Bagging and Random Forests

David S. Rosenberg

CDS, NYU

April 16, 2019

Contents

Ensemble Methods: Introduction

Ensembles: Parallel vs Sequential

- Ensemble methods combine multiple models
- **Parallel ensembles**: each model is built independently
    - e.g. bagging and random forests
    - Main Idea: Combine many (high complexity, low bias) models to reduce variance
- **Sequential ensembles**:
    - Models are generated sequentially
    - Try to add new models that do well where previous models lack

The Benefits of Averaging

A Poor Estimator

- Let $z, z_1, \dots, z_n$ i.i.d. $\mathbb{E}z = \mu$ and $\operatorname{Var}(z) = \sigma^2$.
- We could use any single z_i to estimate μ.
- Performance?
- Unbiased: $\mathbb{E}z_i = \mu$.
- Standard error of estimator would be σ.
    - The **standard error** is the standard deviation of the sampling distribution of a statistic.
    - $\operatorname{SD}(z) = \sqrt{\operatorname{Var}(z)} = \sqrt{\sigma^2} = \sigma$.

Variance of a Mean

- Let $z, z_1, \ldots, z_n$ be i.i.d. with $\mathbb{E}z = \mu$ and $\text{Var}(z) = \sigma^2$.
- Let's consider the average of the z_i's.
  - Average has the same expected value but smaller standard error:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right] = \mu \qquad \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right] = \frac{\sigma^2}{n}.$$

- Clearly the average is preferred to a single z_i as estimator.
- Can we apply this to reduce variance of general prediction functions?

Averaging Independent Prediction Functions

- Suppose we have B independent training sets from the same distribution.
- Learning algorithm gives B decision functions: $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$
- Define the average prediction function as:

$$\hat{f}_{\text{avg}} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b$$

- What's random here?
- The B independent training sets are random, which gives rise to variation among the $\hat{f}_b$'s.

Averaging Independent Prediction Functions

- Fix some particular $x_0 \in \mathcal{X}$.
- Then average prediction on x_0 is

$$\hat{f}_{\text{avg}}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x_0).$$

- Consider $\hat{f}_{\text{avg}}(x_0)$ and $\hat{f}_1(x_0), \ldots, \hat{f}_B(x_0)$ as random variables
  - Since the training sets were random
- We have no idea about the distributions of $\hat{f}_1(x_0), \ldots, \hat{f}_B(x_0)$ – they could be crazy...
- But we do know that $\hat{f}_1(x_0), \ldots, \hat{f}_B(x_0)$ are i.i.d. And that's all we need here...

Averaging Independent Prediction Functions

- The average prediction on x_0 is

$$\hat{f}_{\text{avg}}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x_0).$$

- $\hat{f}_{\text{avg}}(x_0)$ and $\hat{f}_b(x_0)$ have the same expected value, but
- $\hat{f}_{\text{avg}}(x_0)$ has smaller variance:

$$\begin{aligned}
\operatorname{Var}(\hat{f}_{\text{avg}}(x_0)) &= \frac{1}{B^2}\operatorname{Var}\left(\sum_{b=1}^{B} \hat{f}_b(x_0)\right) \\
&= \frac{1}{B}\operatorname{Var}\left(\hat{f}_1(x_0)\right)
\end{aligned}$$

Averaging Independent Prediction Functions

- Using

$$\hat{f}_{\text{avg}} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b$$

  seems like a win.
- But in practice we don't have B independent training sets...
- Instead, we can use **the bootstrap**....

Bagging

Bagging

- Draw B bootstrap samples $D^1, \ldots, D^B$ from original data $\mathcal{D}$.
- Let $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_B$ be the prediction functions from training on $D^1, \ldots, D^B$, respectively.
- The **bagged prediction function** is a **combination** of these:

$$\hat{f}_{\text{avg}}(x) = \text{Combine}\left(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)\right)$$

- How might we combine
  - prediction functions for regression?
  - binary class predictions?
  - binary probability predictions?
  - multiclass predictions?
- Bagging proposed by Leo Breiman (1996).

Bagging for Regression

- Draw B bootstrap samples $D^1, \ldots, D^B$ from original data $\mathcal{D}$.
- Let $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_B : \mathcal{X} \to \mathbf{R}$ be the real-valued prediction functions from $D^1, \ldots, D^B$, respectively.
- Bagged prediction function is given as

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x).$$

- **Empirically**, $\hat{f}_{\text{bag}}$ often performs similarly to what we'd get from training on B independent samples:
  - $\hat{f}_{\text{bag}}(x)$ has same expectation as $\hat{f}_1(x)$, but
  - $\hat{f}_{\text{bag}}(x)$ has smaller variance than $\hat{f}_1(x)$

Out-of-Bag Error Estimation

- Each bagged predictor is trained on about 63% of the data.
- Remaining 37% are called **out-of-bag (OOB)** observations.
- For ith training point, let

$$S_i = \left\{ b \mid D^b \text{ does not contain } i\text{th point} \right\}.$$

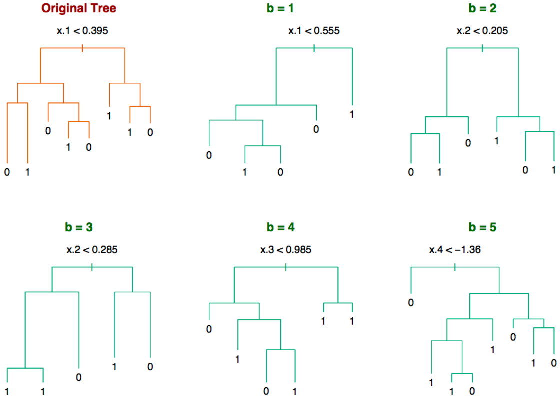
- The **OOB prediction** on x_i is

$$\hat{f}_{\text{OOB}}(x_i) = \frac{1}{|S_i|} \sum_{b \in S_i} \hat{f}_b(x_i).$$

- The OOB error is a good estimate of the test error.
- OOB error is similar to cross validation error – both are computed on training set.

Bagging Classification Trees

- Input space $\mathcal{X} = \mathbf{R}^5$ and output space $\mathcal{Y} = \{-1, 1\}$.

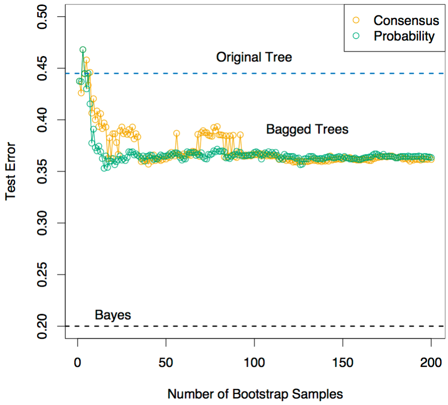


- Sample size $n = 30$
- Each bootstrap tree is quite different
- Different splitting variable at the root
- This high degree of variability from small perturbations of the training data is why tree methods are described as **high variance**.

From HTF Figure 8.9

Comparing Classification Combination Methods

- Two ways to combine classifications: consensus class or average probabilities.



From HTF Figure 8.10

Terms "Bias" and "Variance" in Casual Usage (Warning! Confusion Zone!)

- Restricting the hypothesis space $\mathcal{F}$ "**biases**" the fit
  - **away** from the best possible fit of the training data, and
  - **towards** a [usually] simpler model.
- Full, unpruned decision trees have very little bias.
- Pruning decision trees introduces a bias.
- **Variance** refers to how much the fit changes across different random training sets.
- **Stability** is another term referring to this concept (and I think should be preferred).
  - Low variance = High stability
- If different random training sets give very similar fits, then algorithm has high **stability**.
- Decision trees are found to be high variance (i.e. not very stable).

Conventional Wisdom on When Bagging Helps

- Hope is that bagging reduces variance without making bias worse.
- General sentiment is that bagging helps most when
  - Relatively unbiased base prediction functions
  - High variance / low stability
    - i.e. small changes in training set can cause large changes in predictions
- Hard to find clear and convincing theoretical results on this
- But following this intuition leads to improved ML methods, e.g. Random Forests

Random Forests

Recall the Motivating Principal of Bagging

- Averaging $\hat{f}_1, \ldots, \hat{f}_B$ reduces variance if they're based on i.i.d. samples from $P_{\mathcal{X}\times\mathcal{Y}}$
- Bootstrap samples are
  - independent samples from the training set, but
  - are **not** independent samples from $P_{\mathcal{X}\times\mathcal{Y}}$.
- This dependence limits the amount of variance reduction we can get.
- Would be nice to reduce the dependence between $\hat{f}_i$'s...

Random Forest

Main idea of random forests

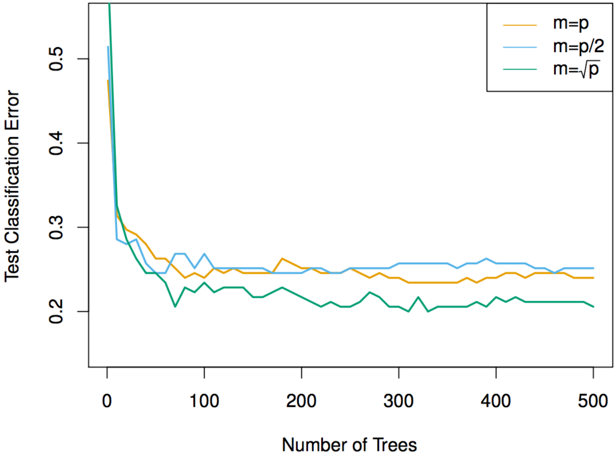
Use **bagged decision trees**, but modify the tree-growing procedure to reduce the dependence between trees.

- **Key step** in random forests:
  - When constructing **each tree node**, restrict choice of splitting variable to a randomly chosen subset of features of size m.
- Typically choose $m \approx \sqrt{p}$, where p is the number of features.
- Can choose m using cross validation.

Random Forest

- Usual approach is to build very deep trees (low bias)
- Diversity in individual tree prediction functions comes from
  - bootstrap samples (somewhat different training data) and
  - randomized tree building
- Bagging seems to work better when we are combining a diverse set of prediction functions.

Random Forest: Effect of m size



From *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Appendix

Variance of a Mean of Correlated Variables

- For $Z, Z_1, \ldots, Z_n$ i.i.d. with $\mathbb{E}Z = \mu$ and $\operatorname{Var}Z = \sigma^2$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \mu \qquad \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{\sigma^2}{n}.$$

- What if Z's are correlated?
- Suppose $\forall i \neq j$, $\operatorname{Corr}(Z_i, Z_j) = \rho$. Then

$$\operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2.$$

- For large n, the $\rho\sigma^2$ term dominates – limits benefit of averaging.