

Data Mining, CSCI 347, Fall 2019
Ensemble Learning, Nov. 18

Ensemble learning – multiple models are trained using the same learning algorithm

Often advantages in:

1. Take training data and derive several training sets
2. Learn a model from each
3. Combine learned models to produce an ensemble of learned models

Bagging – amalgamating output of models generated by training on multi-sets of the original data; the multi-sets are generated using combinations with repetition; bagging decreases the variance in a prediction

Bagging prediction:

1. Categorical class - take a vote (maybe a weighted vote)
2. For numeric class, take average (maybe a weighted average)

Straight average is called bagging

Weighted average is called boosting

Bias – how well the learning method matches the problem. The “persistent” error of a learning algorithm that can’t be eliminated even by taking an infinite number of training sets into account.

Variance – error from a particular dataset. T

Total expected error – bias + variance

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation.