CSCE 633: Machine Learning

Lecture 11: Random Forests

Texas A&M University

10-2-18

Last Time

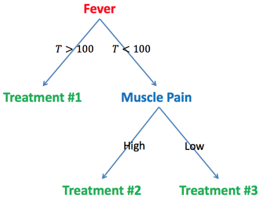
- Decision Trees

Goals of this lecture

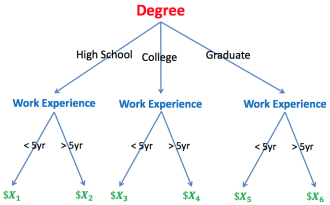
- Random Forest

Many decisions are tree-like structures

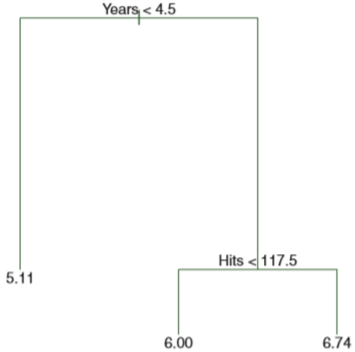
Medical treatment



Salary in a company

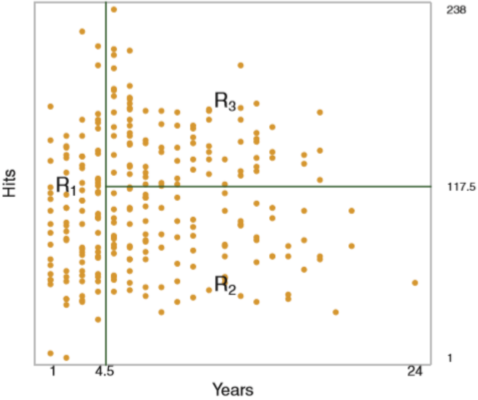


Create a basic tree



make a prediction of e raised to the regression value What is the most important variable?

This partitions our data space



These regions are known as leaves or terminal nodes

Gini Index and Entropy - A Review

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

, which measures the total variance across K classes. This is a measure of node purity.

$$H = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$$

, Entropy which takes a value near 0 if all the $\hat{p}$ are near zero or one - smaller value if node is pure

Decision Trees - A Review

Advantages

- The models are transparent: easily interpretable by human (as long as the tree is not too big)
- Data can contain combination of continuous and discrete features
- Decision tress more closely mirror human decision making than do regressions?
- Graphical representation
- Qualitative predictors without dummy variables!

Disadvantages

- Usually not same level of predictive accuracy as other regression and classification approaches
- Non-robust - small change in data can change a large amount of the final estimated tree
- Solutions? Bagging, Random Forest, Boosting

Random Forests

- We grow many classification trees through bagging & randomization
- Bagging (**B**ootstrap **agg**regat**ing**)
  - Generate independently bootstrap datasets from original data
  - Run a decision tree in each one of them
- Randomize over the set of attributes
  - Before growing a bootstrap decision tree
  - When splitting an interior node of the classification tree
- No pruning (small trees)
- For each sample, each tree "votes" for a class and we perform majority voting for final decision

Advantages

- Very good performance in practice
- Runs efficiently on large data bases
- Runs efficiently on large feature sets
- Gives estimates of the most relevant variables for the problem

Bagging

Bootstrapped Aggregating

- Decision Trees suffer from high variance
- If we split data in half, tree could be very different on both halves

Bagging

Bootstrapped Aggregating

- Bootstrap aggregation (bagging) reduces variance!

Bootstrapped Aggregating

- Bootstrap aggregation (bagging) reduces variance!
- Given n independent observations $Z_1, \cdots, Z_n$ each with variance σ^2

Bagging

Bootstrapped Aggregating

- Bootstrap aggregation (bagging) reduces variance!
- Given n independent observations $Z_1, \cdots, Z_n$ each with variance σ^2
- Variance of mean $\bar{Z} = \frac{\sigma^2}{n}$

Bagging

Bootstrapped Aggregating

- Bootstrap aggregation (bagging) reduces variance!
- Given n independent observations $Z_1, \cdots, Z_n$ each with variance σ^2
- Variance of mean $\bar{Z} = \frac{\sigma^2}{n}$
- What if we apply this to decision trees? (Classification and Regression Trees - CART)

Bagging

Bootstrapped Aggregating

- Take B different training sets

Bagging

Bootstrapped Aggregating

- Take B different training sets
- Train f^1 on training set 1

Bagging

Bootstrapped Aggregating

- Take B different training sets
- Train f^1 on training set 1
- Train f^2 on training set 2

Bagging

Bootstrapped Aggregating

- Take B different training sets
- Train f^1 on training set 1
- Train f^2 on training set 2
- ...

Bagging

Bootstrapped Aggregating

- Take B different training sets
- Train f^1 on training set 1
- Train f^2 on training set 2
- ...
- Can average result over B trees for a single low-variance model from $\hat{f}^1(x)$, $\hat{f}^2(x)$,$\cdots$, $\hat{f}^B(x)$ as $\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$

Bagging

Bootstrapped Aggregating

- Take B different training sets
- Train f^1 on training set 1
- Train f^2 on training set 2
- $\ldots$
- Can average result over B trees for a single low-variance model from $\hat{f}^1(x)$, $\hat{f}^2(x)$, $\cdots$, $\hat{f}^B(x)$ as $\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$
- But where do we come up with B Training sets?

Bagging

- Take B different bootstraps of our one dataset

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{*B} \hat{f}^{b}(x)$

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{*B} \hat{f}^{b}(x)$
- Turns out, you can grow these trees without pruning

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{*B} \hat{f}^{b}(x)$
- Turns out, you can grow these trees without pruning
- Regression - average the values from each tree

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{*B} \hat{f}^{b}(x)$
- Turns out, you can grow these trees without pruning
- Regression - average the values from each tree
- Classification - majority vote from each tree

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{*B} \hat{f}^{b}(x)$
- Turns out, you can grow these trees without pruning
- Regression - average the values from each tree
- Classification - majority vote from each tree
- Test error can be plotted as a function of B

Bagging

- Take B different bootstraps of our one dataset
- $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{*B} \hat{f}^{b}(x)$
- Turns out, you can grow these trees without pruning
- Regression - average the values from each tree
- Classification - majority vote from each tree
- Test error can be plotted as a function of B
- B is not a critical parameter (will see shortly) so large B does not mean we overfit

Out of Bag Error

- If we repeatedly fit bootstrapped subsets (say 2/3 of data)

Out of Bag Error

- If we repeatedly fit bootstrapped subsets (say 2/3 of data)
- Each time we are left with 1/3 of the data we can call out of bag

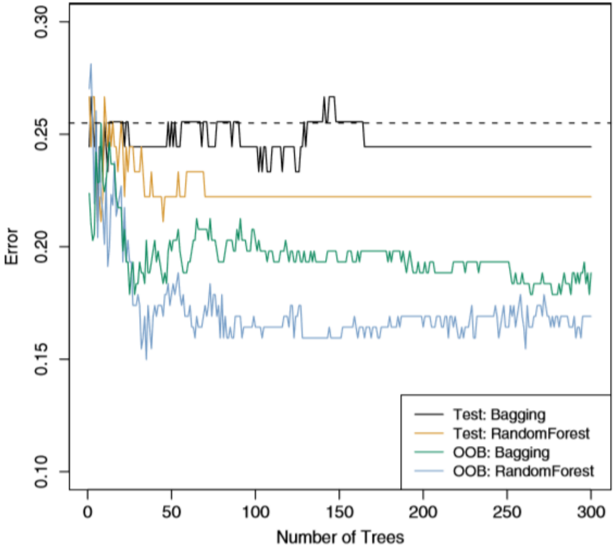
Out of Bag Error

- If we repeatedly fit bootstrapped subsets (say 2/3 of data)
- Each time we are left with 1/3 of the data we can call out of bag
- We can estimate error for this - called Out of Bag Estimation

Example: Heart Dataset

```
> summary(data)
       X              Age             Sex                 ChestPain       RestBP           Chol            Fbs        
 Min.   :  1.0   Min.   :29.00   Min.   :0.0000   asymptomatic:144   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.: 76.5   1st Qu.:48.00   1st Qu.:0.0000   nonanginal  : 86   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :152.0   Median :56.00   Median :1.0000   nontypical  : 50   Median :130.0   Median :241.0   Median :0.0000  
 Mean   :152.0   Mean   :54.44   Mean   :0.6799   typical     : 23   Mean   :131.7   Mean   :246.7   Mean   :0.1485  
 3rd Qu.:227.5   3rd Qu.:61.00   3rd Qu.:1.0000                      3rd Qu.:140.0   3rd Qu.:275.0   3rd Qu.:0.0000  
 Max.   :303.0   Max.   :77.00   Max.   :1.0000                      Max.   :200.0   Max.   :564.0   Max.   :1.0000  
                                                                                                                     
    RestECG           MaxHR           ExAng           Oldpeak         Slope             Ca              Thal    
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :1.000   Min.   :0.0000   fixed     : 18  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000   normal    :166  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :2.000   Median :0.0000   reversable:117  
 Mean   :0.9901   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.601   Mean   :0.6722   NA's      :  2  
 3rd Qu.:2.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000                   
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :3.000   Max.   :3.0000                   
                                                                                  NA's   :4                        
  AHD     
 No :164  
 Yes:139  
```

Example: Heart Dataset

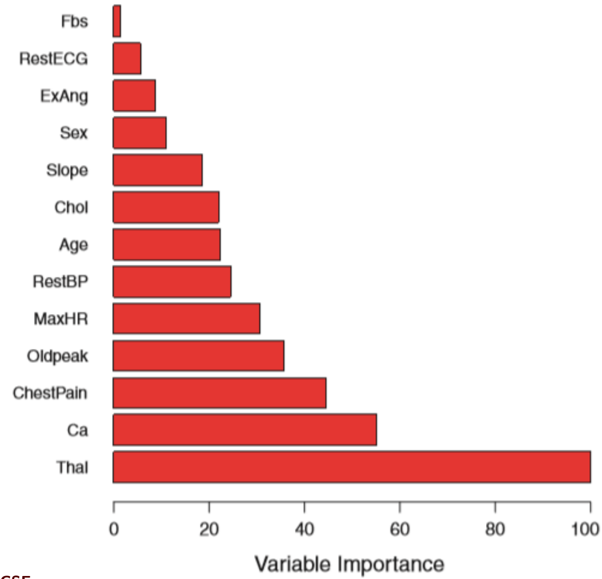


Variable Importance is Lost!

- Interpretting Bagging becomes hard
- No longer possible to decide a variable order from a single tree
- With regression trees - overall summary with reduction in RSS at each split
- With classification - overall summary in reduction in Gini Index at each split
- Relative importance of predictor variable - how often is it in trees?

$$v_j = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(j \in T_m)$$

Example: Heart Dataset



Problems with Bagging

- What if you have a strong predictor and a bunch of moderate predictors?

Problems with Bagging

- What if you have a strong predictor and a bunch of moderate predictors?
- Each time, the first variable is that strong predictor

Problems with Bagging

- What if you have a strong predictor and a bunch of moderate predictors?
- Each time, the first variable is that strong predictor
- is variance really reduced?

Problems with Bagging

- What if you have a strong predictor and a bunch of moderate predictors?
- Each time, the first variable is that strong predictor
- is variance really reduced?
- What if at each split of each tree we only consider a subset m of predictors p? (essentially - randomly eliminate the strong predictor when making some trees)

Random Forest

- set $m \approx \sqrt{p}$

Random Forest

- set $m \approx \sqrt{p}$
- Each time $\frac{p-m}{p}$ predictors aren't even considered

Random Forest

- set $m \approx \sqrt{p}$
- Each time $\frac{p-m}{p}$ predictors aren't even considered
- other predictors have a chance

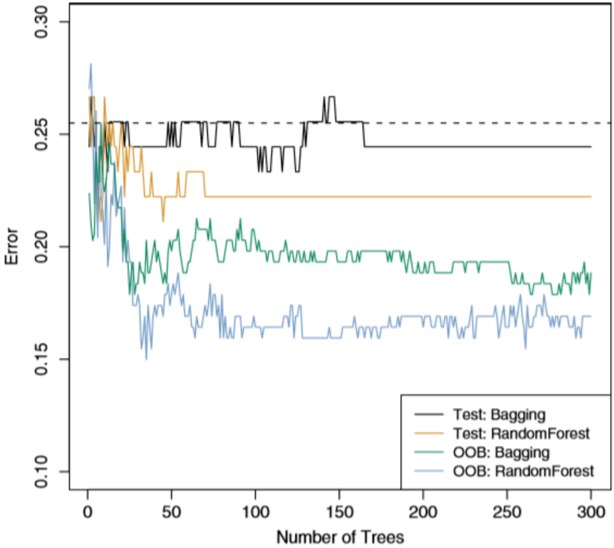
Random Forest

- set $m \approx \sqrt{p}$
- Each time $\frac{p-m}{p}$ predictors aren't even considered
- other predictors have a chance
- Turns out, this process decorrelates trees

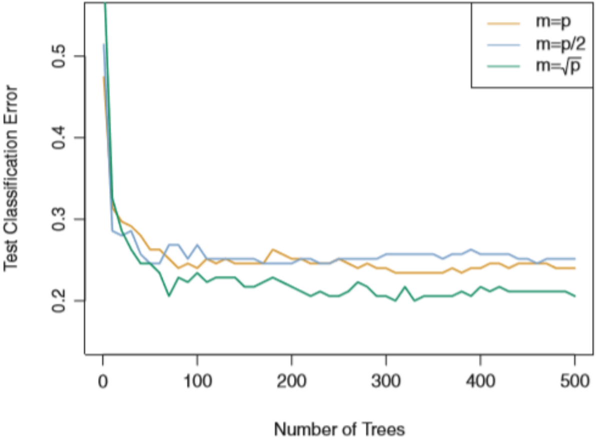
Random Forest

- set $m \approx \sqrt{p}$
- Each time $\frac{p-m}{p}$ predictors aren't even considered
- other predictors have a chance
- Turns out, this process decorrelates trees
- The average tree becomes less variable and thus more reliable

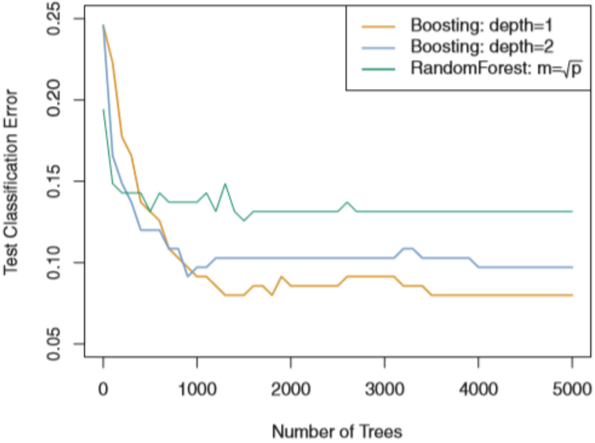
Example: Heart Dataset



RF with different m



RF with different m



What have we learnt so far

Decision Trees

- Hierarchical (tree-like) structure to perform classification/regression
- Tree structure determined by splitting criterion
  - Entropy (measure of uncertainty), gini index, etc.
- Pruning
  - Prevent overfitting by limiting the depth of the tree
  - Avoids perfect performance on train set
  - Pre/Post-pruning
- Main advantage: interpretability

Random Forests

- Tree ensemble
- Bagging & Randomization
- Good peformance in practice

Takeaways and Next Time

- Decision Trees
- Random Forest
- Next Time: Discussion: Random Forest
- Next Time: Lecture: Boosting