Bagging and Random Forests

David Rosenberg

New York University

February 25, 2015

Approximation Error and Estimation Error

- Recall the excess risk decomosition:

$$\textbf{Excess Risk}(\hat{f}_n) = \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}}^*)}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}^*) - R(f^*)}_{\text{approximation error}}$$

- Restricting the hypothesis space $\mathcal{F}$
  - leads to approximation error
  - but helps to reduce estimation error
- Now, we'll switch to the bias/variance terminology more common when discussing the topics of this lecture.

Bias and Variance

- Restricting the hypothesis space $\mathcal{F}$ "**biases**" the fit
  - **towards** a simpler model and
  - **away** from the best possible fit of the training data.
- Full, unpruned decision trees have very little bias.
- Pruning decision trees introduces a bias.
- **Variance** describes how much the fit changes across different random training sets.
- Decision trees are found to be high variance.

Bias and Variance

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$
- From Homework #1, recall that for square error, the bayes prediction function is
$$f^*(x) = \mathbb{E}\left[Y \mid X = x\right]$$
- Let's consider a prediction function $\hat{f}$ trained on a random set of data.
- $\hat{f}$ is random because training data is random.

Excess Risk for Square Error

- Excess risk, conditional on $X = x$ is

$$\begin{aligned}\text{ExcessRisk}(\hat{f} \mid X = x) &= \underbrace{\mathbb{E}\left[\left(Y - \hat{f}(x)\right)^2 \mid X = x\right]}_{\text{Risk of } \hat{f}} \\ &\quad \underbrace{-\mathbb{E}\left[(Y - f^*(x))^2 \mid X = x\right]}_{\text{Risk of } f^*}\end{aligned}$$

- Can show

$$\text{ExcessRisk}(\hat{f} \mid X = x) = \mathbb{E}\left[\left(\hat{f}(x) - f^*(x)\right)^2\right].$$

- What's random?

Bias-Variance Decomposition for Excess Risk

- Prediction $\hat{f}(x)$ for any fixed input x has bias and variance.

$$\begin{aligned}\text{Bias}(\hat{f}(x)) &= \mathbb{E}\left[\hat{f}(x)\right] - f^*(x)\\ \operatorname{Var}\left(\hat{f}(x)\right) &= \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]\end{aligned}$$

- Can show **bias-variance decomposition** for excess risk:

$$\mathbb{E}\left[\left(\hat{f}(x) - f^*(x)\right)^2\right] = \left[\text{Bias}(\hat{f}(x))\right]^2 + \operatorname{Var}\left(\hat{f}(x)\right)$$

- Could we reduce variance without increasing bias?

Variance of a Mean

- Let $Z_1, \ldots, Z_n$ be independent r.v's with mean μ and variance σ^2.
- Suppose we want to estimate μ.
- We could use any single Z_i to estimate μ.
- Variance of estimate would be σ^2.
- Let's consider the average of the Z_i's.
- Average has the same expected value but smaller variance:

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

- Can we apply this to reduce variance of prediction models?

Averaging Independent Prediction Functions

- Suppose we have B independent training sets.
- Let $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$ be the prediction models for each set.
- Define the average prediction function as:

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

- The average prediction function has lower variance than an individual prediction function.
- But in practice we don't have B independent training sets...

Variability of an Estimator

- Suppose we have a random sample $X_1, \ldots, X_n$.
- Compute some function of the data, such as

$$\hat{\mu} = \phi(X_1, \ldots, X_n).$$

- We want to put error bars on $\hat{\mu}$, so we need to estimate $\operatorname{Var}(\hat{\mu})$.
- Ideal scenario:
    - Attain B samples of size n.
    - Compute $\hat{\mu}_1, \ldots, \hat{\mu}_B$.
    - The sample variance of $\hat{\mu}_1, \ldots, \hat{\mu}_B$ estimates $\operatorname{Var}(\hat{\mu})$
- Again, we don't have B samples. Only 1.

The Bootstrap Sample

Definition

A **bootstrap sample** from $\mathcal{D} = \{X_1, \ldots, X_n\}$ is a sample of size n drawn *with replacement* from $\mathcal{D}$.

- In a bootstrap sample, some elements of $\mathcal{D}$
  - will show up multiple times,
  - some won't show up at all.
- Each X_i has a probability $(1 - 1/n)^n$ of not being selected.
- Recall from analysis that for large n,

$$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} \approx .368.$$

- So we expect ~63.2% of elements of $\mathcal{D}$ will show up at least once.

The Bootstrap Method

Definition

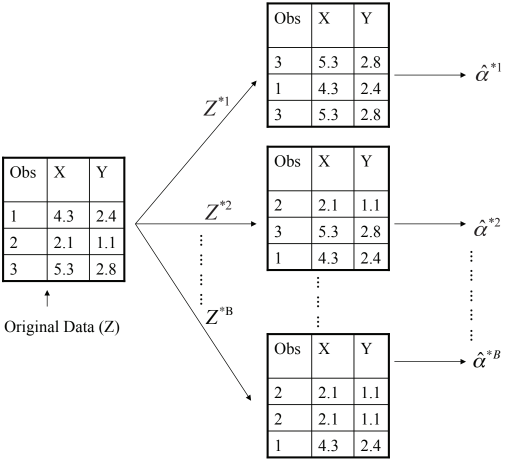
A **bootstrap method** is when you *simulate* having B independent samples by taking B bootstrap samples from the sample $\mathcal{D}$.

- Given original data $\mathcal{D}$, compute B bootstrap samples $D^1, \ldots, D^B$.
- For each bootstrap sample, compute some function

$$\phi(D^1), \ldots, \phi(D^B)$$

- Work with these values as though $D^1, \ldots, D^B$ were independent.
- **Amazing fact:** Things usually come out very close to what we'd get with independent samples.

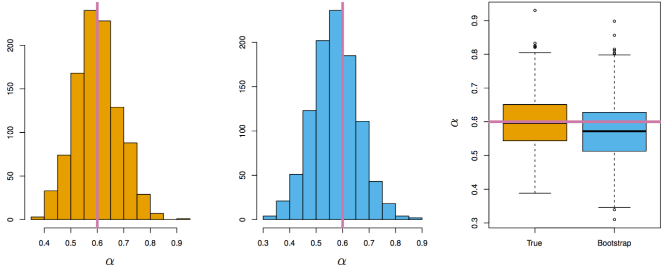
The Bootstrap Method



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.

Independent vs Bootstrap Samples

- Original sample size $n = 100$ (simulated data)
- $\hat{\alpha}$ is a complicated function of the data.
- Compare values of $\hat{\alpha}$ on
  - 1000 independent samples of size 100, vs
  - 1000 bootstrap samples of size 100



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.

Bagging

- Suppose we had B independent training sets.
- Let $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$ be the prediction models from each set.
- Define the average prediction function as:

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

- But we don't have B independent training sets.
- **Bagging** is when we use B bootstrap samples as training sets.
- Bagging estimator given as

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

  where $\hat{f}_b^*$ is trained on the b'th bootstrap sample.
- Bagging proposed by Leo Breiman (1996).

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

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

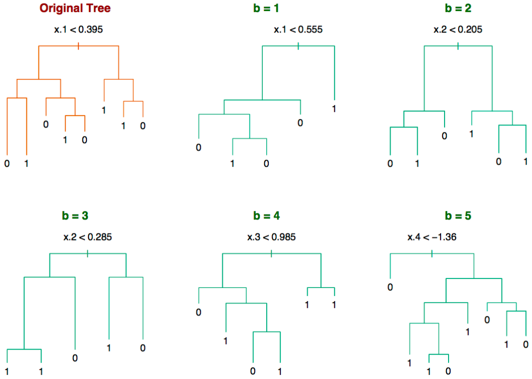
- The **OOB prediction** on x_i is

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

- The OOB error is a good estimate of the test error.
- For large enough B, OOB error is like cross validation.

Bagging Trees

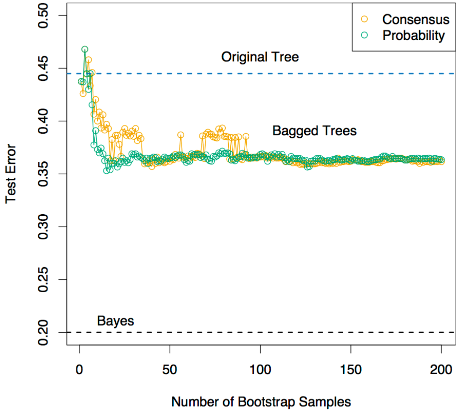
- Input space $\mathcal{X} = \mathbf{R}^5$ and output space $\mathcal{Y} = \{-1, 1\}$.
- Sample size $N = 30$ (simulated data)



From ESL Figure 8.9

Bagging Trees

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



From ESL Figure 8.10

Variance of a Mean of Correlated Variables

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

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

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

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

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

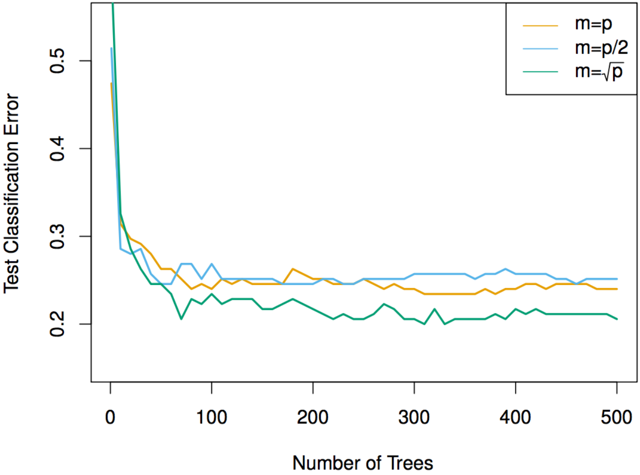
Random Forest

Main idea of random forests

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

Random Forest: Effect of m size

- See movie in Criminisi et al's PowerPoint: `http://research.microsoft.com/en-us/um/people/antcrim/ACriminisi_DecisionForestsTutorial.pptx`