CMPE 188-Machine
Learning and Big data
Homework

Instructor: Jahan Ghofraniha

Ensemble Methods using scikit-learn

Reading assignment

Review the lecture notes for ensemble techniques.

Coding Assignment

1. Download startup failure dataset and its description.
2. Perform standard EDA to get familiar with the dataset.
3. Use the sample code for the Ensemble classifier and modify it to work with the Startup dataset.
4. Compare the performance a decision tree, bagging classifier, random forest and a boosting classifier using all default settings and configuration used in the sample code.
5. Modify the random forest classifier tree depth hyper-parameter for the depth of 2-7 and analyze and comment on the results of the impact of changing the tree depth on the performance (replace the max_leaf_nodes=16 with max_depth = 2 (change from 2-7)).
6. For the Adaboost classifier, modify the learning rate to a higher rate and a low rate and analyze and comment on the results (you need to experiment with the learning rate to figure out what range makes sense).
7. Compare the performance of all models (all in steps 2-4) once again this time using cross-validation. Analyze the results and compare with the manual approach (steps 2-4)
8. Include your code, the results and explanation of the results either as a Jupyter notebook (.ipynb) plus a pdf of the notebook or as a .py plus a pdf document and upload to Canvas before the deadline.