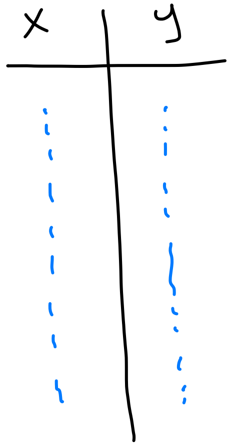
CS 307

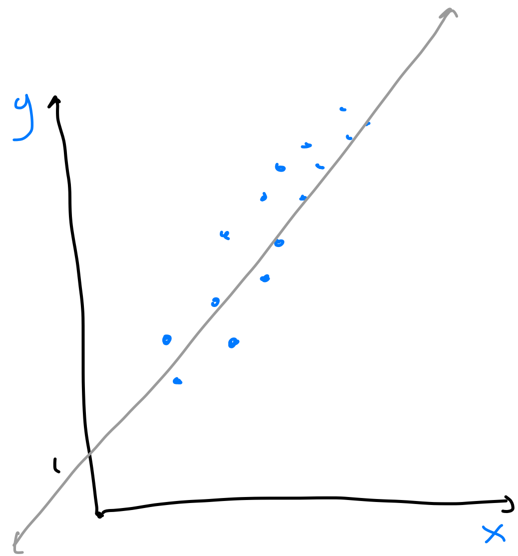
Fall 2023

Dalpiaz

$\boxed{\text{Week 05}}$

Some Linear Regression Review





Assume

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

"Fit"

"least squares"

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

"Predict"

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

"Multiple" Linear Regression

y	x_1	x_2	....	x_p

Assume

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

"Fit"

$$\min_{\beta} \sum_{i=1}^{n} \Big(y_i - \big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \big) \Big)^2$$

$$\min_{\beta} \sum \Big(y_i - \big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \big) \Big)^2$$

"Predict"

$$\hat{f}(\underline{x}) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

Best Subset Selection

p features, linear regression

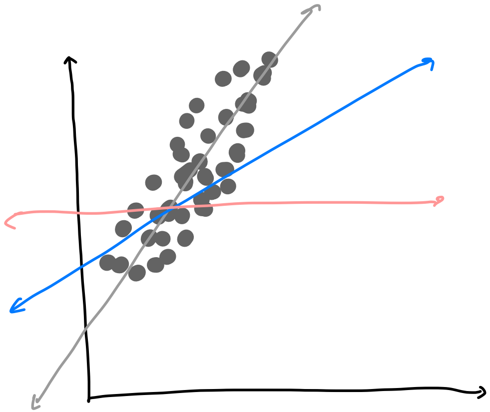
# Features	# Models	Models
p	1	$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$
$p-1$	$\binom{p}{2}$	TOO MANY TO LIST
$\vdots$	$\vdots$	
2	$\binom{p}{2}$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$, ...
1	p	$Y = \beta_0 + \beta_1 x_1 + \varepsilon$, $Y = \beta_0 + \beta_2 x_2 + \varepsilon$, ..., $Y = \beta_0 + \beta_p x_p + \varepsilon$
0	1	$Y = \beta_0$
	2^p	

To save on computation → "Search"

Step Direction + **Metric**

Step Direction	Metric
Backwards →	AIC → Default in R's step( )
Forwards	BIC
Both	Cp
	Etc

Related to LOOCV asymptotically



TRUE MODEL $Y = 2 + 5x + \varepsilon$

LEAST SQUARES

BIAS

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2 \longrightarrow \hat{B}_1 = 5.2$$

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2$$

SUBJECT TO $|B_1| < 3 \longrightarrow \hat{B}_1 = 3$

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2$$

SUBJECT TO $|B_1| < 0 \longrightarrow \hat{B}_1 = 0$

VAR

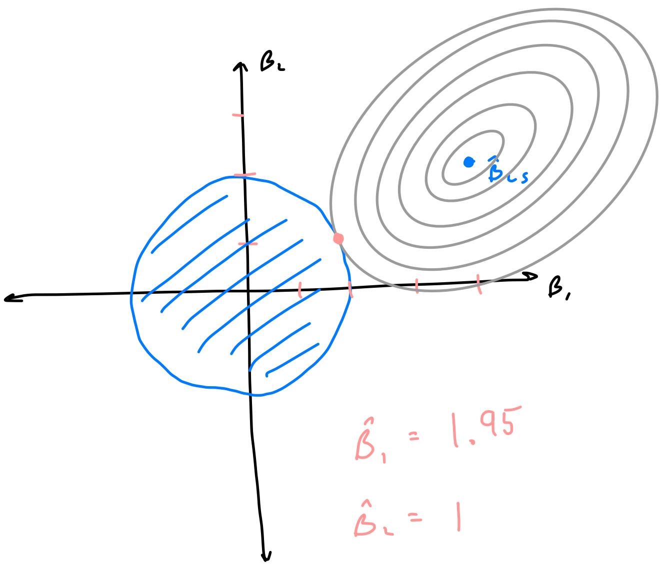
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad \hat{\beta}_1 = 4, \quad \hat{\beta}_2 = 2$$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \right)^2$$

SUBJECT TO $\beta_1^2 + \beta_2^2 \leq 4$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) \right)^2$$

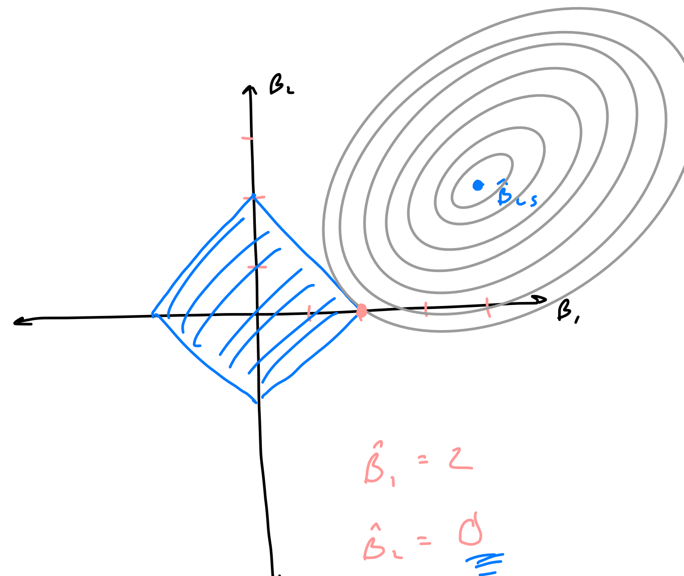
SUBJECT TO $|\beta_1| + |\beta_2| \leq 2$



$\hat{\beta}_1 = 1.95$

$\hat{\beta}_2 = 1$

RIDGE



$\hat{\beta}_1 = 2$

$\hat{\beta}_2 = 0$

LASSO

LS / ERROR   CONSTRAINT

?

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \le s$$

RIDGE

"BUDGET"

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s$$

LASSO

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \ne 0) \le s$$

BEST SUBSET SELECTION

RIDGE AND LASSO ARE GREAT WHEN p IS LARGE

LASSO DOES SELECTION!

$$\min \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

$s = 0 \longrightarrow \beta_0$

$s = \infty \longrightarrow \text{OLS}$

$s \longleftrightarrow \lambda$

$$\min \left[\underbrace{\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2}_{\text{ERROR}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{PENALTY}} \right]$$

TUNING

TRADEOFF

$\lambda = 0 \longrightarrow \text{OLS}$

$\lambda = \infty \longrightarrow \beta_0$

AS LS ERROR ↓, PENALTY ↑

A note about scaling

y	X	X^*
⋮	900,000	-1
⋮	⋮	⋮
⋮	1,000,000	0
⋮	⋮	⋮
⋮	1,100,000	1

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \hat{\beta}_1 = 0.001$$

$$Y = \beta_0 + \beta_1 x^* + \varepsilon \qquad \hat{\beta}_1 = 1000$$

$$x^* = \frac{x - \bar{x}}{SD[x]}$$

Big effect on

$$\sum_{j=1}^{p} \beta_j^2$$

$X_1, X_2, \ldots X_n$ FROM SOME DISTRIBUTION

$X_1, \ldots X_n$ IND

$$\mathbb{E}[X_i] = \mu$$

$$\mathrm{Var}[X_i] = \sigma^2$$

$$\bar{X} = \frac{1}{n} \sum X_i$$

"SAMPLE MEAN"

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

$$\mathrm{Var}[\bar{X}] = \sigma^2/n + [\text{COVARIANCE TERM}]$$

IF NOT IND

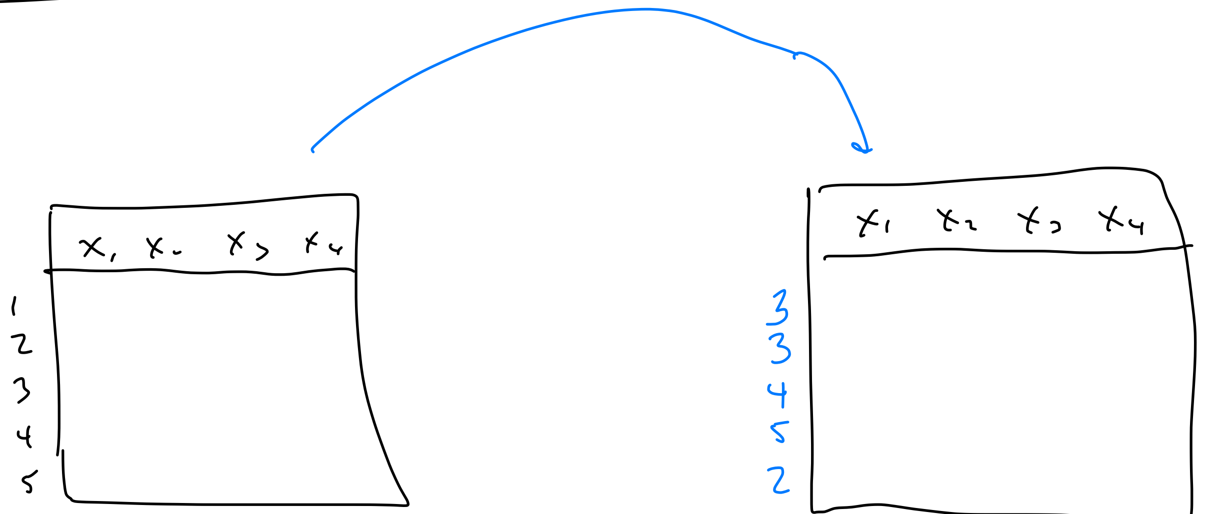
ENSEMBLE METHODS

- FIT MANY MODELS TO DATA

BASE LEARNERS

- MAKE PREDICTIONS FROM EACH
- AVERAGE PREDICTIONS

Bootstrap



Original Data

Bootstrap resample

w/ replacement

same shape

BAGGING

w/ training data

Fix some number of resamples, B

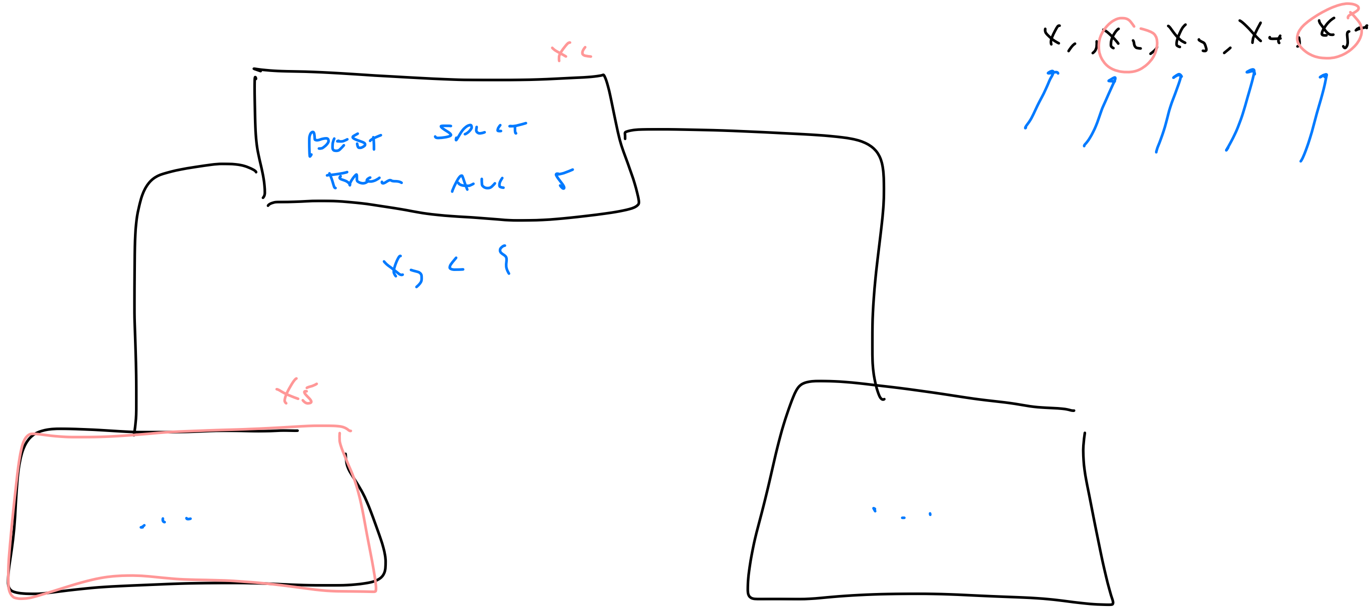
For each B

- Create bootstrap resample ← Random
- Fit base learner ← Same for each

Add randomness here

To make prediction

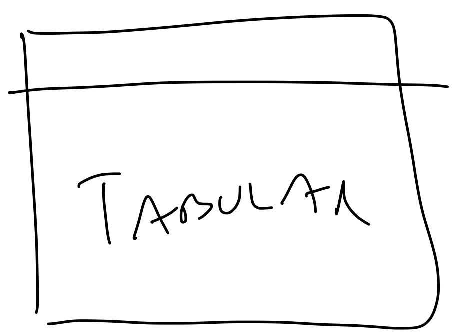
- Make prediction w/ each learner
- Ave predictions



MAX FEATURES = 1

DEEP LEARNING IS MOSTLY HYPE

KAGGLE IS UNHELPFUL



WIN W/ ENSEMBLES

RF

BOOSTING → xgboost