Lecture 20 - Bagging and RFs

Last time: trees (CARTs)

classification and regression trees

Categorical Variables

Consider a q-level categorical variable, a "split" is dividing categories into two groups

$$1, 2, 3, \dots, q$$

split

$\boxed{2, 3, 5}$ $\quad$ $\boxed{1, 7, q}$

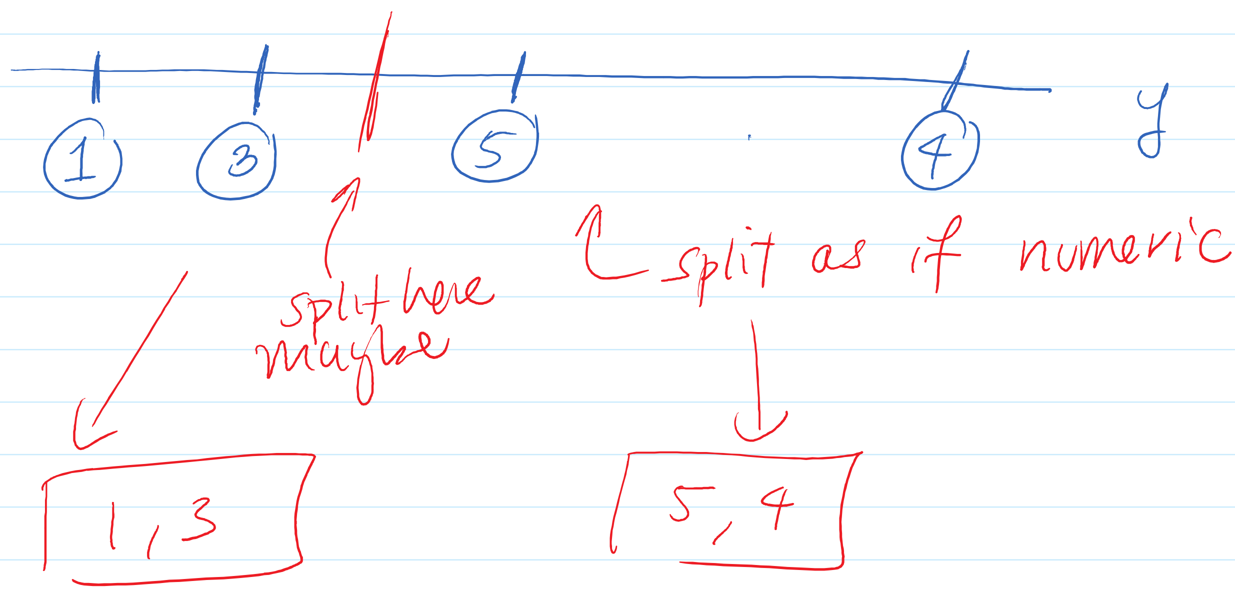
divide categories into two groups

Problem: If I have q levels then I have $2^{q-1} - 1$ possible "splits"

problem if q is large.

Soln:

If y is numeric (regression problem) I can order all my q levels by their mean value of y



tune at this find the optimal spit.

Classification tree

→ Binary classification:

can do similar procedure to regr. tree

Order by % in class 1.

This finds optimal split for Gini or Entropy.

→ Multi-class: no such trick.

Warning: trees tend to like to split on

Categorical vars w/ many levels.
Easy to overfit. Careful.

Missing Data : handle missing data well.

Categorical vars: add a "missing category"

Numeric: Keep track of "surrogate splits"

i.e. splits using other vars that give similar divisions.

↳ idea: use surrogate split from another var if I'm missing values for one.

Problem w/ CARTs: high variance but low bias. i.e. easy to overfit.

Quick Recap of $\bar{X}$.

If X_n are from same dist. w/ a mean μ and a variance σ^2. $Var(X_n) = \sigma^2$ $\mathbb{E}[X_n] = \mu$

Consider: $$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

Properties:

Properties:

① $\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n]$

$= \frac{1}{N}\sum_{n=1}^{N}\mu = \frac{1}{N}N\mu = \mu.$

$$\rightarrow \boxed{\mathbb{E}[\bar{X}] = \mu}$$

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

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

$= \frac{1}{N^2}\left[\sum_{n=1}^{N}\underbrace{\mathrm{Var}(X_n)}_{\sigma^2} + \sum_{n\neq n'}\underbrace{\mathrm{Cov}(X_n, X_{n'})}_{\sigma^2\rho}\right]$

If $\mathrm{Cor}(X_n, X_{n'}) = \rho \Leftrightarrow \mathrm{Cov}(X_n, X_{n'}) = \sigma^2\rho$

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

$$* \quad \boxed{= \frac{\sigma^2}{N} + \frac{(N-1)\sigma^2\rho}{N}}$$

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

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

$$\circledast \quad \boxed{= \rho\sigma^2 + \frac{\sigma^2}{N}(1-\rho)} \longleftarrow$$

If $\boxed{\rho = 0}$ then $\boxed{\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N}}$. $\longleftarrow$

Baggins: (Ensemble Method)

(1) Draw B samples (w/ replacement) from my training data: $\{(x_n, y_n)\}$

$$S_1, S_2, S_3, \ldots, S_B$$

each of size N

(2) Train a method $\hat{f}_b$ on each sample S_b

for $b = 1, \ldots, B$

(3) Combine $\hat{f}_b$s to make a Bagged method $\hat{f}$

(i) Regression:

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

(ii) Classification:

$\hat{f}(x) = \text{plurality class among all } \hat{f}_1(x), \hat{f}_2(x), \dots, \hat{f}_B(x)$

$= \text{majority vote among } \hat{f}_b.$

---

why does this help?

For regression

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

① $\underbrace{bias(\hat{f}(x))} = E[\hat{f}(x)] - y$ basically $\bar{X}$

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

$$= E[\hat{f}_b] - y = \underbrace{bias(\hat{f}_b)}$$

bias is unchanged

Bagging doesn't change bias

② $Var(\hat{f}(x)) = Var\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\right)$

$$\boxed{= \rho\sigma^2 + (1-\rho)\frac{\sigma^2}{B}}$$

$$\boxed{\qquad \frac{\ }{B}}$$

when $\mathrm{cor}(\hat{f}_b, \hat{f}_{b'}) = \rho$

and $\mathrm{Var}(\hat{f}_b) = \sigma^2$

If we can choose $\hat{f}$s so that $\boxed{\rho \approx 0}$

then

$$\mathrm{Var}(\hat{f}) \approx \frac{\sigma^2}{B} = \frac{\boxed{\mathrm{Var}(\hat{f}_b)}}{B}$$

reduced the variance

So

$$\boxed{MSE = \underbrace{bias^2}_{\text{leaves unchanged}} + \underbrace{Var}_{\text{reduces this}}}$$

Idea: (1) choose a method w/ low bias but high variance

(2) reduce variance through bagging.

Random Forest: Bagged CART.

RF Algo:

① Fit B trees:

For $b = 1, \dots, B$

helps make $\rho \approx 0$

ⓘ draw a subsample from training data

ⓘⓘ grow a CART on subsample but each time I split I consider a random subset of covs to split on.

② bag them together for prediction.

Classification: A little more complicated. Bagging good trees helps. Bagging bad trees can (potentially) hurt.

Out-of-Bag Error (OOB)

When I fit a RF it trains $\{\hat{f}_b\}$ on subsamples.

① For any $\hat{f}_b$ there are some training data not used to train it

② Flip side: For any particular training pt (x,y) there are some $\hat{f}_b$s that don't use it.

Idea: For some (x,y) I bag only $\hat{f}_b$s that don't use (x,y) to train them, then my train pt (x,y) is basically a test pt as far as this new bagged method goes

Can do: predict a test err for (x,y) predicting using OOB methods.

OOB error: do this for each point and calculate test error this way.

↖ an est. of test err.