Lecture 23: Random Forests

Problem w/ CARTs is that they are easy to overfit (grow really large tree)

Tend to be low bias/high variance

Recap properties of means

If I have data X_n w/ the same dist each having mean μ and var. σ^2 and correlation among any two is ρ

Consider $\bar{X} = \frac{1}{N}\sum_n X_n$.

Properties: (1) $\mathbb{E}\bar{X} = \mathbb{E}\left[\frac{1}{N}\sum_n X_n\right]$

$$= \frac{1}{N}\sum_n \mathbb{E}X_n = \frac{1}{N}N\mu = \mu$$

(2) $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{N}\sum_n X_n\right)$

$$= \frac{1}{N^2}\mathrm{Var}\left(\sum_n X_n\right)$$

$$= \frac{1}{N^2}\left(\sum_n \mathrm{Var}(X_n) + \sum_{i\neq j}\mathrm{Cov}(X_i, X_j)\right)$$

$$= \frac{1}{N^2}\left(\sum_n \sigma^2 + \sum_{i \neq j} \sigma^2 \rho\right)$$

$$= \frac{1}{N^2}\left(N\sigma^2 + N(N-1)\sigma^2\rho\right)$$

$$= \frac{\sigma^2}{N} + \frac{N-1}{N}\sigma^2\rho$$

$$= \frac{\sigma^2}{N} + \sigma^2\rho - \frac{\sigma^2\rho}{N}$$

$$\boxed{Var(\bar{x}) = \sigma^2\rho + \frac{\sigma^2}{N}(1-\rho)}$$

If $\rho = 0$ then $Var(\bar{x}) = \sigma^2/N$.

Bagging: Ensemble Method

Bootstrap Aggregating

combine multiple methods into a better one

① Draw a series of bootstrap samples

Assume I have traing data $\{(x_n, y_n)\}_{n=1}^{N}$

Draw B bootstrap samples

For $b = 1, \ldots, B$

I draw a (sub-)sample of N of traing points w/ replacement.

Call these bootstrap samples $S_1, S_2, \ldots, S_B$

② Train a method on each sample S_b

For $b = 1, \dots, B$

$$\hat{f}_b = \text{method fit on } S_b$$

③ Combine these $\hat{f}_b$ to make a bagged overall method $\hat{f}$

(i) Regression: $\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$

(ii) Classification: $\hat{f}(x) =$ most common predicted class among $\hat{f}_b(x)$ (plurality)

Why does this work?

For regression

$$MSE(\hat{f}) = Bias(\hat{f})^2 + Var(\hat{f})$$

Bias of my bagged estimator

$$Bias(\hat{f}) = \mathbb{E}[\hat{f}(x)] - y$$

$$= \mathbb{E}\left[\frac{1}{B}\sum_b \hat{f}_b(x)\right] - y$$

If each $\hat{f}_b$ has the same bias then

$$= \mathbb{E}[\hat{f}_b(x)] - y$$
$$= \text{Bias}(\hat{f}_b)$$

My bias unchanged by Bagging.

If $\text{Var}(\hat{f}_b) = \sigma^2$ and $\text{cor}(\hat{f}_b, \hat{f}_{b'}) = \rho$
then

$$\text{Var}(\hat{f}) = \rho\sigma^2 + (1-\rho)\frac{\sigma^2}{B}$$

If we can build these $\hat{f}_b$ so they're approx. uncorrelated, then $(\rho \approx 0)$

$$\text{Var}(\hat{f}) = \sigma^2/B$$

So bagging keeps bias unchanged and reduces variance thus

$$\text{MSE} = \text{bias}^2 + \text{Var}$$

goes down.

$\{1, 2, 3, 4\}$

$\{1, 1, 3, 4\}$, $\{2, 3, 3, 4\}$

$\hat{f}_1$ $\hat{f}_2$

Works best if applied to a method w/ low bias but high variance.
B/c I can reduce the var. through bagging.

Random Forest : Basically bagged set of decision trees.

RF algorithm:

(1) Fit B trees

For $b = 1, \dots, B$

help reduce correlation among individual trees

(i) Draw bootstrap sample from training data

(ii) Fit a CART on the bootstrap sample — but each time I make a split in my tree I consider splitting on a random subset of vars.

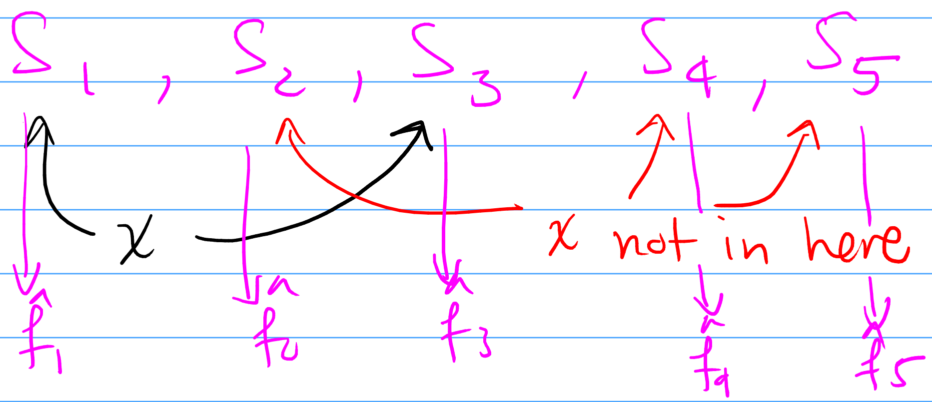
(ii) helps avoid splitting on the same vars in diff. trees

(2) bag the trees

Out-of-Bag Error = basically an estimate of my test error

When I generate my bootstrap samples

For any training point x – x is in some bootstrap samples but not others



Consider bagging only those $\hat{f}$s not containing x in their training sample $\rightarrow \hat{f}_{-x}$

For these $\hat{f}$s $(\hat{f}_{-x})$ x is essentially a test point – not used to train them

So I can predict the corresp. y as $\hat{y}^{OOB} = \hat{f}_{-x}(x)$

So if I do this for all training pts $\hat{y}^{OOB}_n$

I can get a essentially test err by calc. err of these OOB ests.