CS 559: Machine Learning: Fundamentals and Applications

HW 8 Due: 4/22/2024 Monday 11:59 p.m.

- The assignment must be individual work and must not be copied or shared. Any tendency to cheat/copy evidence will lead to a 0 mark for the assignment.
- Students must only use Pandas, NumPy, matplotlib, and Spacy if the coding problem does not specify libraries/packages. Use of libraries other than the specified will be penalized.
- All problems must be submitted in a single notebook file. Do not work in the lecture notebook file.

1 Stacking Method [50 pts]

In the stacking method, the meta-learner predicts the target using the predicted values from the base learners as features. The objective of this assignment is to build a final model to predict the salary of baseball players. Scikit-learn machine learning algorithms can be used to learn, but Scikit-learn StackingRegressor can not be used.

a. Load the provided data set, **Hitters.csv**.

b. [15 pts] The data set is raw data set and it needs to be trained before applying base learners. There are a few missing targets and three discrete features, while the rest are continuous. Impute the missing target values with a mean value. Convert the text features to integers. Split the data set into train and test sets in the ratio of 8 to 2. Calculate the standard deviation of the test target. This will be used as a target error value.

c. [15 pts] Write a method that returns the new features from the base learners. Base learners can be any algorithms discussed in the lecture. There are no limits on base learner trials, but only three base learners will be collected and used in the meta-learning process. Explain why you choose them as base learners.

d. [15 pts] Then build a meta-learner. Investigate which base learner is the highest contributor and explain why it is so. This should be done with a train data set, not the test data set. Check if the MSE value is within 10% of the target error chosen from Question a. If not, determine the dominant type of the error, either bias or variance. Then, improve the model by upgrading the train set and regularizing the meta learner until the MSE value falls within 10% of the target error.

e. [5 pts] Once the train modeling is done, apply the test data set to predict the target and report the test MSE value. Determine if it is within 10% of the target error.