Lec: 10
01.18.24

"Resampling Method"

2 types of Resampling:
i) cross-validation
ii) Bootstrap

overfitting: memorizing but not learning → doing bad
underfitting: complex model
optimum fit: between over and underfitting

① Cross-validation
* validation set approach
(Test-set, val-set, train-set)

Training
↓
enough data to learn
80%
70%
90%

Testing
20%
30%
10%

↓
validation set

validat^{n} - set → we will use to test the model
During training period

Trai	val set	test set
70 :	10 :	20

training → 200 times
first step ↳ checking / monitoring
iteratn ↳
যদি ভালো না হয়, then model ↳
change needed

classical → Logistic and Linear
training process time is less
multiple iteratn

validat^{n} is applicable in deep-NN.

Advantages: faster, simple and easy to understand

Disadvantages: (i) split reduced the size of the data (Losing data)
(ii) Training and testing error
(3) " random split → easy
(iii) sample in the test data
overestimate

2. K- fold validation:

instead of splitting

We divide data according to k value.

Model 1 is going to be tested with 2, 3, 4, 5 and validated by (1)

↓

Avg accuracy of all 5 - model

If it is classificatn → 3 cat 2 dog

↓

cat

variant of k-fold validatn:

LOOCV

↓

Leave one out cross-validation

↓

Splitting data according to the no of sample.

If n = 6

M. [][][][][][] Test with (1)

Disadvantages → slow

Advantages → → accurate

→ all the data are getting tested

best value of k = 3, 5, 7, 10
5 and 7 → Good accuracy

5 models → unique test
rest train
LOO → (one is getting left-out)

(2) Bootstrap approaches:
↓
General statistical technique
(Decision tree)

we want to quantify an associat" between estimation

↔ Sampling with replacement
randomly select"
↓
then new dataset

disadvantage: → same sample can be in the same dataset
→ Performance low
might not reflect the real data
→ underestimate predict"

How to remove overlapping?

sampling without replacement

Summary: Resampling

↓

cross-validation (2 method)

↙ k-fold validation

↘ Bootstrap

↓

used in decision trees → not good

** in resampling

Lec: 12
02-26-24

Model select^n and Regularizat^n

#

Dataset

	x_1	x_2	...	x_p	y
1					
2					
...					

Resampling method
↓
focuses on sample

But in model select^n
↓
Focusing on the predictors

{Subset select^n
shrinkage}

$$\{Y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p\}$$

Dataset → more predictors than the sample

$p = 50$
$n = 20$
} Difficult to control the variance
↓
Feature select^n needed
↓
i) Subset-select^n method:
selecting some of the predictors:

Best subset selection:

1st technique:

Best

① Model with no predictors

↓

null model

$M_0 = \beta_0$ (intercept / mean of resp)

② fitting all combinat^n of p and k

$k=0$

$k=1$ $M_1 = \beta_0 + \beta_1 X_1$

$M_2 = \beta_0 + \beta_1 X_2$

$M_3 = \beta_0 + \beta_1 X_3$

x_1	x_2	x_3	y

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X$

$k=2$ $M_4 = X_1 + X_2$

$M_5 = X_1 + X_3$

$M_6 = X_2 + X_3$

} finding the best one

↓

if this, $k=3$

$k=3$ $M_7 = X_1 + X_2 + X_3$

③ Step 3 → finding best model.

M2, M6 and M7 comparison by cross-validatn

↓

selecting the best one

P = 50 predictrs

k = 1 (all possible combindn)

↓

dis: i) Time consuming

Advantage: i) Best possible model is guaranteed (as we are trying all possible combindn)

second technique of Subset-selection:

Getting ideas from sequential moveing

A lot of predictors

↓

so overfitting may occurs

↓

so instead of moving at a time

— Best subset selectn
= forward stepwise select

① $M_0 = \beta_0$

first predictor X_1 →

$M_1 = \beta_0 + \beta_1 X_1$

$M_2 = \beta_0 + X_2$ → if accuracy ↓↓ than M1 ↓ moving to M_3

$M_3 = \beta_0 + X_3 + X_1$

if this is good $k = 1$

$M_4 = \beta_0 + X_3 + X_2$

$k = 3$

$M_5 = \beta_0 + X_3 + X_2 + X_1$

if this is good $k = 3$

① $M_0 = \beta_0$

② $M_1 = \beta_0 + X_1$ $k = 1$

$M_6 = \beta_0 + X_2 + X_1 + X_3$

not at this point

M4 is good we found

Step 3 :

M_2 vs M_4

↓

By cross-validation

↓

best model select^n

Advantage: Best subset select^n ↑↑ good

Backward stepwise selection:

{Model with all predictors}

① $M_p = \beta_0 + X_1 + X_2 + X_3$

② $k = P - 1, \quad k = 2$

better ✓ $M_1 = \beta_0 + X_1 + X_2$ (without 1 Predictor from M_p)

~~M_2 M_2 $X_1 + X_3$~~

~~$M_3 = X_2 + X_3$~~

$M_3 = X_2 + X_3$

$k = 1,$

$M_4 = X_1$ ✓ better

~~M_3~~ $= X_2$

$k_0 \quad M_0 = \beta_0$

(3) M_1 vs M_4 y cross-validation

↓

best select

(i) Dis [fitting all the predictors at a time → not viable]

[Differences are in forward and backward processes]

(ii) Dis might loss the best one model

Shrinkage methods:

Lec: 13
02-28-24

we need essential predictors
↓
Systematic way
↓
3 - classes are:

i) ~~Best~~ subset selection | Shrinkage | Dimension Reduct

- Best subset
- forward stepwise
- Backward "
- Hybrid "

Hybrid : X_1 X_2 X_3 X_4

Evaluation criteria

forward → " → then backwards →
evaluation → then forward

F → E → B → E → F → E → B → E → f

E → ✓ R^2 values and b B)

F : $y = X_1$ ✓ | $y = X_2 + X_3$ ✓ | $y = X_2 + X_3 + X_4$
$y = X_2$ ✓ | $y = X_2 + X_4$ | $y = X_2 + X_3 +$ X_1 ✓
$y = X_3$ | $y = X_2 + X_1$ |
$y = X_4$ |

M_1 | M_2 best. | M_3 best

p values → x_2, x_3, x_4 → > 0.03

R^2 → >

If the model meets the criteria
↓
then stop
↓
Otherwise we will go backwards
↓

$y = x_2 + x_3$ ✓	$y = x_2$
$y = x_1 + x_2$	$y = x_3$ ✓
$y = x_1 + x_3$	
M_1	M_2

→ Evaluation → If good
↓
Then stop

Shrinkage ~~method~~ :
Forward

Problem solving quer
Time ↓↓ → forward
Accuracy → Hybrid
If we have Time → Best subset

sales target → Dec : 2nd
target → next year

time ~~2nd~~ 2(n)
↓
forward ~~for~~ best
Accuracy → Hybrid
↑
in case of

Best subset
↓
best model is guaranteed
↓
if plenty time ~~2nd best~~

*Limited time and resources
↓
backward won't be a good model

Shrinkage method:

↓

Get rid of unnecessary predictors

↓

those that aren't statistically significant

↓

we want to shrink the less coefficient values.

model 6

→ Ridge

→ Lasso

Ridge regression →

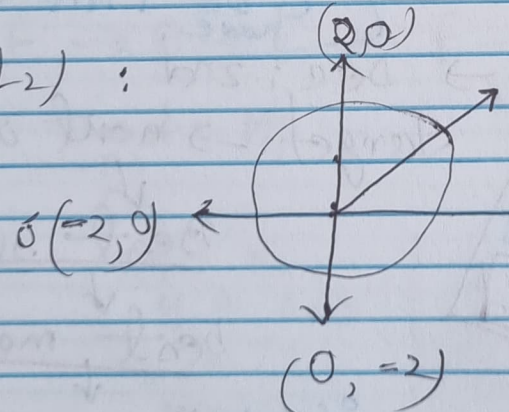
$y = \beta_0 + \beta_1 x_1$ | $y_2 = \beta_0 + 5.6(x_1)$

$x_1 = 0.5$

$\beta_1 = 0.5$, $P > 0.05$

$\beta_2 = 5.6$ → significant impact

Ridge (L_2) :



$L_2 \text{ norm} = \sqrt{\beta_0^2 + \beta_1^2}$

Shrinking the coefficient.

Disadva : → still have all the Predictors

Lasso :

(L1) shrinking the co-efficient to 0

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$\beta_1 = 0$

$y = \beta_0 + \beta_2 X_2$

pushing towards 0 → very small numbers will get to the zero

but in ridge → 0 এর কাছাকাছি যাবে but not exactly 0.

* Dimension Reduction methods :

↓

Before the creation of model.

↓

feature selection

(1) rows × columns → 4/3

p = 50, n = 200	n × p, 200 × 50

We want to ↓ the number of predictors columns

= Data এর variability কিভাবে হ্রাস predictor removal.

200 × 50
200 × 20

* PCR (Principal components Regression):

PCA:

way of putting together → predictors

step 1: Normalize the data

↓

same range in terms of number

① Age

200 → Highest

0 → newborn

Salary (per year)

100,000

0

Model will এটিকে significant but age কে underestimate করতে

এই জন্য numerical value কে normalize করা,

2nd step 2: finding the largest variance

↓

highest co-efficient of y

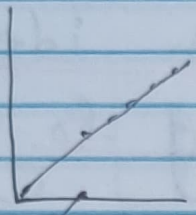
↓

Principle components finding

wei	height	age	salary

↓

PC1	age



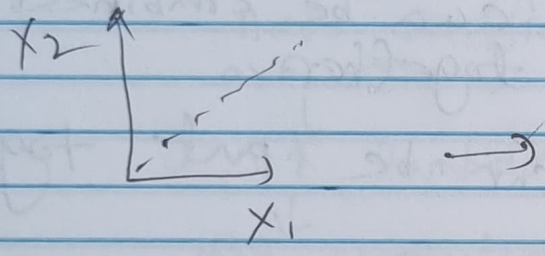
→ 2 - features highly correlated

↓

will be put together

↓

to form one component



will be the same

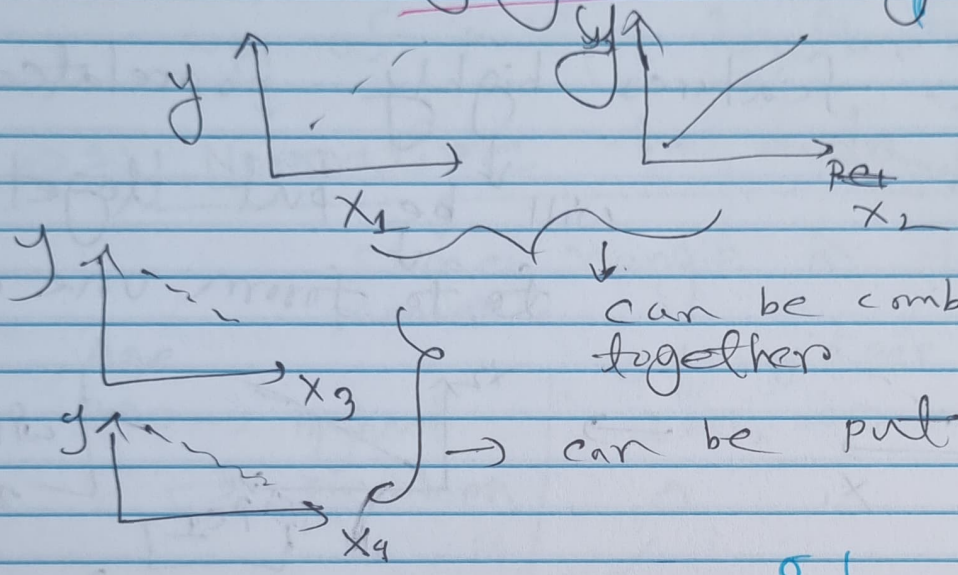
↓

we are not missing any important data

↔ Partial Least sq :

PCA → we do not use y

* We go through the PCA idea while paying attention to y

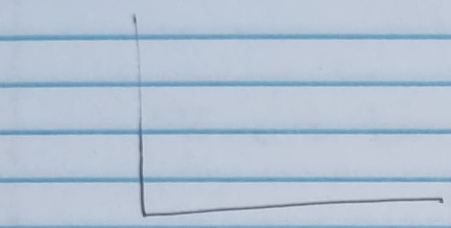


PCA → Unsupervised

PLS → Supervised

Lec: 14
03-04-24

"coding"



↑↑ log value (log(100)
provided → log(10)

−10 −5 0

Here coefficients are 0

↑ lambda

avg coefficients will be lower as its punishing the model to shrink the co-efficient

normalizing the data and providing to model

Lasso 'n' → stable
ridge → stable

Grid-search
↓
best model searching with the best lambda value

default → PCA → (3)

CV MSE:
6 com and 17 com
↓
sgood
↓
but 6 com
↓
otherwise fine↑↑

PCA	PLS
6	12
Accuracy + efficiency	have the computation power
+ not enough computation	