows: base learners (10, 25, 50) and learning rates (0.0001, 0.001, 0.01).

   (c) **[10 points]** Is it possible to improve the performance of KNN using an AdaBoost ensemble classifier that uses KNN as the base classifier? Give the detailed (qualitative) reasons on your assertions.

2. **[50 points] Bagging:**

   (a) **[10 points]** List two different ways for developing bagging ensembles and describe each method in a few words.

   (b) Implement Bagging KNN classifiers **from scratch**. Feel free to use your code for KNN from Assignment 1. Use the same dataset from Assignment 1, Question 1. Split the dataset in the following way: 70% (training data) and remaining (30%) for validation dataset. Number of neighbors = 19, range of number of base learners = [2, 25]. Use bootstrap sampling: each bootstrap sample used to learn each base learner is generated using random sampling with replacement, and should be of the same size as the training dataset used (70%).

      i. **[20 points]** Submit your code for Bagging KNN classifier.

      ii. **[10 points]** Report the training accuracy and the best number of base learners. Plot accuracy vs number of base learners.

iii. **[10 points]** Does Bagging on KNN help improve the accuracy? Give the detailed reasons for your assertions.

- For further clarification on the bootstrap sampling technique details, please check Section 2.2.1 (IID Resampling) in the paper: "A Resampling Technique for Relational Data Graphs" posted on resources and references on the assignment page.