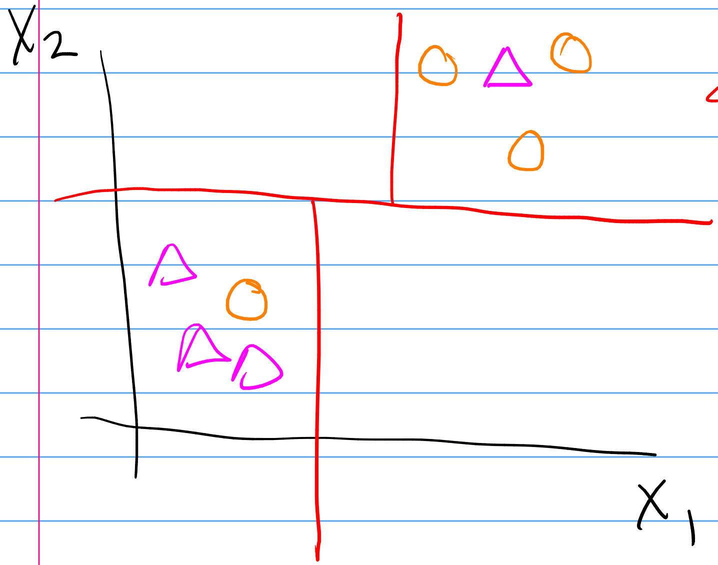
Lecture 20: Classification Trees and RFs



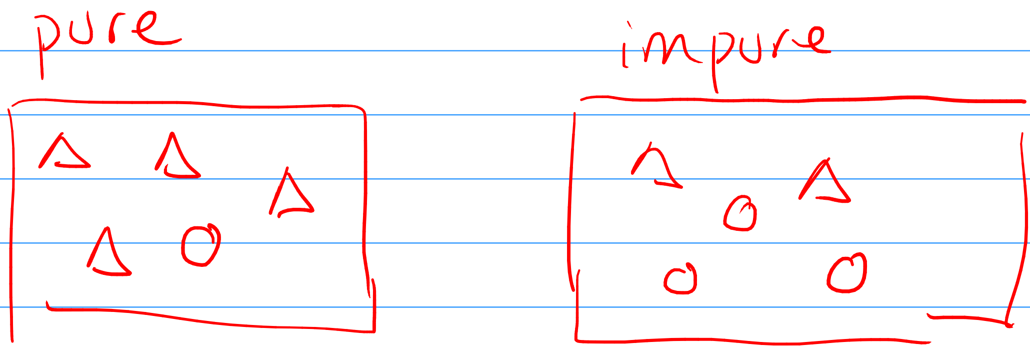
in each rectangle, predict majority class

What makes a good split for class tree?

Regression ⇝ RSS

Classification ⇝ reduce node impurity
increase node purity

Ex.



Node impurity measures: p_k = pct. of class k in my rectangle

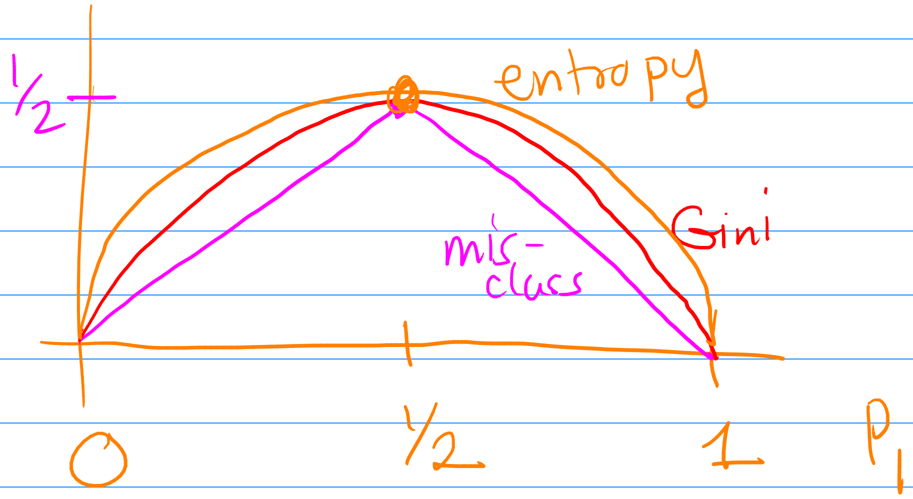
(1) mis-class rate: $1 - p_{\hat{k}}$, $\hat{k} = \arg\max_k p_k$
= maj. class

② Gini-Index

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

③ Entropy: $\sum_k p_k \log p_k$

$K=2$



Categorical Vars

Splitting a cat var is just dividing cats into two groups

1, 2, 3, 4, 5

3, 5, 2   1, 4

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

CARTs can deal w/ missing data very nicely.

cat. vars, just add a "missing" category

numeric vars. – Keep track of "surrogate" splits using other vars that divide the data similarly

---

Problem w/ CART they are really easy to over-fit

Tend to be low bias, high variance.

Recap of properties of means

If I have X_n all w/ the same mean μ, and same variance σ^2.

Let ρ the correlation among them.

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

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

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

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

$$= \frac{1}{N^2}\Big[\sum_n \underbrace{\mathrm{Var}(X_n)}_{\sigma^2} + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)\Big]$$

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cor}(X_i, X_j) \cdot \sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)} = \rho\sigma^2$$

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

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

$$= \ldots$$

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

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

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

Bagging: Ensemble Method

Bootstrap Aggregating

combining many learning methods

① **Draw a series of Bootstrap samples**

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

Sample B bootstrap samples

For $b = 1, \ldots, B$

draw a sample of N training pts w/ replacement

Call these samples $S_1, \ldots, S_B$

② **Train a series of methods on each resample**

For $b = 1, \ldots, B$

$\hat{f}_b$ = method fit to S_b

3 combine thes to make a ensemble $\hat{f}$

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)

Binary:

If $\hat{f}_b(x) \in \{\pm 1\}$

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

Why does this work?

For regression

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

$$bias(\hat{f}) = \mathbb{E}\hat{f} - f$$

For bagging in regression

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

$$= \mathbb{E}\left[\hat{f}_b(x)\right] - f(x)$$

$$= bias(\hat{f}_b)$$

So bagging doesn't change bias.

However,

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

$$\left[\mathrm{Var}(\hat{f}_b) = \sigma^2, \quad \mathrm{Cor}(\hat{f}_b, \hat{f}_{b'}) = \rho\right]$$

and so if $\rho \approx 0$ then

$$\mathrm{Var}(\hat{f}) \approx \mathrm{Var}(\hat{f}_b)/B.$$

So bagging reduces variance.

So to make bagging effective we need

1. to build $\hat{f}_b$ that are $\approx$ uncorrelated
2. want to bag $\hat{f}_b$ that are high var and low bias (e.g. trees!)

Random Forest : basically bagged set of trees (bagged randomized trees)

RF: algo

(1) Fit B trees (randomized trees)

For $b = 1, \dots, B$

help make ρ between trees smaller

(i) bootstrap sample from training data

(ii) Fit a randomized tree to this bootstrap sample — at each point where I split I only consider a random subset of variables

fit a huge tree

(2) bag the trees

Out-of-bag Error (OOB)

estimate of my test error

When I bootstrap sample, for any point x it will end up in some of my bootstrap samples, and not others

x not in these samples

$S_1 \quad S_2 \quad S_3 \quad \dots \quad S_B$

x is in these sample

Consider bagging only those trees trained on S_b not including x : $\hat{f}_{-x}$

As far as $\hat{f}_{-x}$ is concerned, x is a validation point and so

$$\hat{y}_{OOB} = \hat{f}_{-x}(x)$$

then $y - \hat{y}_{OOB}$ is an est of my test/val error for y

So if I do this for all pts, I can est test err using these OOB predictions.
