

# Bootstrap, Bagging, and Random Forests

David Rosenberg

New York University

March 22, 2017

# Bias and Variance

# Parameters

- Suppose we have a probability distribution  $P$ .
- Often want to estimate some characteristic of  $P$ .
  - e.g. expected value, variance, kurtosis, median, etc...
- These things are called **parameters** of  $P$ .
- A **parameter**  $\mu = \mu(P)$  is any function of the distribution  $P$ .
- Question: Is  $\mu$  random?
- Answer: Nope. For example if  $P$  has density  $f(x)$  on  $\mathbf{R}$ , then mean is

$$\mu = \int_{-\infty}^{\infty} xf(x) dx,$$

which is just an integral - nothing random.

# Statistics and Estimators

- Suppose  $\mathcal{D}_n = (x_1, x_2, \dots, x_n)$  is an i.i.d. sample from  $P$ .
- A **statistic**  $s = s(\mathcal{D}_n)$  is any function of the data.
- A statistic  $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$  is a **point estimator** of  $\mu$  if  $\hat{\mu} \approx \mu$ .
- Question: Are statistics and/or point estimators random?
- Answer: Yes, since we're considering the data to be random.
  - The function  $s(\cdot)$  isn't random, but we're plugging in random inputs.

# Examples of Statistics

- Mean:  $\bar{x}(\mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n x_i$ .
- Median:  $m(\mathcal{D}_n) = \text{median}(x_1, \dots, x_n)$
- Sample variance:  $\sigma^2(\mathcal{D}_n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}(\mathcal{D}_n))^2$

Fancier:

- A data histogram is a statistic.
- Empirical distribution function.
- A confidence interval.

# Statistics are Random

- Statistics are random, so they have probability distributions.
- The distribution of a statistic is called a **sampling distribution**.
- We often want to know some **parameters** of the sampling distribution.
  - Most commonly the mean and the standard deviation.
- The standard deviation of the sampling distribution is called the **standard error**.
- Question: Is standard error random?
- Answer: Nope. It's a parameter of a distribution.

# Bias and Variance for Real-Valued Estimators

- Let  $\mu : P \mapsto \mathbf{R}$  be a real-valued parameter.
- Let  $\hat{\mu} : \mathcal{D}_n \mapsto \mathbf{R}$  be an estimator of  $\mu$ .
- We define the **bias** of  $\hat{\mu}$  to be  $\text{Bias}(\hat{\mu}) = \mathbb{E}\hat{\mu} - \mu$ .
- We define the **variance** of  $\hat{\mu}$  to be  $\text{Var}(\hat{\mu}) = \mathbb{E}\hat{\mu}^2 - (\mathbb{E}\hat{\mu})^2$ .
- An estimator is **unbiased** if  $\text{Bias}(\hat{\mu}) = \mathbb{E}\hat{\mu} - \mu = 0$ .

Neither bias nor variance depend on a specific sample  $\mathcal{D}_n$ . We are taking expectation over  $\mathcal{D}_n$ .

# Estimating Variance of an Estimator

- To estimate  $\text{Var}(\hat{\mu})$  we need estimates of  $\mathbb{E}\hat{\mu}$  and  $\mathbb{E}\hat{\mu}^2$ .
- Instead of a single sample  $\mathcal{D}_n$  of size  $n$ , suppose we had
  - $B$  independent samples of size  $n$ :  $\mathcal{D}_n^1, \mathcal{D}_n^2, \dots, \mathcal{D}_n^B$
- Can then estimate

$$\mathbb{E}\hat{\mu} \approx \frac{1}{B} \sum_{i=1}^B \hat{\mu}(\mathcal{D}_n^i)$$

$$\mathbb{E}\hat{\mu}^2 \approx \frac{1}{B} \sum_{i=1}^B [\hat{\mu}(\mathcal{D}_n^i)]^2$$

and

$$\text{Var}(\hat{\mu}) \approx \frac{1}{B} \sum_{i=1}^B [\hat{\mu}(\mathcal{D}_n^i)]^2 - \left[ \frac{1}{B} \sum_{i=1}^B \hat{\mu}(\mathcal{D}_n^i) \right]^2.$$



## Putting “Error Vars” on Estimator

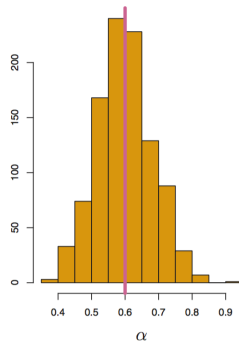
- Why do we even care about estimating variance?
- Would like to report a confidence interval for our point estimate:

$$\hat{\mu} \pm \sqrt{\widehat{\text{Var}}(\hat{\mu})}$$

- (This confidence interval assumes  $\hat{\mu}$  is unbiased.)
- Our **estimate of standard error** is  $\sqrt{\widehat{\text{Var}}(\hat{\mu})}$ .

# Histogram of Estimator

- Want to estimate  $\alpha = \alpha(P)$  for some known  $P$ , and some complicated  $\alpha$ .
- Point estimator  $\hat{\alpha} = \hat{\alpha}(\mathcal{D}_{100})$  for samples of size 100.
- Histogram of  $\hat{\alpha}$  for 1000 random datasets of size 100:



## Practical Issue

- We typically get only one sample  $\mathcal{D}_n$ .
- We could divide it into  $B$  groups.
- Our estimator would be  $\hat{\mu} = \hat{\mu}(\mathcal{D}_{n/B})$ .
- And we could get a variance estimate for  $\hat{\mu}$ .
- But the estimator itself would not be as good as if we used all data:

$$\hat{\mu} = \hat{\mu}(\mathcal{D}_n).$$

- Can we get the best of both worlds?
  - A good point estimate AND a variance estimate?

# The Bootstrap

# The Bootstrap Sample

## Definition

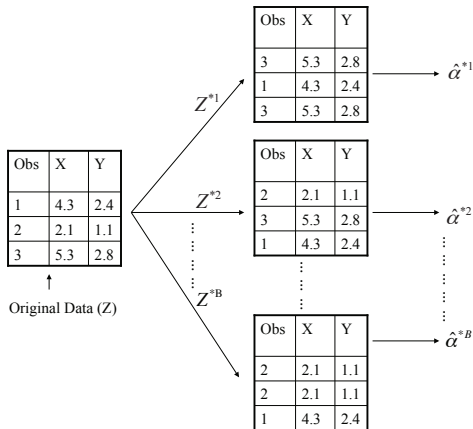
A **bootstrap sample** from  $\mathcal{D}_n = \{x_1, \dots, x_n\}$  is a sample of size  $n$  drawn *with replacement* from  $\mathcal{D}_n$ .

- 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  $\sim 63.2\%$  of elements of  $\mathcal{D}$  will show up at least once.

# The Bootstrap Sample



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.

# The Bootstrap Method

## Definition

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

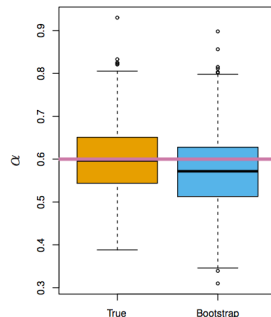
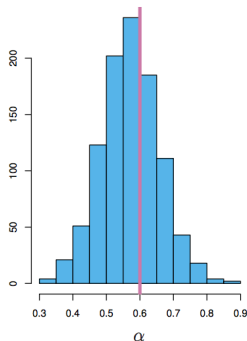
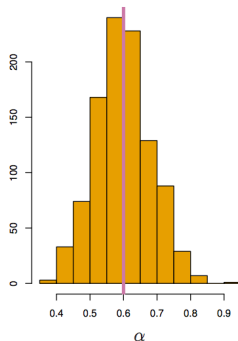
- Given original data  $\mathcal{D}_n$ , compute  $B$  bootstrap samples  $D_n^1, \dots, D_n^B$ .
- For each bootstrap sample, compute some function

$$\phi(D_n^1), \dots, \phi(D_n^B)$$

- Work with these values as though  $D_n^1, \dots, D_n^B$  were i.i.d.  $P$ .
- **Amazing fact:** Things often come out very close to what we'd get with independent samples from  $P$ .

# Independent vs Bootstrap Samples

- Want to estimate  $\alpha = \alpha(P)$  for some known  $P$ . (some complicated  $\alpha$ )
- Point estimator  $\hat{\alpha} = \hat{\alpha}(\mathcal{D}_{100})$  for samples of size 100.
- Histogram of  $\hat{\alpha}$  based on
  - 1000 independent samples of size 100, vs
  - 1000 bootstrap samples of size 100





# The Bootstrap in Practice

- Suppose we have an estimator  $\hat{\mu} = \hat{\mu}(\mathcal{D}_n)$ .
- To get error bars, we can compute the “bootstrap variance”.
  - Draw  $B$  bootstrap samples.
  - Compute empirical variance of  $\hat{\mu}(\mathcal{D}_n^1), \dots, \hat{\mu}(\mathcal{D}_n^B)$ ..
- Could report

$$\hat{\mu}(\mathcal{D}_n) \pm \sqrt{\text{Bootstrap Variance}}$$

# The Benefits of Averaging

# A Lousy Estimator

- Let  $Z, Z_1, \dots, Z_n$  i.i.d.  $\mathbb{E}Z = \mu$  and  $\text{Var}Z = \sigma^2$ .
- We could use any single  $Z_i$  to estimate  $\mu$ .
- Performance?
  - Unbiased:  $\mathbb{E}Z_i = \mu$ .
  - Variance of estimator would be  $\sigma^2$ .

## Variance of a Mean

- Let  $Z, Z_1, \dots, Z_n$  i.i.d.  $\mathbb{E}Z = \mu$  and  $\text{Var}Z = \sigma^2$ .
- Let's consider the average of the  $Z_i$ 's.
  - Average has the same expected value but smaller variance:

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

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

# Averaging Independent Prediction Functions

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

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

- What's random here?

# Averaging Independent Prediction Functions

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

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

- Consider  $\hat{f}_{\text{avg}}(x)$  and  $\hat{f}_1(x), \dots, \hat{f}_B(x)$  as random variables. (They are.)
- $\hat{f}_1(x), \dots, \hat{f}_B(x)$  are i.i.d.
- $\hat{f}_{\text{avg}}(x)$  and  $\hat{f}_b(x)$  have the same expected value, but
- $\hat{f}_{\text{avg}}(x)$  has smaller variance:

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

# 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  $D^1, \dots, D^B$  from original data  $\mathcal{D}$ .
- Let  $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_B$  be the decision functions for each set.
- The **bagged decision function** is a **combination** of these:

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

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

# Bagging for Regression

- Draw  $B$  bootstrap samples  $D^1, \dots, D^B$  from original data  $\mathcal{D}$ .
- Let  $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_B : \mathcal{X} \rightarrow \mathbf{R}$  be the predictions functions for each set  $x$ .
- Bagged prediction function is given as

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

- If bootstrap samples were independent draws from  $P$ ,
  - $\hat{f}_{\text{bag}}(x)$  would have the same expectation as  $\hat{f}_1(x)$ , but
  - $\hat{f}_{\text{bag}}(x)$  would have smaller variance.
- Empirically: Often get a similar effect for bagging.

## 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  $i$ th training point, let

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

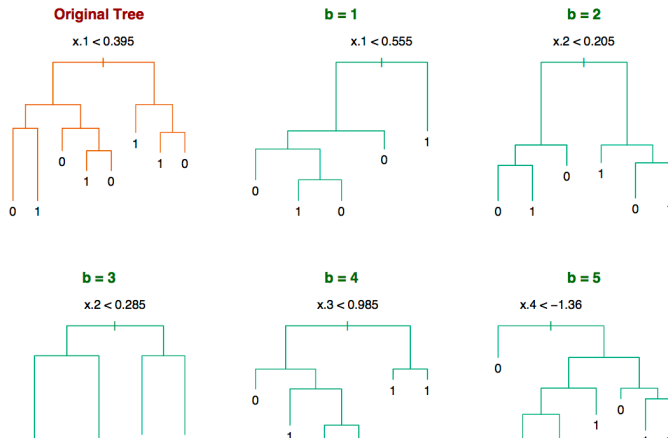
- The **OOB prediction** on  $x_i$  is

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

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

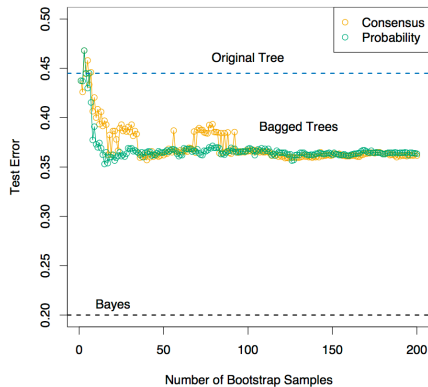
# Bagging Trees

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



# Bagging Trees

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



From ESL Figure 8.10

# Terms “Bias” and “Variance” in Casual Usage

- 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.

## Conventional Wisdom on When Bagging Helps

- Bagging does nothing to eliminate bias.
- Hope is that bagging reduces variance.
- General sentiment is that bagging helps most when
  - Relatively unbiased base predictions
  - High variance
    - e.g. small changes in training set can cause large changes in predictions
- I'm not aware of solid theory on this...
- Empirical observation
  - Bagging trees works well.
  - Trees have high variance and low bias.
  - QED?

# Random Forests



## Recall the Motivating Principal of Bagging

- Averaging  $\hat{f}_1, \dots, \hat{f}_B$  reduces variance, if they're based on i.i.d. samples.
- Bootstrap samples are not independent.
- This probably limits the amount of variance reduction we can get.
- Would be nice to reduce the dependence between  $\hat{f}_i$ 's...

# Variance of a Mean of Correlated Variables

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

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

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

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

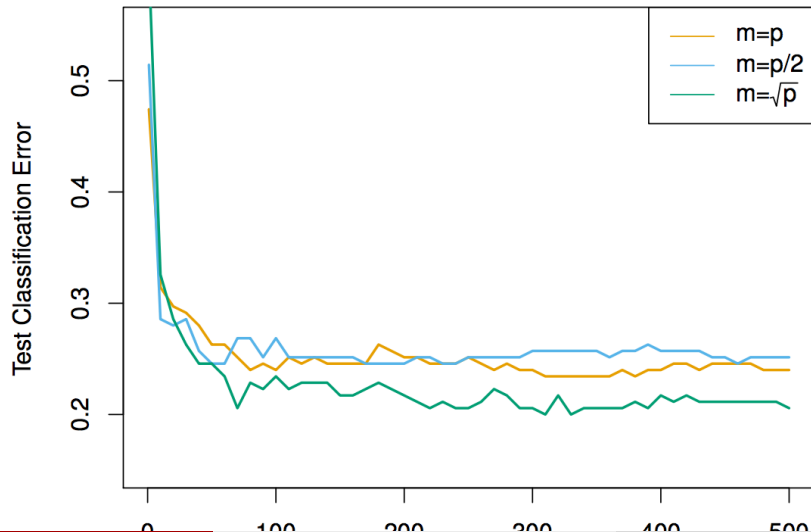
- For large  $n$ , the  $\rho \sigma^2$  term dominates – limits benefit of averaging.

# 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

## 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](http://research.microsoft.com/en-us/um/people/antcrim/ACriminisi_DecisionForestsTutorial.pptx)

# TensorForest: “Scalable random forests on TensorFlow”

- New tree building algorithm  $\rightarrow$  “ExtraTrees”
- batches of points keep arriving to a tree building job
- from the first  $\#$ features (tunable) points we receive, we sample random (feature,feature value) pairs. These are our (only) candidate splits.
- When we have accumulated enough points (atmost 2000) we decide on the best split.
- These points could then be pushed down to the nodes of the lower layers, but I think they’re thrown out.
- Next we wait for more points to arrive, push them down to leaf nodes... ..
  - (can actually use this to change presentation of random forests, and even tree building)
  - maintain a set of feature split point pairs...
  - could even have a more generic set of splitting hypotheses