Bagging and Random Forests

He He
Slides based on Lectures 11a - 11b from David Rosenberg's course materials
(https://github.com/davidrosenberg/mlcourse)

CDS, NYU

April 6, 2021

Bagging and Random Forests

Recap: statistic and point estimator

- Observe data $\mathcal{D} = (x_1, x_2, \ldots, x_n)$ sampled i.i.d. from a parametric distribution $p(\cdot \mid \theta)$.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
  - E.g., sample mean, sample variance, histogram, empirical data distribution
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is a **point estimator** of θ if $\hat{\theta} \approx \theta$.

Review questions

In frequentist statistics,

- Is θ random?
- Is $\hat{\theta}$ random?
- Is the function $s(\cdot)$ random?

Recap: bias and variance of an estimator

- Statistics are random, so they have probability distributions.
- The distribution of a statistic is called a **sampling distribution**.
- The standard deviation of the sampling distribution is called the **standard error**.
- What are some parameters of the sampling distribution we might be interested in?

  Bias $\text{Bias}(\hat{\theta}) \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{\theta}\right] - \theta.$

  Variance $\text{Var}(\hat{\theta}) \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{\theta}^2\right] - \mathbb{E}^2\left[\hat{\theta}\right].$

- [discussion]Is bias and variance random?
- [discussion]Why do we care about variance?

Variance of a Mean

Using a single estimate may have large standard error

- Let $\hat{\theta}(\mathcal{D})$ be an unbiased estimator: $\mathbb{E}\left[\hat{\theta}\right] = \theta$, $\text{Var}(\hat{\theta}) = \sigma^2$.
- We could use a single estimate $\hat{\theta} = \hat{\theta}(\mathcal{D})$ to estimate θ.
- The standard error is $\sqrt{\text{Var}(\hat{\theta})} = \sigma$.

Average of estimates has smaller standard error

- Consider a new estimator that takes the average of i.i.d. $\hat{\theta}_1, \ldots, \hat{\theta}_n$ where $\hat{\theta}_i = \hat{\theta}(\mathcal{D}^i)$.
- Average has the same expected value but smaller standard error:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right] = \theta \quad \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right] = \frac{\sigma^2}{n} \tag{1}$$

Averaging Independent Prediction Functions

Let's apply *averaging* to reduce variance of prediction functions.

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

$$\hat{f}_{\text{avg}} \stackrel{\text{def}}{=} \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b \tag{2}$$

- [discussion]What's random here?
- Concept check: What's the distribution of $\hat{f}$ called? What do we know about the distribution?

Averaging reduce variance of predictions

- 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 (see 1):

$$\operatorname{Var}(\hat{f}_{\text{avg}}(x_0)) = \frac{1}{B}\operatorname{Var}\left(\hat{f}_1(x_0)\right)$$

- Problem: in practice we don't have B independent training sets...

The Bootstrap Sample

How do we simulate multiple samples when we only have one?

- A **bootstrap sample** from $\mathcal{D}_n = (x_1, \ldots, x_n)$ is a sample of size n drawn *with replacement* from $\mathcal{D}_n$.
- Some elements of $\mathcal{D}_n$ will show up multiple times, and some won't show up at all.

[discussion]How similar are the bootstrap samples?

- Each x_i has a probability of $(1 - 1/n)^n$ of not being selected.
- Recall from analysis that for large n,

$$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} \approx .368. \tag{3}$$

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

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 $D_n^1, \ldots, D_n^B$.
- For each bootstrap sample, compute some function

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

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

Independent vs Bootstrap Samples

- Want to estimate $\alpha = \alpha(P)$ for some unknown P and some complicated α.
- 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

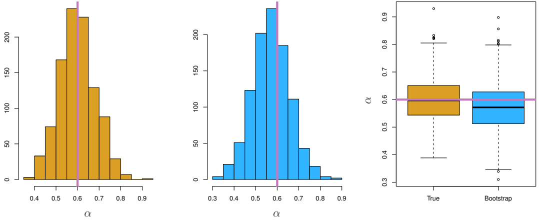


Figure 5.10 from *ISLR* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Side note: Bootstrap in Practice

We can use bootstrap to get error bars in a cheap way.

- Suppose we have an estimator $\hat{\theta} = \hat{\theta}(\mathcal{D}_n)$.
- To get error bars, we can compute the "*bootstrap variance*".
  - Draw B bootstrap samples.
  - Compute sample variance of $\hat{\theta}(\mathcal{D}_n^1), \ldots, \hat{\theta}(\mathcal{D}_n^B)$..
  - Could report
$$\hat{\theta}(\mathcal{D}_n) \pm \sqrt{\text{Bootstrap Variance}}$$

Ensemble methods

Key ideas:

- *Averaging* i.i.d. estimates reduces variance without making bias worse.
- Can use bootstrap to simulate multiple data samples.

Ensemble methods:

- Combine outputs from multiple models.
  - Same learner on different datasets: ensemble + bootstrap = bagging.
  - Different learners on one dataset: they may make similar errors.
- Parallel ensemble: models are built independently, e.g., bagging
- Sequential ensemble: models are built sequentially, e.g., boosting
  - Try to add new learners that do well where previous learners lack

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:

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

- [discussion]How might we combine
  - prediction functions for regression?
  - binary class predictions?
  - binary probability predictions?
  - multiclass predictions?

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

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

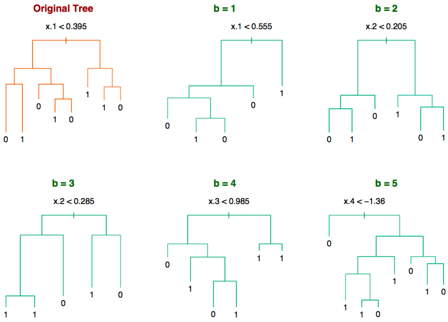
- The **OOB prediction** on x_i is

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

- 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
- **High variance**: high degree of model variability from small perturbations of the training data.
- Conventional wisdom: Bagging helps most when base learners are relatively unbiased but has high variance / low stability $\Longrightarrow$ decision trees.

From HTF Figure 8.9

Variance of a Mean of Correlated Variables

Recall the motivating principle of bagging:

- For $\hat{\theta}_1, \ldots, \hat{\theta}_n$ *i.i.d.* with $\mathbb{E}\left[\hat{\theta}\right] = \theta$ and $\operatorname{Var}\left[\hat{\theta}\right] = \sigma^2$,

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

- What if $\hat{\theta}$'s are correlated?
- Suppose $\forall i \neq j$, $\text{Corr}(\hat{\theta}_i, \hat{\theta}_j) = \rho$. Then

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

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

Correlation between bootstrap samples

- 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.
- Solution: reduce the dependence between $\hat{f}_i$'s by randomization.

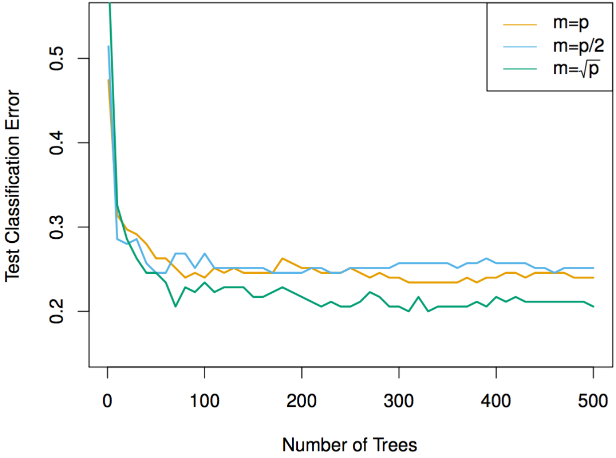
Random Forest

Key idea

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

- Build a collection of trees independently (in parallel).
- When constructing each tree node, restrict choice of splitting variable to a randomly chosen subset of features of size m.
  - Avoid dominance by strong features.
- 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.

Review

- Usual approach is to build very deep trees—low bias but high variance
- Ensembling many models reduces variance
  - Motivation: Mean of i.i.d. estimates has smaller variance than single estimate.
- Use bootstrap to simulate many data samples from one dataset
  - $\implies$ Bagged decision trees
- But bootstrap samples (and the induced models) are correlated.
- Bagging seems to work better when we are combining a diverse set of prediction functions.
  - $\implies$ random forests (randomized tree building)