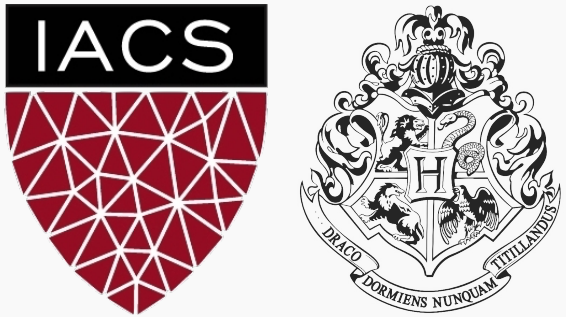
Lecture #16: Boosting

CS 109A, STAT 121A, AC 209A: Data Science

Pavlos Protopapas Kevin Rader



Lecture Outline

Review

Gradient Boosting

AdaBoost

Review

Bags and Forests of Trees

Last time we examined how the short-comings of single decision tree models can be overcome by ensemble methods - making one model out of many trees.

We focused on the problem of training large trees, these models have low bias but high variance.

We compensated by training an ensemble of full decision trees and then averaging their predictions - thereby reducing the variance of our final model.

Bags and Forests of Trees

- Bagging:
  - create an ensemble of full trees, each trained on a bootstrap sample of the training set;
  - average the predictions
- Random forest:
  - create an ensemble of full trees,each trained on a bootstrap sample of the training set;
  - in each tree and each split, randomly select a subset of predictors, choose a predictor from this subset for splitting;
  - average the predictions

Note that the ensemble building aspects of both method are embarrassingly parallel!

Gradient Boosting

AdaBoost