Lecture 18 -- Random Forests

Thursday, November 7, 2024 3:29 PM

To determine splits for a classification tree:

try to decrease node impurity

Impurity Measures

(1) misclassification rate:

$1 - p_{\hat{k}}$, $\hat{k}$ = maj. class in node

pct. in maj. class

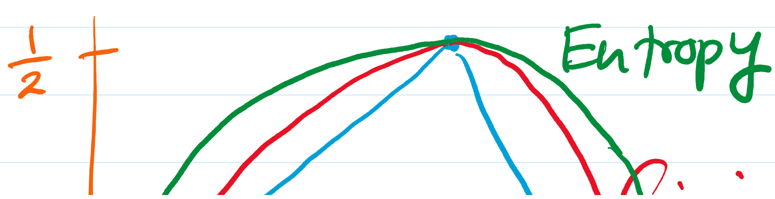
(2) Gini Index:

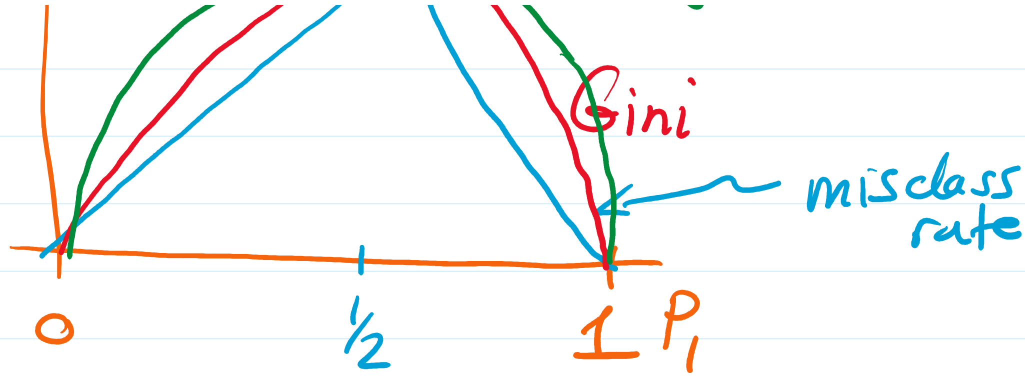
$$\sum_k p_k (1 - p_k)$$

pct in class k

(3) Entropy: $\sum_k p_k \log(p_k)$

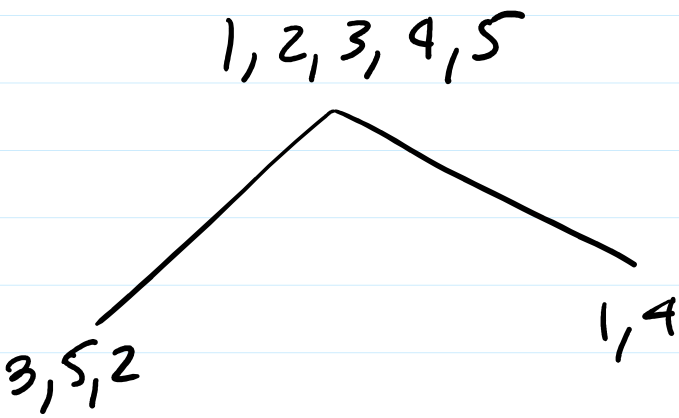
K = 2





Categorical Vars

Splitting a cat var is just dividing cats into two groups



If I have q levels then there are $2^{q}-1$ possible splits.

Missing Values

CARTs can deal with missing values

CARTs can deal with missing values very nicely.

Cat vars just add "missing" category

numeric vars : keep track of "surrogate" splits that divide data similarly

---

Problems w/ CARTs

they are really easy to overfit

Tend to be low bias, high variance

Recap of Means

If I have X_ns all have same mean μ and variance σ^2

and if they are pairwise correlated w/ correlation ρ

Consider $\bar{X} = \frac{1}{N} \sum^{N} X_i$

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

① $E[\bar{X}] = E\left[\frac{1}{N} \sum_n X_n\right]$

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

$= \mu$

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

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

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

$Cov(X_i, X_j) = \sigma^2 \underbrace{Cor(X_i, X_j)}_{\rho}$

$\downarrow$

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

$\vdots$

$$\vdots$$

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

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

If $\rho = 1$ then $Var(\bar{x}) = \sigma^2$

Bagging: Ensemble Method

combing many methods together

Bootstrap Aggregating

(1) Draw a series of bootstrap sample from traing data

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

Sample B bootstrap samples

For $b = 1, \ldots, B$

$S_b \leftarrow$ sample w/ replacement of N train pairs from traing

$S_b \leftarrow$ sample w/ replacement of N traing pairs from traing data

② Train a collection of methods on each resample (bootstrap sample)

For $b = 1, \ldots, B$

$\hat{f}_b \leftarrow$ ML method fit on S_b

③ Combine these methods together

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

(ii) Classification : $\hat{f}(x) =$ most common pred. class among all $\hat{f}_b(x)$ [plurality]

Binary: if $\hat{f}_b(x) \in \{-1, 1\}$

Binary: If $f_b(x) \in \{-1, 1\}$

$$\hat{f}(x) = \text{sign}\left(\sum_{b=1}^{B} \hat{f}_b(x)\right)$$

Why is this reasonable?

For regression

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

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

For bagging:

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

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

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

same $\forall b$

(some σ_B

$$= \frac{1}{B} B\, E[\hat{f}_1(x)] - f(x)$$

$$= E[\hat{f}_1(x)] - f(x)$$

$$= \text{Bias}(\hat{f}_1(x))$$

Bias Ensemble = Bias of individual

So bagging doesn't increase bias

However,

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

$$\left[\sigma^2 = Var(\hat{f}_1),\ \rho = Cor(\hat{f}_i, \hat{f}_j)\right]$$

and so if we can make $\rho \approx 0$

then $Var(\hat{f}) \approx Var(\hat{f}_1)/B$.

So bagging reduces variance.

So bagging reduces variance.

<u>Summary</u>: bagging

(1) leaves bias unchanged

(2) reduces variance

(more if constituent $\hat{f}_b$ are uncorrelated)

So to make a really good bagged method I want my $\hat{f}_b$ to be <u>low bias</u> and <u>high variance</u>.