Ensemble Models

David Robinson

Ensemble Models

Ensemble modeling is a technique in machine learning that combines multiple models to achieve better predictive performance.

Common Ensemble Techniques

- **Bagging** reduces variance by training models on different random samples and averaging their predictions, such as random forest.
- **Boosting** reduces bias by sequentially building models that correct errors made by the previous one, such as AdaBoost and Gradient Boosting.
- **Stacking** combines preditions from different types of strong learners by training a meta-model on their outputs to improve final prediction accuracy.

Expected Test Error

$$\mathbb{E}_{D\sim P_n(x,y)\sim P}[(f_D(x)-y)^2] = \text{Variance} + \text{Bias} + \text{Noise}$$

where

- **Variance** $= \mathbb{E}_{x,D}[(f_D(x)-\bar{f}(x))^2]$ measures the variability of the predictions from model trained on subset D, $f_D(x)$, around the average prediction $\bar{f}(X)$.
- **Bias** $= \mathbb{E}_x[(\bar{f}(x)-\bar{y}(x))^2]$ measures the difference between the average model prediction $\bar{f}(x)$ and the true value $\bar{y}(x)$.
- **Noise** $= \mathbb{E}_{x,y}[(\bar{y}(x)-y)^2]$ represents the randomness in the data.

Random Forest

Random Forest is a bagging-based ensemble method.

1. Draw m samples from the original dataset D.
2. Train an independent decision tree for each sample.
3. At each node split within a tree, randomly select a subset of $k \leq d$ features, where d is the total number of features, and choose the best split only from this subset