Linear Regression

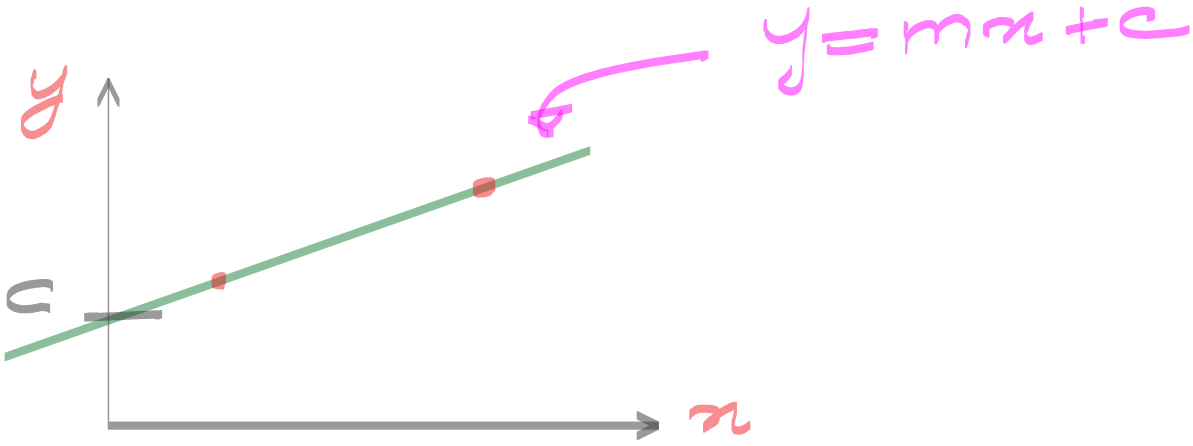
07 December 2022 20:35

* Equation of Line :

$$y = mx + c$$

$$y = a + bx$$

$$y = \theta_1 + \theta_2$$



m ⟶ slope of line

c ⟶ intercept

$$y = 5x + 3$$

If $x = 2$, $y = 13$

$x = 5$, $y = 28$

* Types of LR

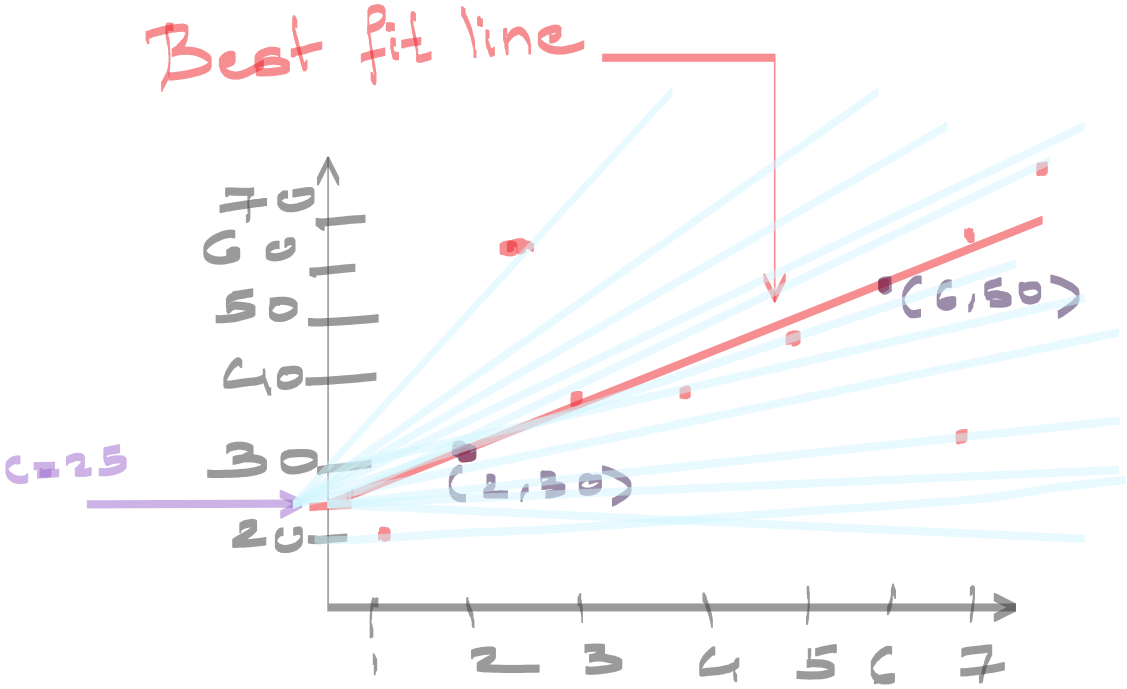
1. Simple LR

1. Simple LR

$$y = mx + c$$

2. Multiple LR

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n + c$$



(x) Exp	(y) Sal
2	30
3	35
4	35
5	45
6	50
7	60
8	70

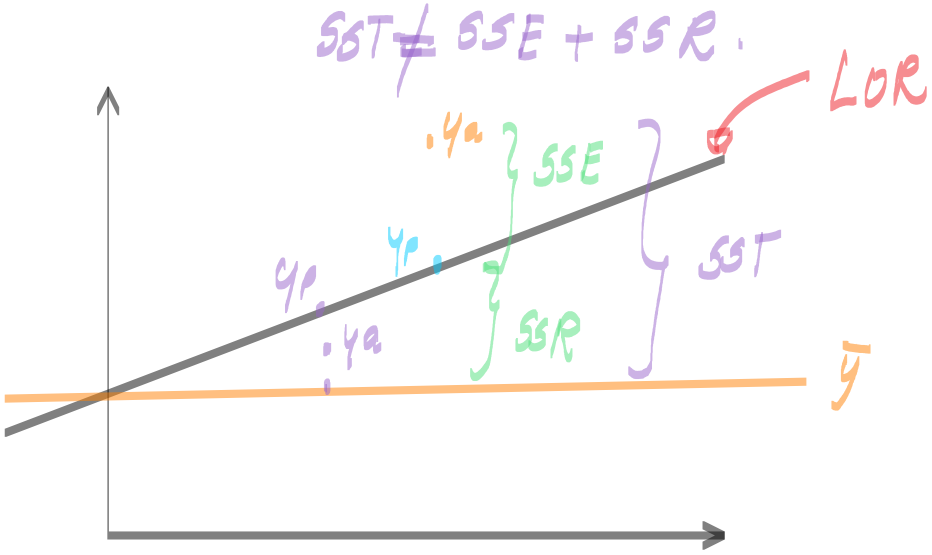
$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{50 - 30}{6 - 2} = \frac{20}{4} = 5$$

$$y = 5x + 25$$ ← Linear Model

* Blue lines → infinite number of possibilities

Linear Regression

08 December 2022 07:31



MSE → Scale Variant.

Ht.	Temp.
5.	150
5.2	160
5.4	170
6	183
6.1	185

MSE 25-36

MSE 2890

R-squared → R^2 → Scale Invariant.

0 to 1 , -ve

1. $R^2 = 1$ $\quad R^2 = 1 - \frac{SSE}{SST}$

$= 1 - \frac{0}{SST} = 1$

2. $R^2 = 0$ $\quad R^2 = 1 - 0$ $\quad SSE = SST$

2. $R^2 = 0$ $\quad R^2 = 1 - 0 = 1$, $\quad SSE = SST$

3. $R^2 = +ve$, $\quad SSE < SST$

$R^2 = 1 - 0.1 = 0.9$

4. $R^2 = -ve$, $\quad SSE > SST$

$R^2 = 1 - 1.4 = -0.4$

Non-linear.

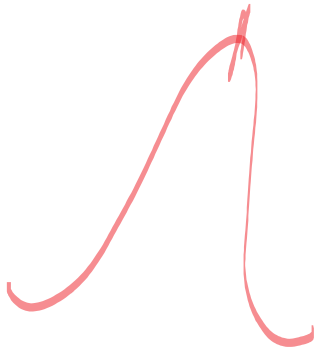
$$R^2 = 1 - \frac{SSE}{SST}$$

$SSE \rightarrow (y_a - y_p)^2$

$SST \rightarrow (y_a - \bar{y})^2$

Exp	Sal
2	30
3	40
4	50
5	60
6	70

y_a, $\bar{y} = 50$



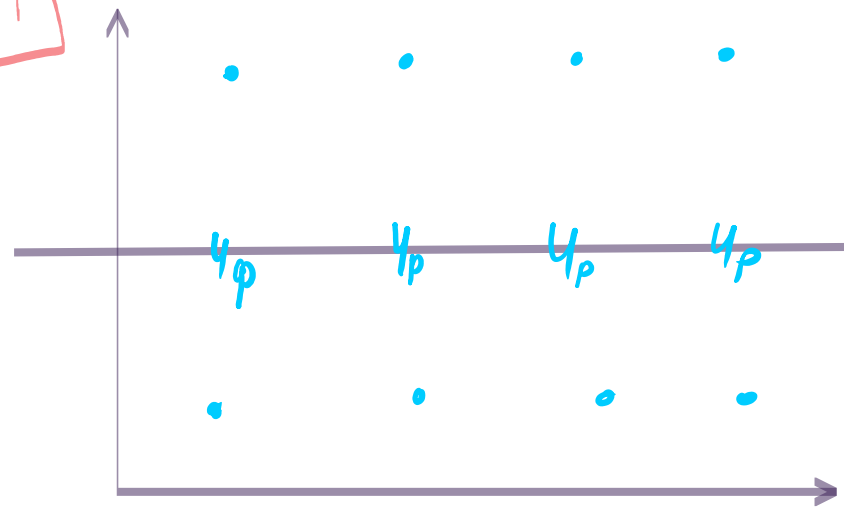
$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

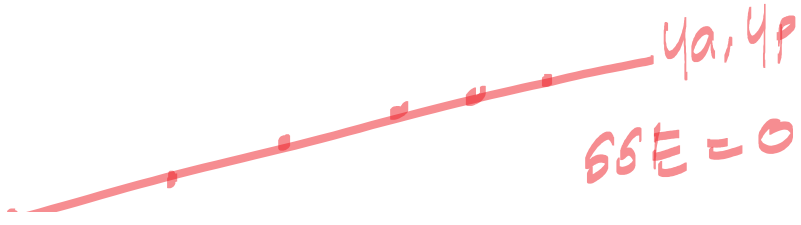
$$R^2 = \frac{SST - SSE}{SST} \rightarrow \frac{\text{Explained Var}}{\text{Total Var}}$$

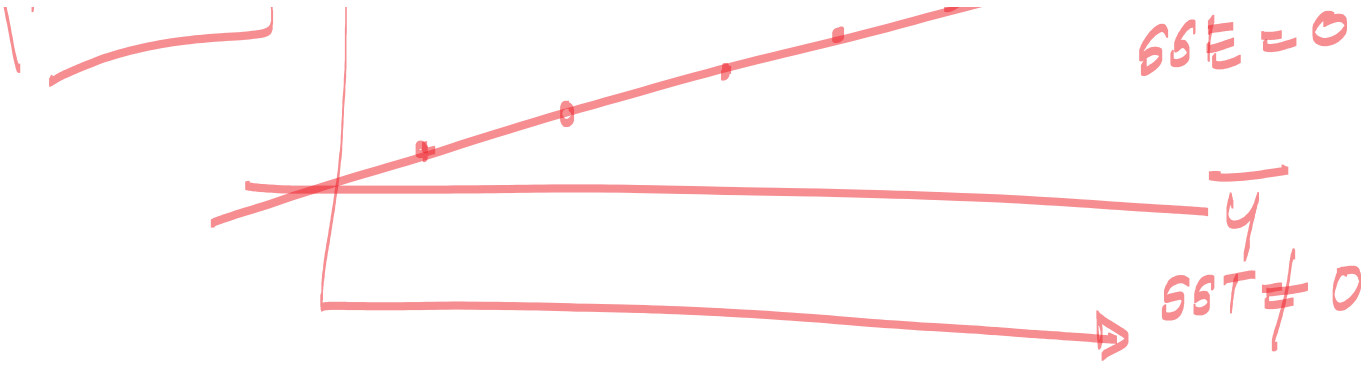
$SSE = 0$

$SSE = SST$



$R^2 = 1$





R-squared

Coeff. of Corr.

$R^2 = R \times R \longrightarrow$ SLR

$R^2 \neq R \times R \longrightarrow$ MLR

(r_2 - score)

$R^2 \longrightarrow$ Goodness of Best fit line

x_1	x_2	x_3	x_4	x_5	y
0.9	0.4	0.8	0.7	-0.5	

3 features $\longrightarrow R^2 = 0.85$ } 0.852
(x_1, x_3, x_4)

4 features $\longrightarrow R^2 = 0.86$ }
(x_1, x_2, x_3, x_4)

Adjusted R-squared → $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

n = No of samples (rows)

p = No of Predictors

$R^2 \geq \bar{R}^2$

	R^2	$\bar{R}^2$
	0.85	0.85
① $R = 0.3$	0.86	0.84
② $R = 0.9$	0.87	0.86

$$R^2 = 1 - \frac{SSE}{SST}$$

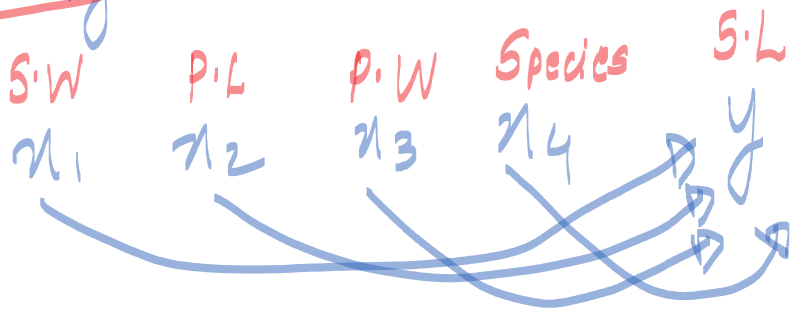
$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \quad \begin{matrix} \leftarrow 999 \\ \leftarrow 994 \end{matrix}$$

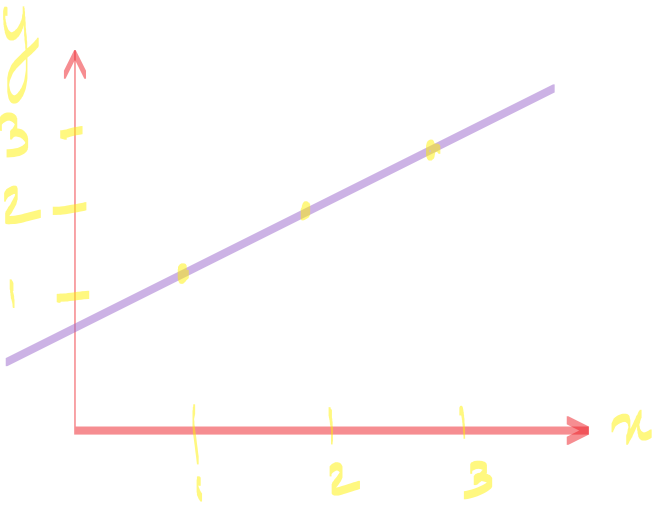
$n = 1000$, $p = 5$

$\boxed{n > p}$

~~5 rows~~
6 columns X

* Linearity.





R-value = 1

150 rows

df → x_1 x_2 x_3 x_4 y

Tr / Test ← x → Df → x_1 x_2 x_3 x_4

Tr / Test ← y → Series → y

Train → x_train (120 rows), y_train (120 rows)

Test → x_test (30 rows), y_test (30 rows)

x: x_1 x_2

y: y

Tr – 1, 2, 3

Tr – 4

Tr – 5

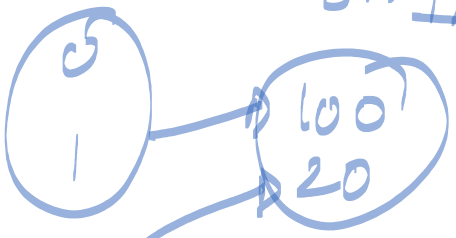
x_train – 1, 4, 5

Tr — 2
3
Tr — 4
Tr — 5

— Train → 3 row → x_train, 2, 4, 5
                → y_train, 2, 4, 5
— Test → 2 row → x_test 1, 3
                → y_test 1, 3

shuffle = True

50
50
50



120 rows
30 rows

100 row → 80 ← 0 → Train
          20 ← 1 → Test

* What is Data Science ?

Train — 4 lines → 3 correct 75%
                  1 incorrect 25%

evaluate

* What is KNN ?

Test — 4 line → 2 correct 50%

* what is RMSE

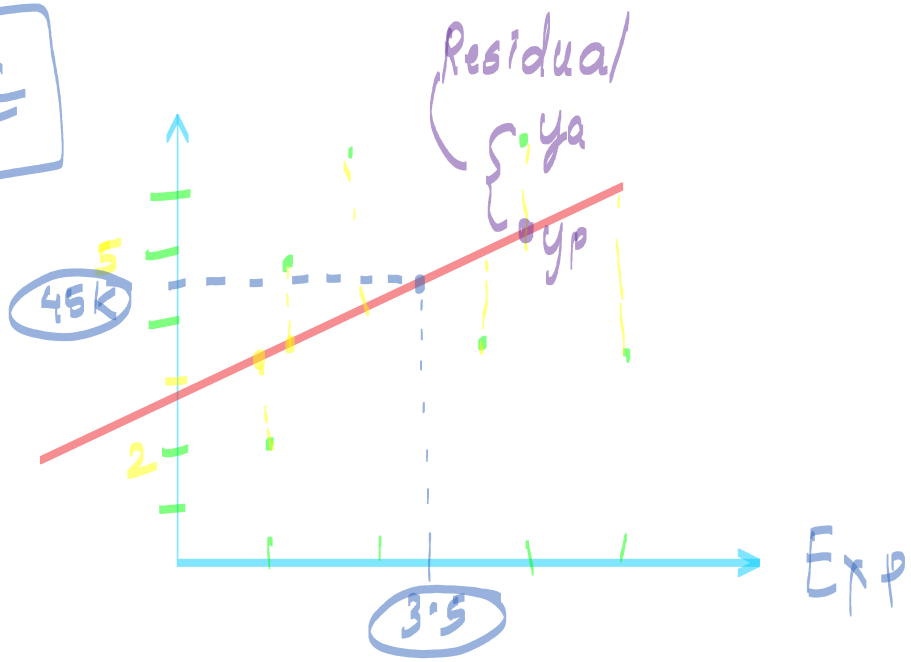
Test — 4 line → 2 correct 50%

Evaluate

unseen

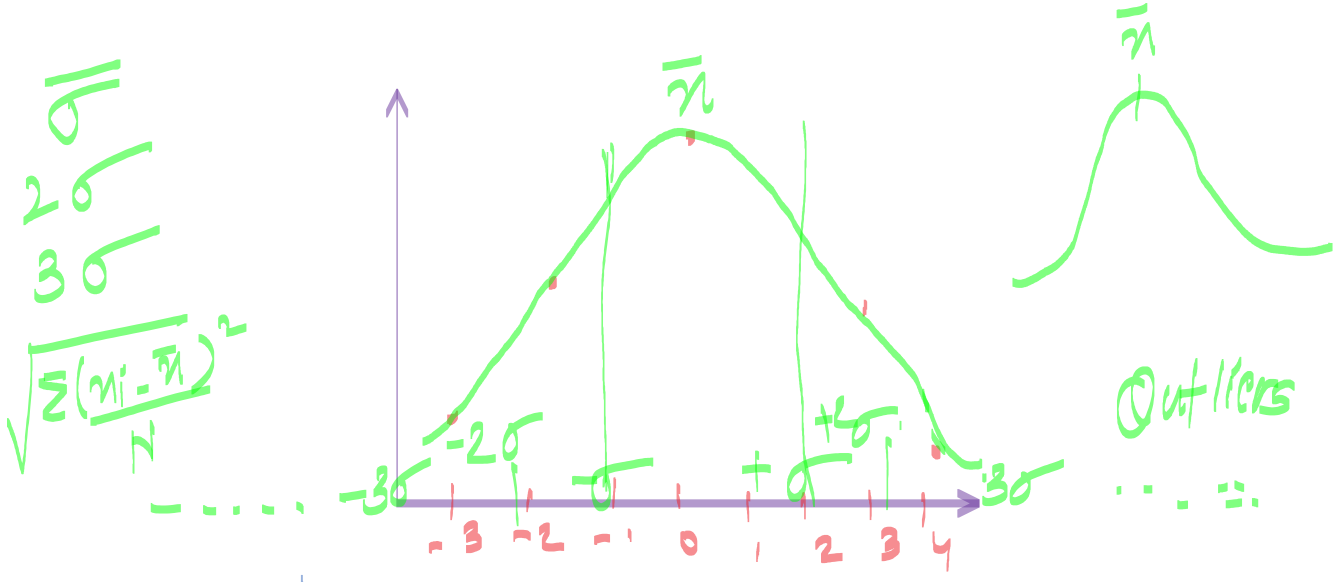
	Area	Bed	Bath	Location	Price
Train	1000	3	2	Baner	2cr
	700	2	1	Wakad	80
	800	1	1	Bale.	75
Test Unseen	650	2	2	Baner	90 L

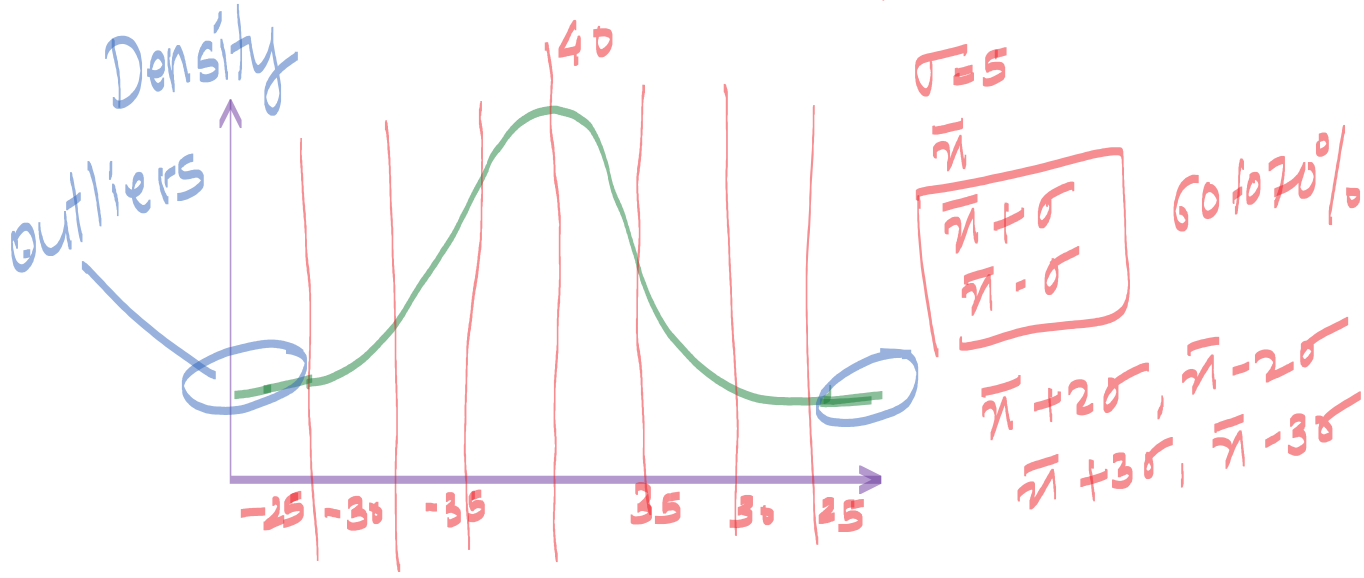
$y = mx + c$

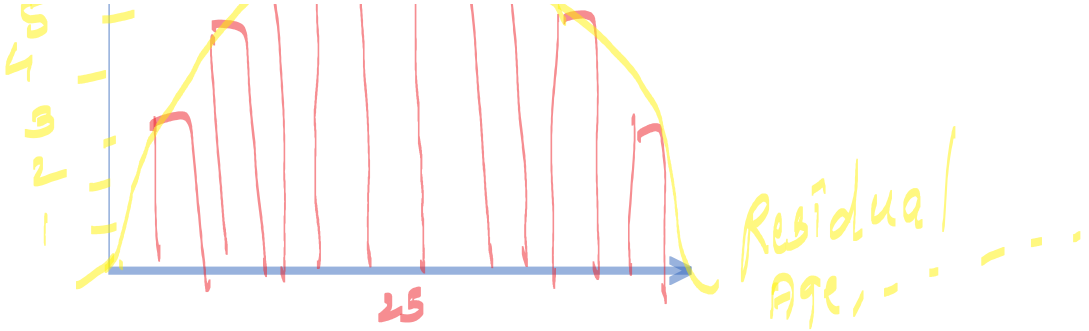


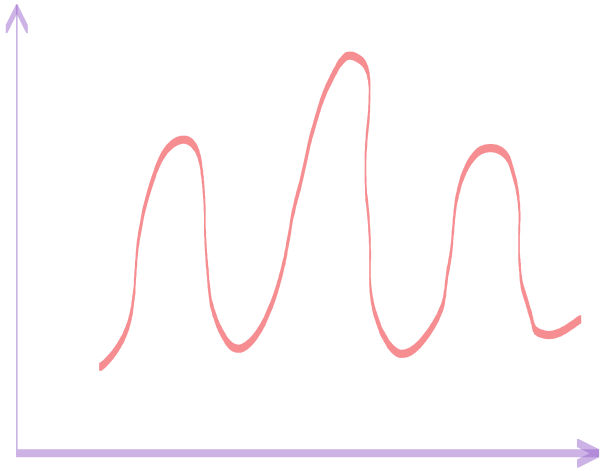
x_1 . . .	ya	yp	Residual
	10	12	-2

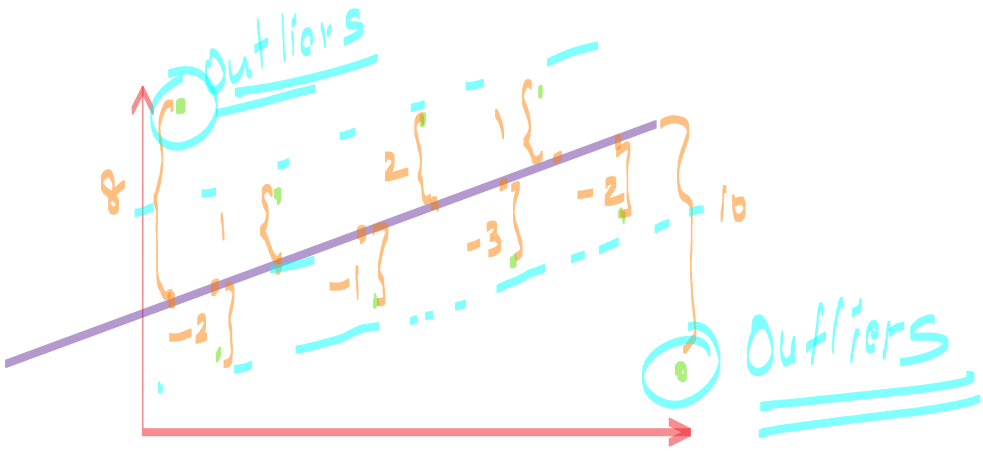
$x_1 \ldots$	y_a	y_p	Residual
	10	12	-2
	20	17	3
	30	34	-4
	25	28	-3
	15	12	3
	45	41	4
	50	50	0

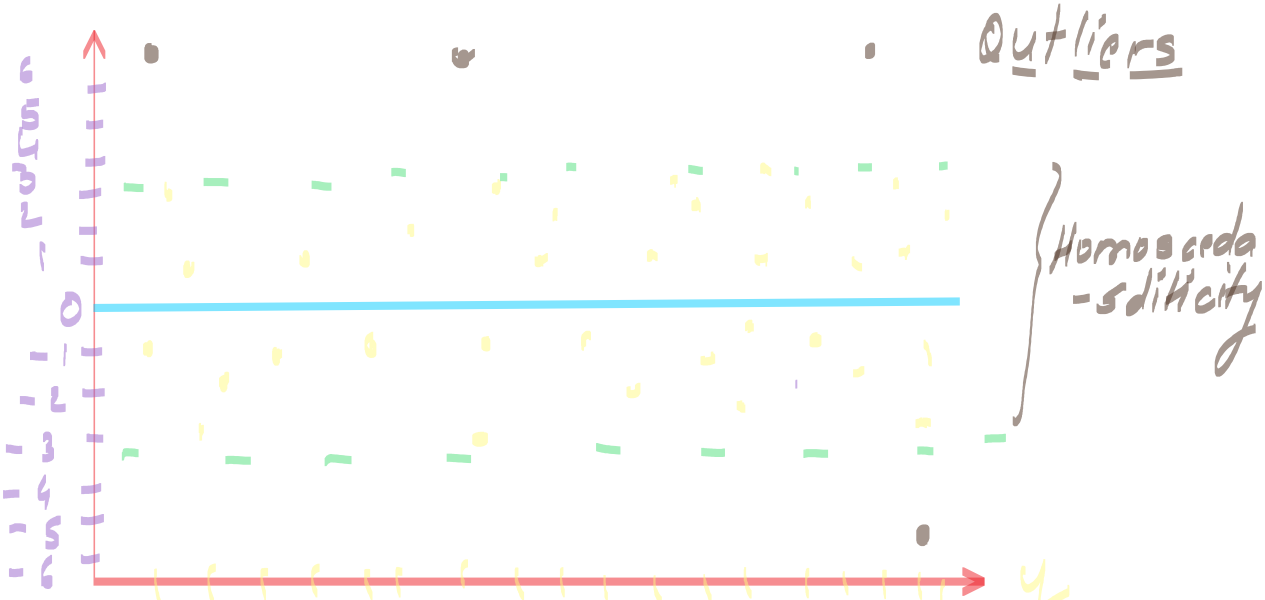


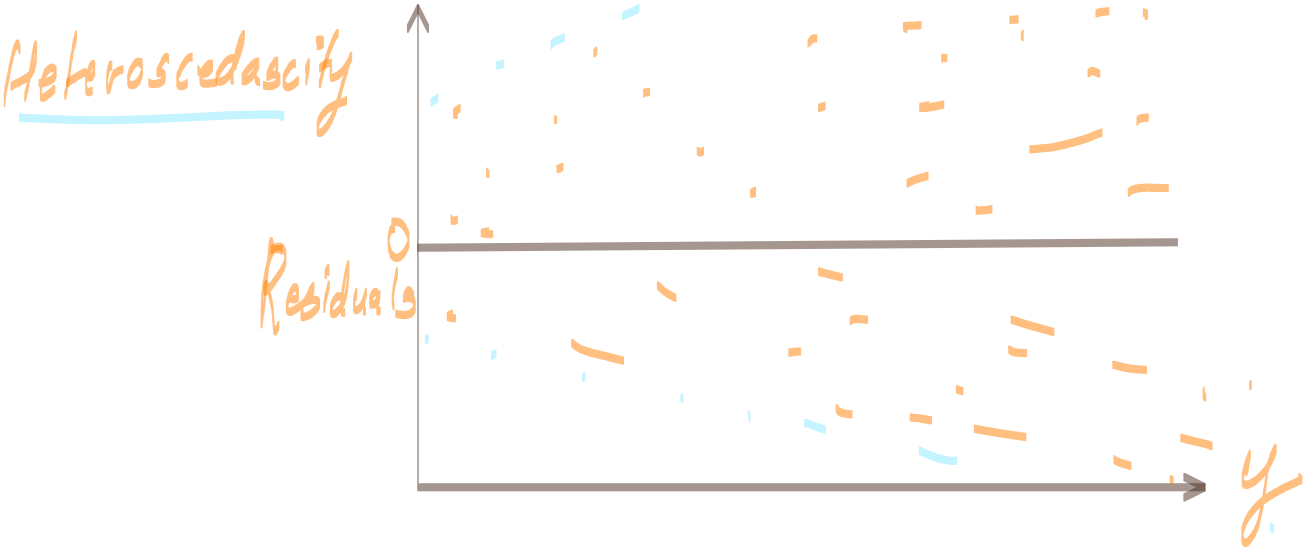



5
4
3
2
1
25
Residual
Age, - - - - -









MSE → square (Differential Term)

$(y_a - y_p)^2$

a (C.F) → (MSE)

$(y_a - y_p)^2$

$\frac{\partial C \cdot F}{\partial m, c} \rightarrow (MSE)$

MAE $\rightarrow$ $|y_a - y_p|$

120

100, 2, 3, 4, 5, 6 $\rightarrow$ Outliers

MSE

16 $\rightarrow$

60 $\rightarrow$ MSE

MSE is less robust $\rightarrow$ highly outliers

4 9 16 . . . 49 $\rightarrow$ MSE

2, 3, 4, ... 7 $\rightarrow$ MAE

MAE

RMSE $\rightarrow$

Price^2 X

Train $R^2 \rightarrow 0.9$

X Test $R^2 \rightarrow 0.6$ } X

1. Simple LR

x y.

$y = mx + c$ G.O.A

2. Multiple LR

x_1 x_2 x_3 x_4 y

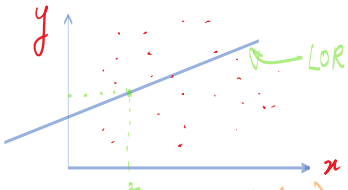
$y = m_1x_1 + m_2x_2 + \cdots m_nx_n + c$

$m = 10$, $c = 2$

$y = 10x + 2$

$x = 2$, $y_p = ?$

$y_p = 10 \times 2 + 2 = 22$



1. Regression → Target (y) continuous.
2. Classification → Categorical

Independent				Target (dependent)
x_1	x_2	x_3	x_4	y
Con	Con	Cat	Cat	Con (Reg)
Con	cat	Con	cat	Cat (Class)

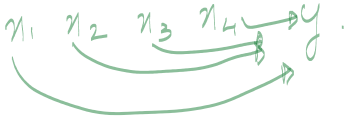
* Assumptions of LR.

1. Linearity

Linear Rel x & y.

Correlation → Coeff or R value

df.corr()



Positive Corr $R \geq 0.7$
Good Predictors

Negative Corr $R < -0.7$
Good Predictors

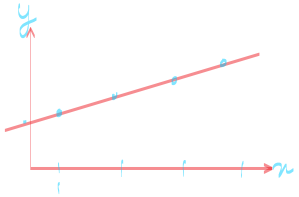
Good Predictors

No corr -0.3 R 0.3

R = 1 → strong +ve
R = -1 → strong -ve
R = 0

$$R = \frac{Cov}{\sigma_x \sigma_y}$$

	y	x_1	x_2	x_3	x_4
y	1	✓	✓	✓	✓
x_1					
x_2					
x_3					
x_4					



R=1

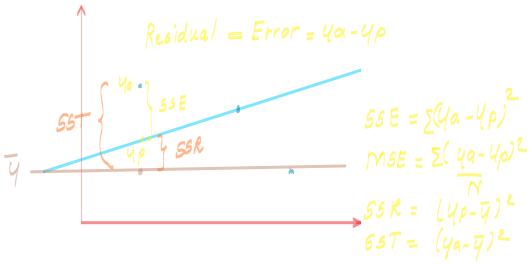
2. No Multicolinearity

x_1 ——
x_2 ——
x_3 ——
x_4 ——
} VIF

1 to inf

x_1, VIF = 1, No corr
VIF = 1 to 5, Moderate
VIF > 5, Highly
VIF > 10, significantly.

Linearity $x_1, x_2 \rightarrow y$
No Multico



$SSE = \sum(y_a - y_p)^2$

$MSE = \frac{\sum(y_a - y_p)^2}{n}$

$SSR = (y_p - \bar{y})^2$

$SST = (y_a - \bar{y})^2$

$R^2 \rightarrow$ R squared (Coeff. of Determination)

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SST = SSE + SSR$$

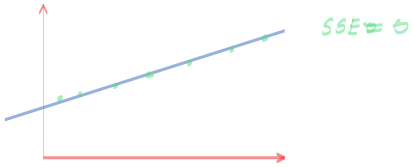
$$= \frac{SST - SSE}{SST} \quad = \frac{SSR}{SST}$$

$R^2 \rightarrow$ 0 to 1

$R^2 \rightarrow$ -ve (Non linear)

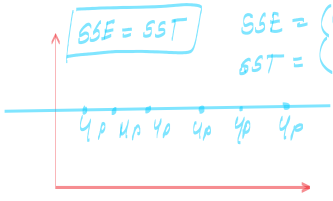
1. $R^2 = 1$, $SSE = 0$

$$R^2 = 1 - \frac{SSE}{SST} = 1$$



2. $R^2 = 0$ $SSE = SST$

$$R^2 = 1 - \frac{SSE}{SST}$$



3) $R^2 = +ve$ $SSE < SST$

$$R^2 = 1 - \frac{SSE}{SST}$$

4) $R^2 = -ve$, $SSE > SST$

Non Linear

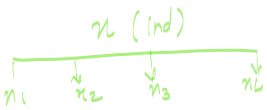
Species SL.
x_1 y

50 0 -
50 1 -
50 2 -

80% Train - 120
0(50), 1(50), 2(20)

20% Test
2(30)

shuffle = True



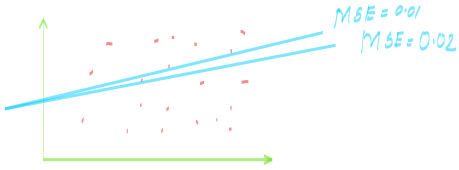
150 rows

x_train
120 rows

x_test
30 rows

y_train
120 rows

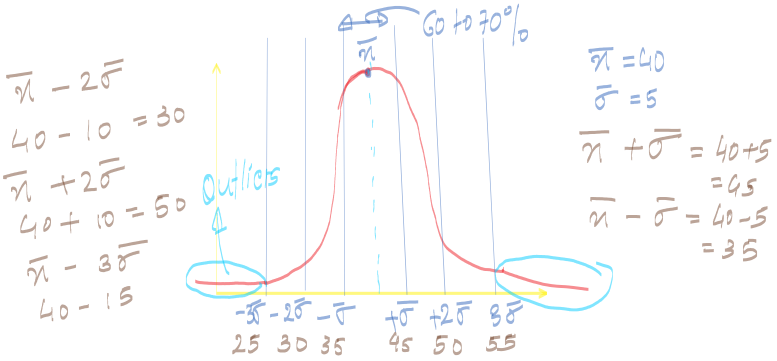
y_test
30 rows



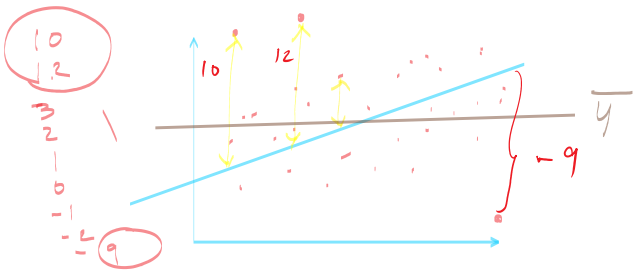
1. What is Data Science ? (Train)
→ 5 line → 4 lines → 80%

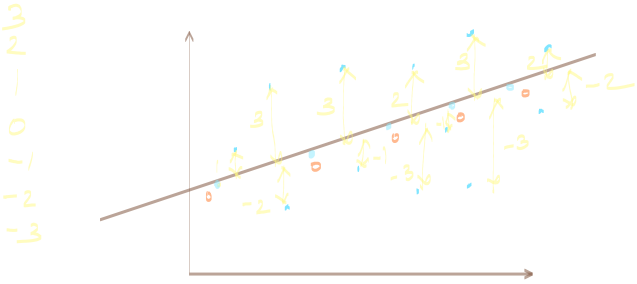
2. What is KNN ?
5 line → 2 line → Test (Unseen)

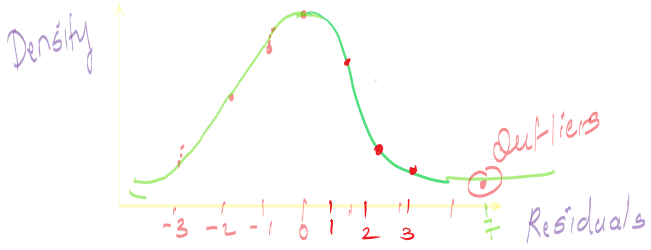
	x_1	x_2	x_3	x_4	y_a	y_p
Train	1	2	3	4	20	22
Test	2	3	4	5	24	25

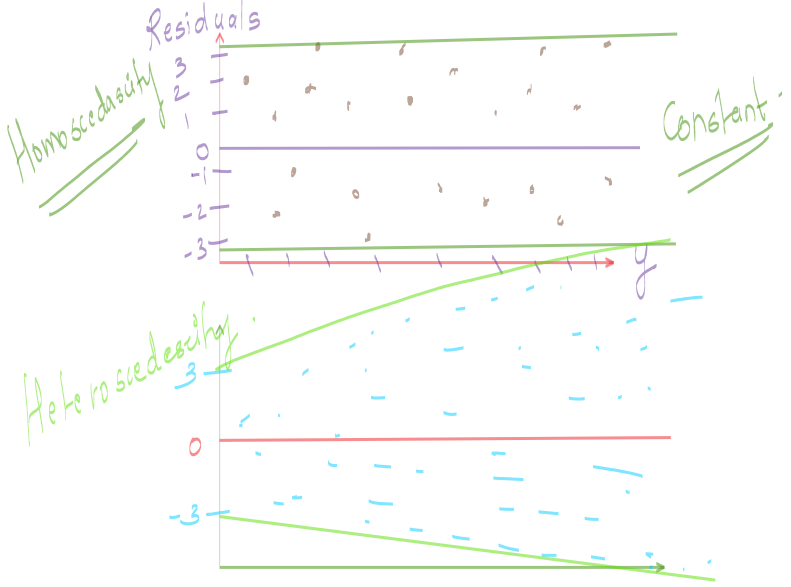


$\bar{x} = 40$
$\bar{\sigma} = 5$
$\bar{x} + \bar{\sigma} = 40 + 5 = 45$
$\bar{x} - \bar{\sigma} = 40 - 5 = 35$









(CF) MSE → Differential fn $(y_a - y_p)^2$

MAE → $|y_a - y_p|$ cm, Price scale variant

RMSE → $\sqrt{MSE}$

MSE: length cm cm^2

Price Rs Rs^2

MAE 2 3 - - - - 5 → 3 → 10

MSE 4 9 25 → 21 → 100

→ less robust (Highly sensitive)

C.F → MSE

* Outliers

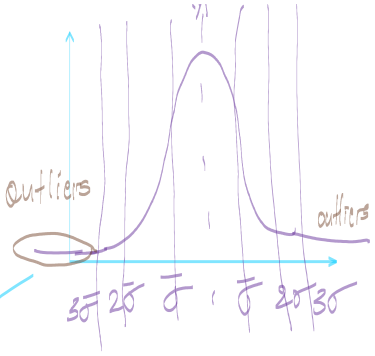
* Outliers

C.F → MSE

Mean $_{without\_out}$ = 4.5

Mean $_{with\_out}$ = 50

2
3
4
5
6
100
7
8



* Missing

Continuous → mean ✓
→ median

Categorical → mode.

$$R^2 = 1 - \frac{SSE}{SST}$$

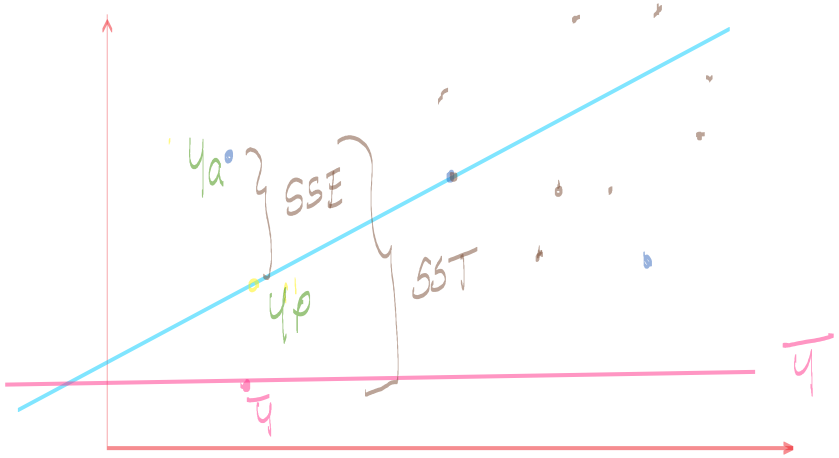
$\bar{R}^2$ → Adjusted R^2

✓ x_1 → $R = 0.85$
✓ x_2 → $R = -0.3$ ✓

x_1 → $R^2 = 0.85$
x_2 → $R^2 = 0.86$.

Adjusted $\bar{R}^2$

$\bar{R}^2 = 0.82$ ↓
$\bar{R}^2 = 0.81$



$$SSE = (y_a - y_p)^2$$

$$SST = (y_a - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

$$= \frac{SST - SSE}{SST}$$

$$= \frac{\text{Total var} - \text{Unexplained var}}{\text{Total}}$$

$$R^2 = \frac{\text{Explained Var}}{\text{Total Var.}}$$

$R^2 = 1$ | $R^2 = 0$ | $R^2 = +ve$ | $R^2 = -ve$

$SSE = 0$ | $SSE = SST$ | $SSE < SST$ | $SSE > SST$

Non - Linear

* $\bar{R}^2 \longrightarrow$ Adjusted.

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

$n > p$

Simple $\longrightarrow$ $R^2 = R \times R$

Multiple $\longrightarrow$ $R^2 \neq R \times R$

n = No. of samples (rows)

p = No. of predictors (features)

$r_1 = 0.85$ | $r_2 = -0.2$

$R^2 = 0.85$

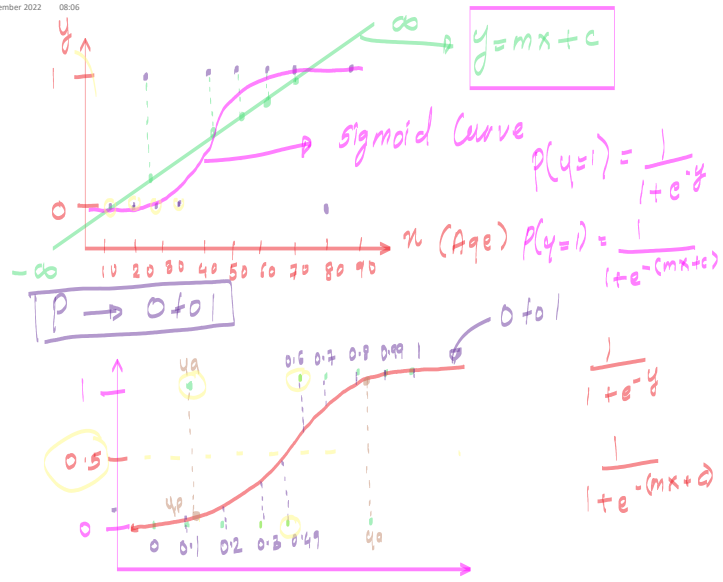
$\bar{R}^2 = 0.83$

$\bar{R}^2 = 0.82$

$R^2 = 0.85$

$R^2 = 0.86$

$\overline{R}^2 = 0.82$



$P(y=1) = 0.8$

$P(y=0) = 1 - 0.8 = 0.2$

$p \geq 0.5 \rightarrow$ Class 1

$p < 0.5 \rightarrow$ Class 0

Logistic $\leftarrow \log\left(\frac{p}{1-p}\right) \rightarrow$ odds, logit

x	ya	yp		
10	0	0.4	0	✓
20	0	0.2	0	✓
30	1	0.8	1	✓
40	0	0.9	1	✗ misclassified
50	1	0.7	1	✓
60	1	0.3	0	✗ misclassified
70	0	0.4	0	✓

$p \geq 0.5 \rightarrow 1$

$p < 0.5 \rightarrow 0$

7 → 5 C, 2 M

$\frac{5}{7} \times 100$

$\frac{7}{10} \times 10$

$\frac{x}{1-x} \rightarrow$ odds

$\log\left(\frac{p}{1-p}\right) \Longrightarrow$ log odds, logit Fn

Cost Function → MSE → $\frac{(y_a - y_p)^2}{N}$ → Linear Reg

C.F → Log Loss → Logistic Regression

$$\text{Log Loss} = -\frac{1}{N}\left[\sum y_a \log(p) + (1-y_a)\log(1-p)\right]$$

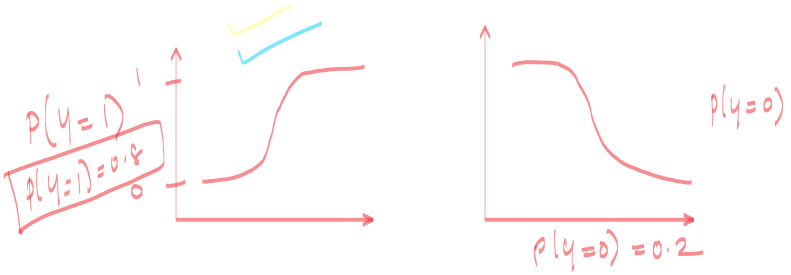
N

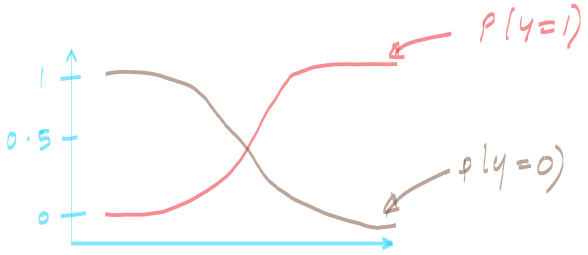
$\frac{1}{1+e^{-y}} \Rightarrow 0 \text{ to } 1$

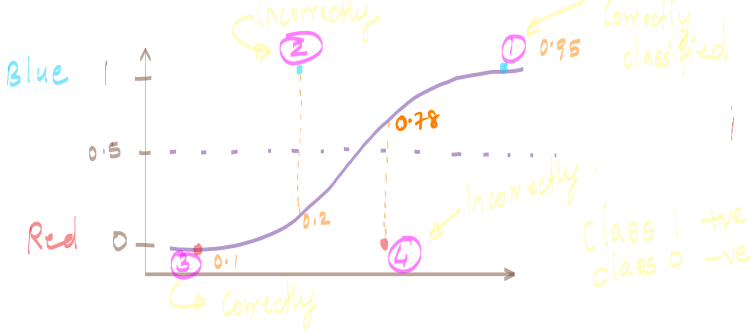
$e = 2.78$

0 to 1

$\frac{1}{1+e^{0.8}}$, $\frac{1}{1+e^{4}}$, $\frac{1}{1+e^{-0.4}}$, $\frac{1}{1+e^{10}}$







$y_a = 1$

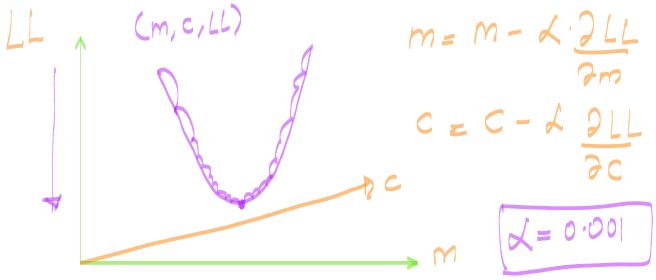
$-\frac{1}{N}\left[\sum \underbrace{y_a}_{1} \log P + \underbrace{(1-y_a)}_{0} \log(1-P)\right]$

Class 1 $\longrightarrow$ LL $\longrightarrow -\log P$

$y_a = 0$

$-\frac{1}{N}\left(\underbrace{y_a \log P}_{0} + \underbrace{(1-y_a)}_{0} \log(1-0)\right)$

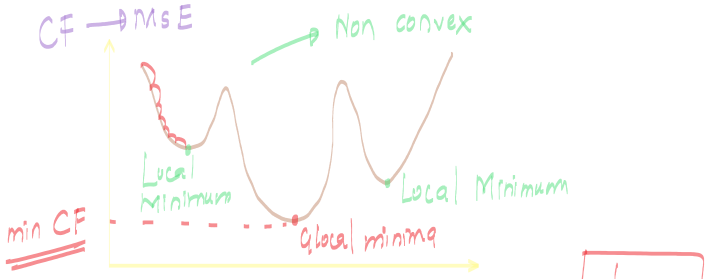
Class 0 $\longrightarrow$ LL = $-\log(1-P)$

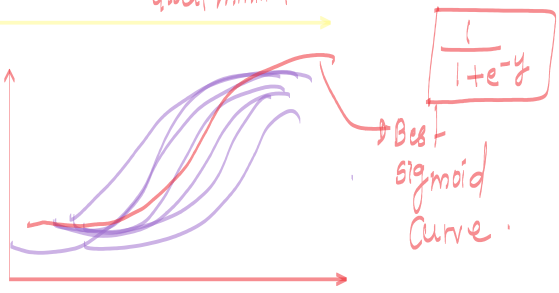


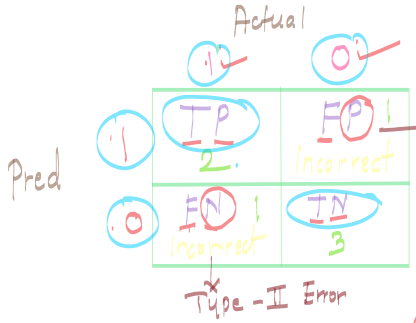
$$m = m - \alpha \cdot \frac{\partial LL}{\partial m}$$

$$c = c - \alpha \frac{\partial LL}{\partial c}$$

$\alpha = 0.001$







Type I Error

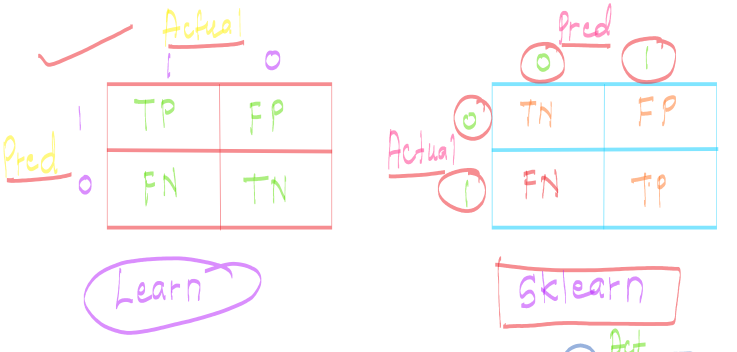
class 1 → +ve
class 0 → -ve

7 pts

Correct = 5 / 7 × 100 = 70%

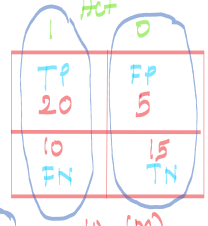
✓ Accuracy Score = $\frac{TP + TN}{TP + FN + TN + FP}$ (Correct / Total pts)

$= \frac{2+3}{2+1+3+1} = \frac{5}{7} = 0.7 \times 100 = 70\%$



Learn

sklearn



① $TPR = \frac{TP}{TP + FN}$

Sensitivity
Recall

Total → 50 → class 1 (30) → 20 → 10 (M)

Sensitivity
Recall

Total → 50 → class1 (30 → 20 → 10 (M))
→ class0 (20 → 15 → 5 (M))

	FN	TN

(2) $FNR = \frac{FN}{TP+FN}$

(3) $TNR = \frac{TN}{TN+FP}$ Specificity

(4) $FPR = \frac{FP}{TN+FP}$

$Recall = \frac{TP}{TP+FN}$

$Precision = \frac{TP}{TP+FP}$

Pred \ Actual	1	0
1	TP	FP
0	FN	TN

Precision (→ TP, FP)
Recall (→ TP, FN)

Classification Report

	Precision	Recall	f1 score
class0	—	—	—
class1	—	-	—

Accuracy score = $\frac{TP+TN}{TP+TN+FP+FN}$

Pred \ Act	1	0
1	25	4
0	5	16

Precision = $\frac{25}{25+4} = \frac{25}{29}$

Recall = $\frac{25}{25+5} = \frac{25}{30}$

(1) 900 — 100
(0) (1)

	1	0
1	100	0
0	0	900

Acc = $\frac{100+900}{100+900} = 100\%$

(2)

	40	50
	60	850

Acc = $\frac{40+850}{1000} = \frac{890}{1000} = 89\%$

(3)

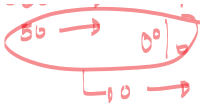
	1	0
1	900	100
0	0	0

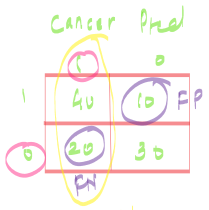
Acc = $\frac{900}{100} = 90\%$ ✓

50 — 50
70 — 30
60 — 40
65 — 35

90 — 10
80 — 20
} Imbalanced data

500 → 100%
50 → 0%





Recall → Increase = $\frac{TP}{TP+FN}$ ↓ ↑

(2) Spam - Filter

1 → Spam
0 → Not Spam

Precision

Offer letter

Not spam → Spam

Adv

Spam → Not spam

(3) f1-score

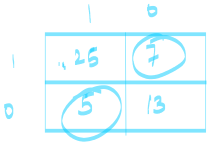
f_β-score = $(1+\beta^2)\frac{P \times R}{(\beta^2 \times P)+R}$

$\beta = 1$, f_1-score = $\frac{2PR}{P+R}$ → Harmonic mean

$x \& y = \frac{2xy}{x+y}$

$\beta = 0.5$ $f_{0.5} = \frac{1.25\ P \times R}{0.25\ P+R}$

$\beta = 2$ $f_2 = \frac{5PR}{4P+R}$



Precision = $\frac{25}{32}$ = 0.8

Recall = $\frac{25}{30}$ = 0.83

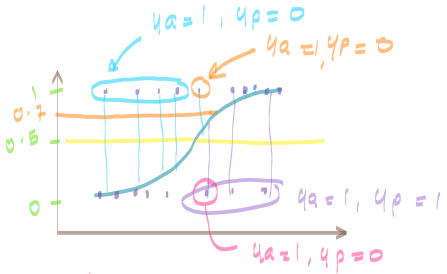
Precision = $\frac{25}{25}$ = 1

Recall = $\frac{30}{30}$ = 1

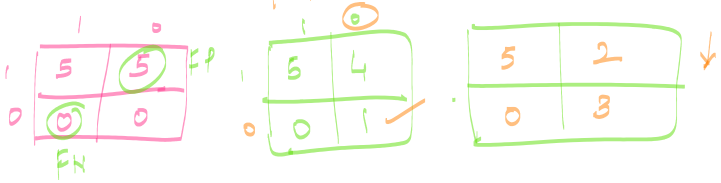
Loan → Approved / Declined

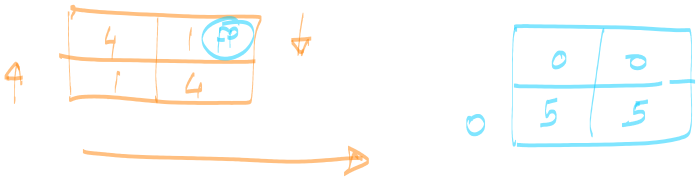
FP & FN

f_1-score ↓



ya	p	yp(0)	yp(0.1)	yp(0.2)	yp(0.4)	y(0.6)	yp(1)
1	0.8						0
1	0.95						0
1	0.42					0 X	0
0	0.33				0 ✓	0 ✓	0
1	0.72						0
1	0.65					0 X	0
0	0.45						0
0	0.56						0
0	0.15				0 ✓	0 ✓	0
0	0.26				0 ✓	0 ✓	0





$y_{p(0)} \longrightarrow y_{p(1)}$

FP ↓ ⟶ Precision ↑

FN ↑

$y_{p(0)} \longleftarrow y_{p(1)}$

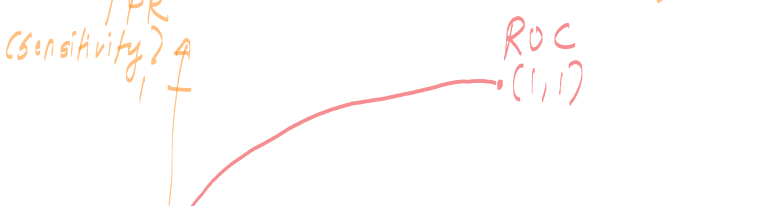
FN ↓ ⟵ Recall ↑

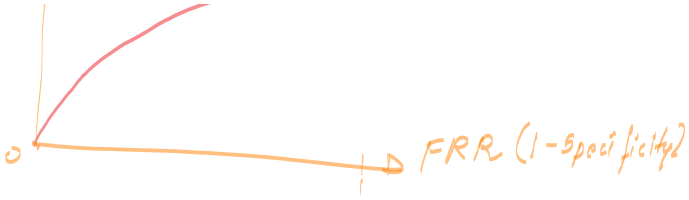
ROC & AOC

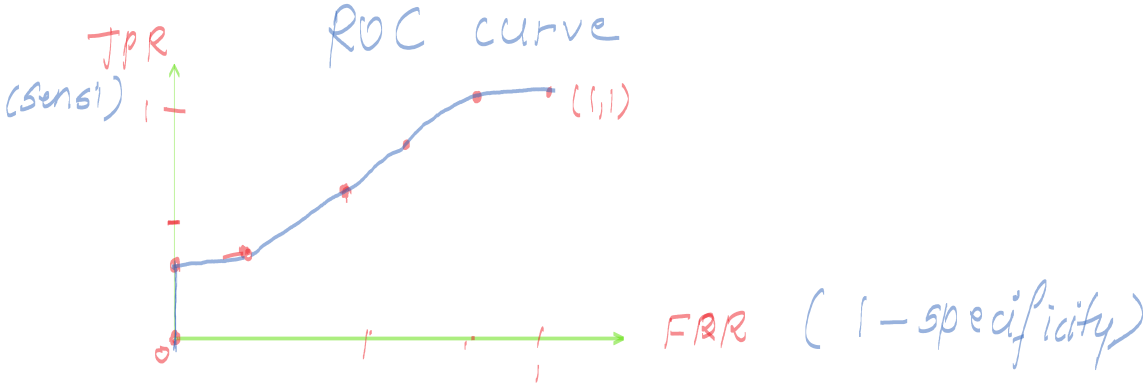
$$TNR = \frac{15}{20} = \frac{TN}{TN+FP} = 0.75$$

$$FPR = \frac{FP}{TN+FP} = \frac{5}{20} = 0.25$$

$$FPR = 1 - TNR = 1 - \text{specificity}$$




0
FRR (1-Specificity)



① Th = 0

	1	0
1	5	5
0	0	0

$TPR = \frac{5}{5+0} = 1$

$FPR = \frac{5}{5+0} = 1$

② Th = 0.2

	1	0
1	5	4
0	0	1

$TPR = 1$

$FPR = \frac{4}{4+1} = \frac{4}{5} = 0.8$

③ Th = 0.4

4	3
1	2

$TPR = \frac{4}{5} = 0.8$

$FPR = \frac{3}{5} = 0.6$

④ Th = 0.5

3	2
2	3

$TPR = \frac{3}{5} = 0.6$

$FPR = \frac{2}{5} = 0.4$

⑤ Th = 0.7

2	1
3	4

$TPR = \frac{2}{5} = 0.4$

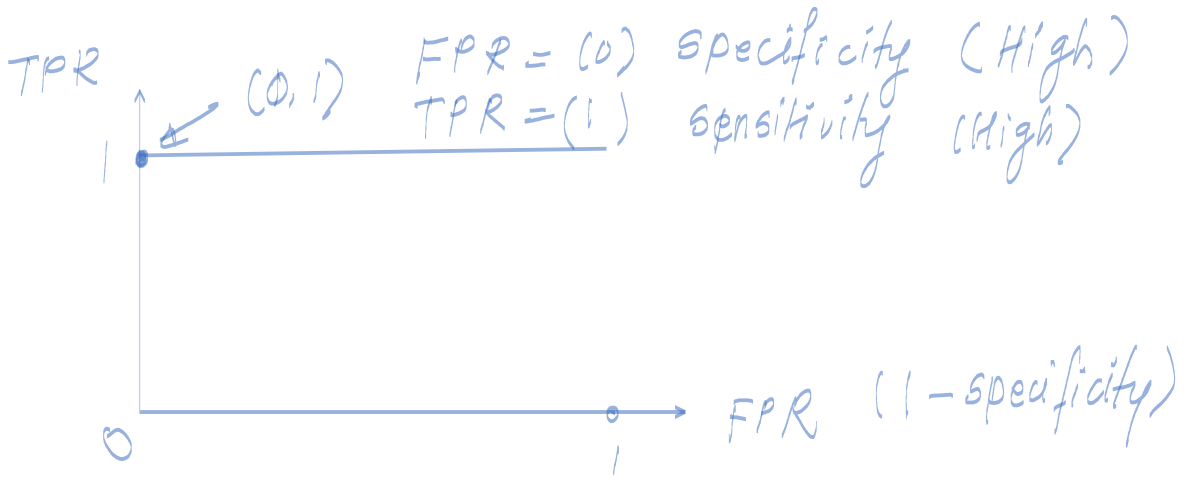
$FPR = \frac{1}{5} = 0.2$

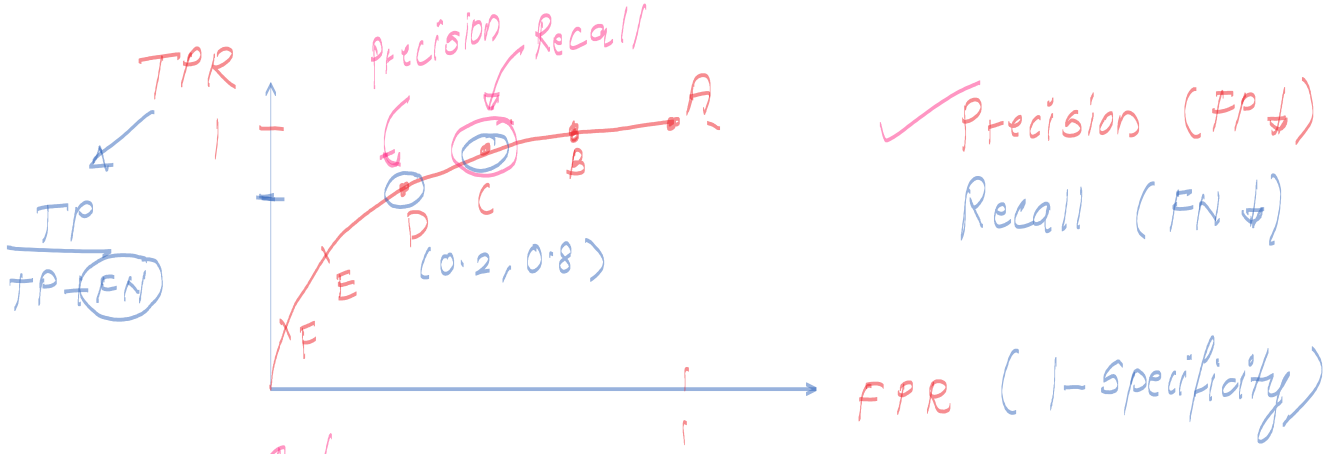
⑥ Th = 0.9

1	0
4	5

$TPR = \frac{1}{5} = 0.2$

$FPR = 0$





✓ Precision (FP ↓)
Recall (FN ↓)

✓

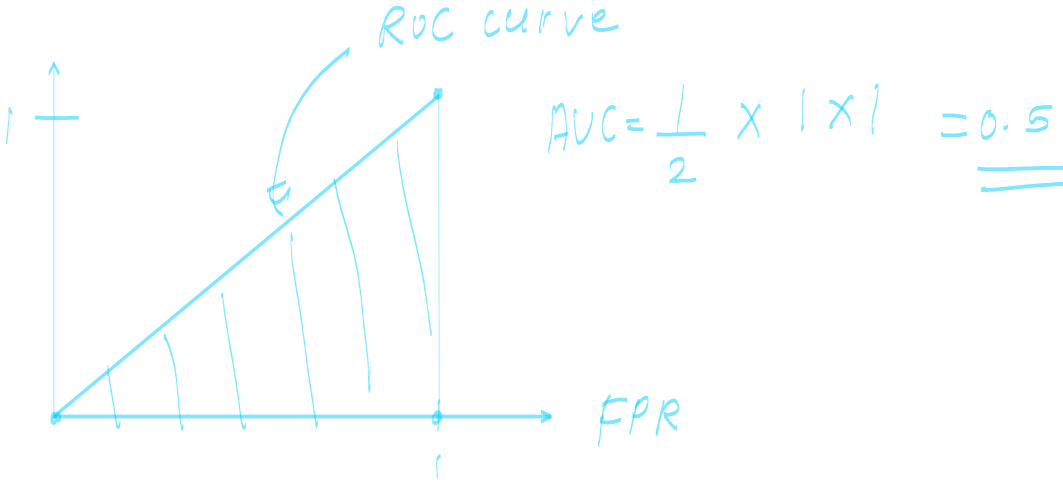
Act

Pred	1	0
1	TP	FP
0	FN	TN

Pred

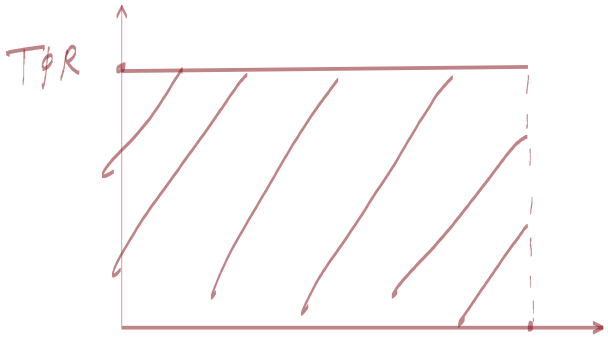
Act	0	1
0	TN	FP
1	FN	TP

sklearn



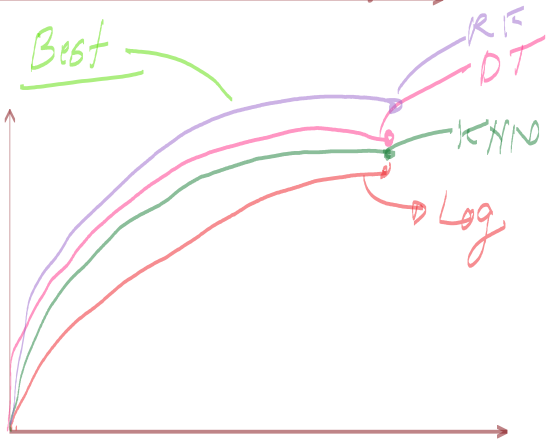
AUC = $\frac{1}{2} \times 1 \times 1$ = 0.5

TPR = 1, FPR = 0



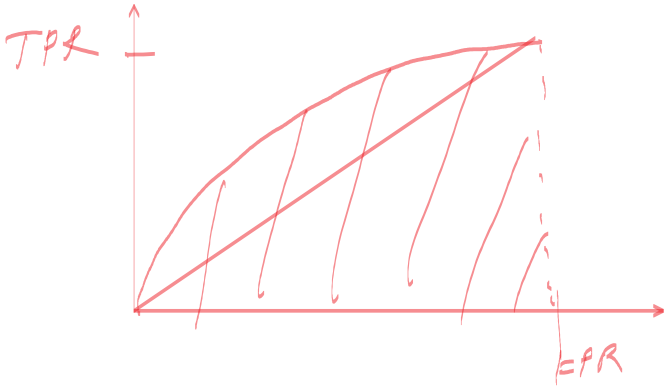
AUC

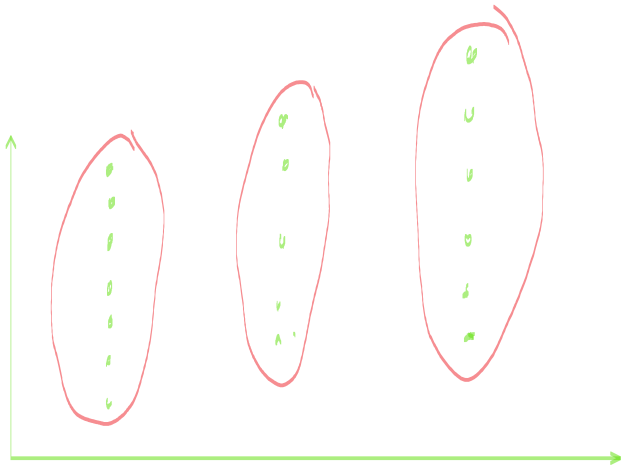
|X| = 1



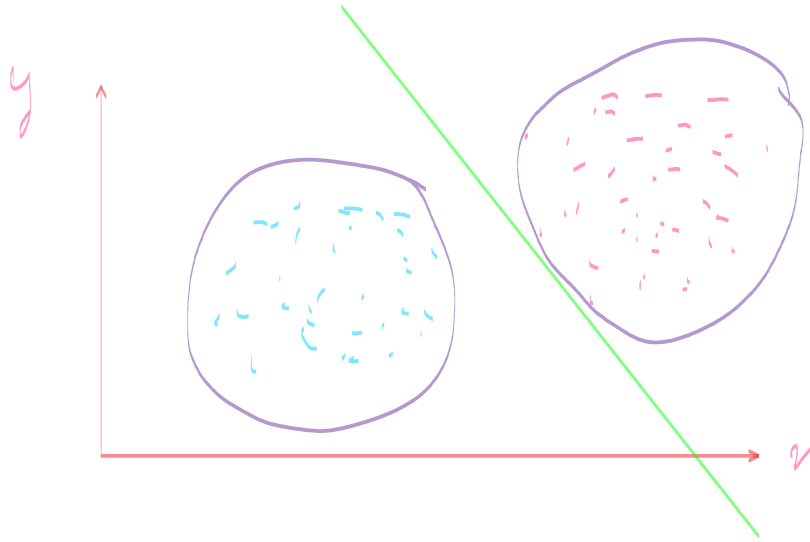
0.5 < AUC < 1

0.8
0.7
0.6

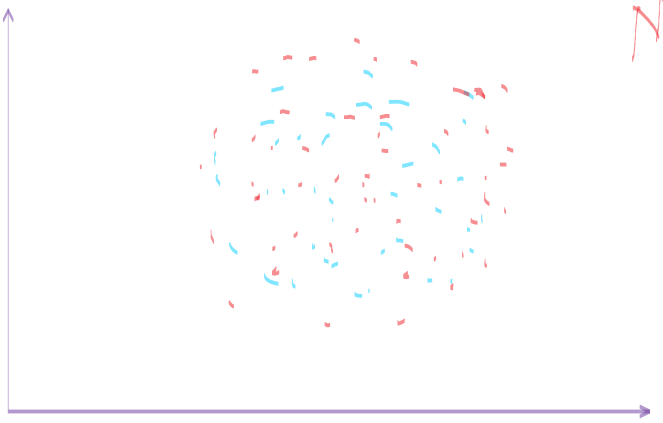




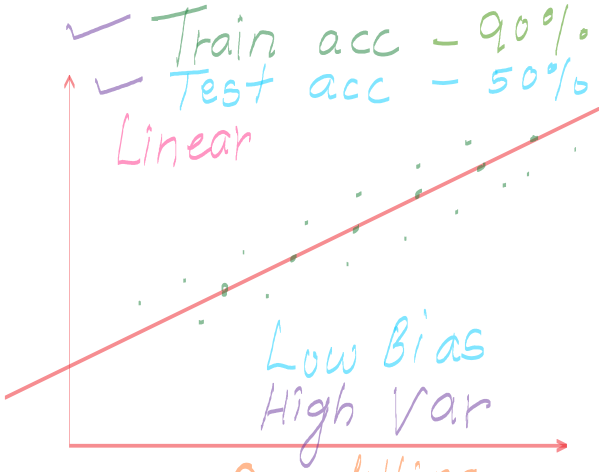
Species

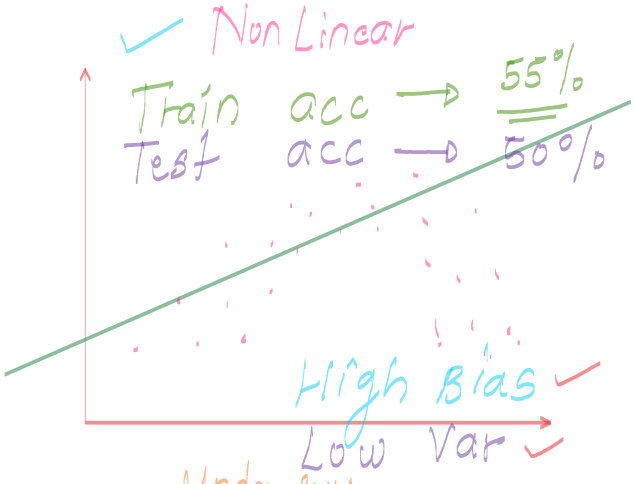


Linearly seperable



Non linearly seperable data





Bias → Low Bias (High Training)
High Bias (Low Training)

Variance → High Var → High Train, Low Test
High Test, Low Train

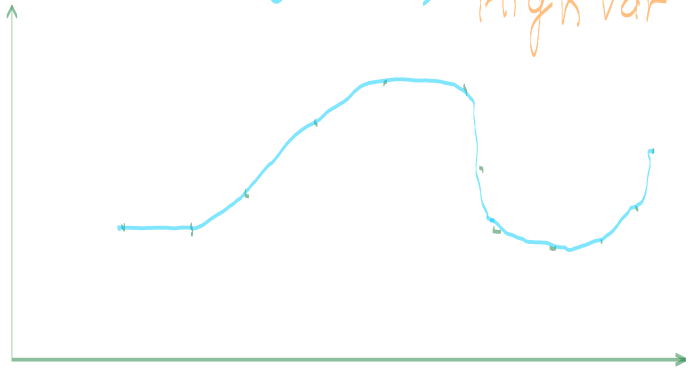
Variance → High ... → High Test, Low Train

Low Var → Low Train, Low Test

→ High Train, High Test

① Over fitting → Low Bias

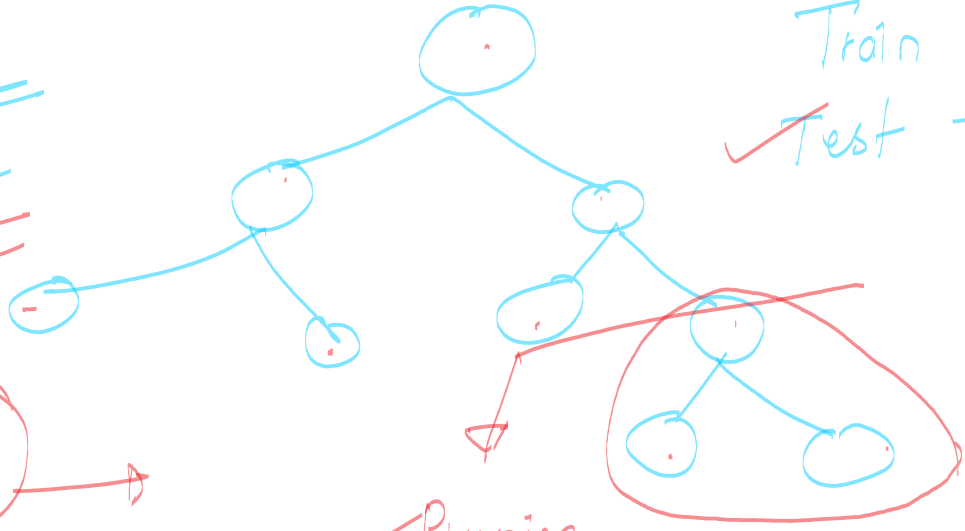
→ High var



Train = 100%

DT

Overfit



Train → 100%

Test → 70%

90%.

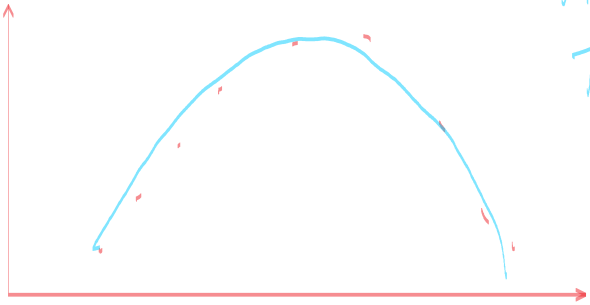
RF →

Pruning

(2) Underfitting

Train - 60%
Test - 58%

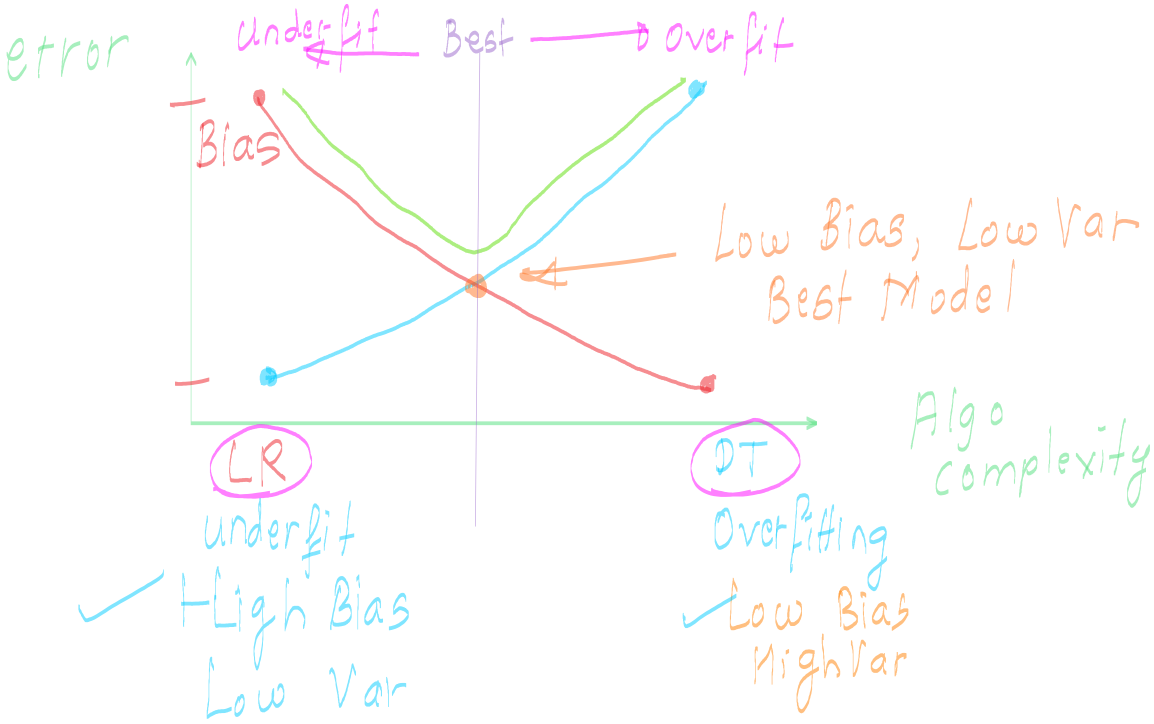
(3) Best Model



Train 90%
Test 85%

Low Bias, Low Var.

* Bias-Variance - Trade off



underfit
High Bias
Low Var

Overfitting
Low Bias
High Var

$$TE = Bias^2 + Var + \text{Irreducible Error}$$

Error

1000 Rows → Train 800 → CV

→ Test 200

Cross Validation

800

Train

200

Test

Unseen.

CV = 4

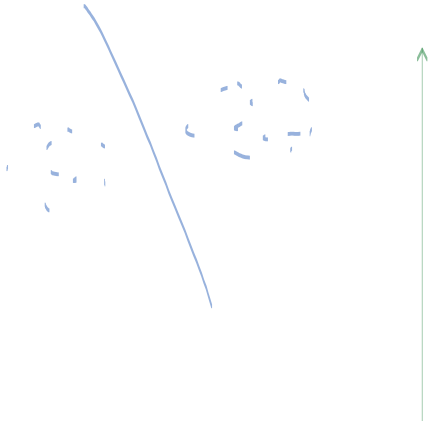
Train

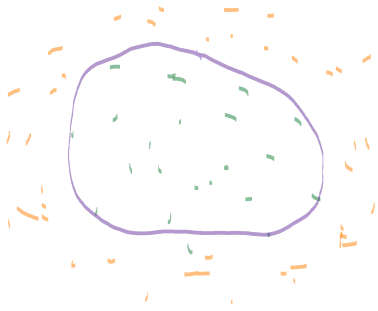
Test (val)

① 200 200 200 200

② 200 200 Test 200 200

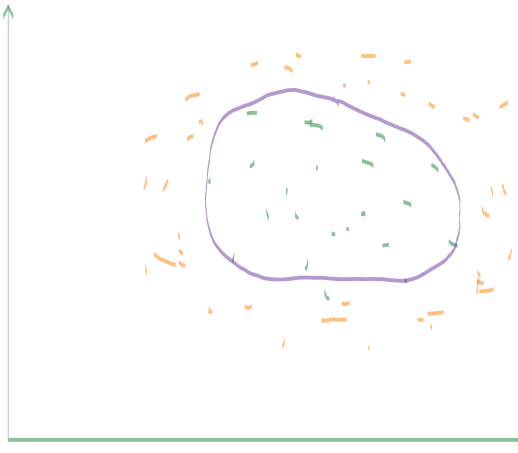
③ 200 Test 200 200 200





Non linear

SVM



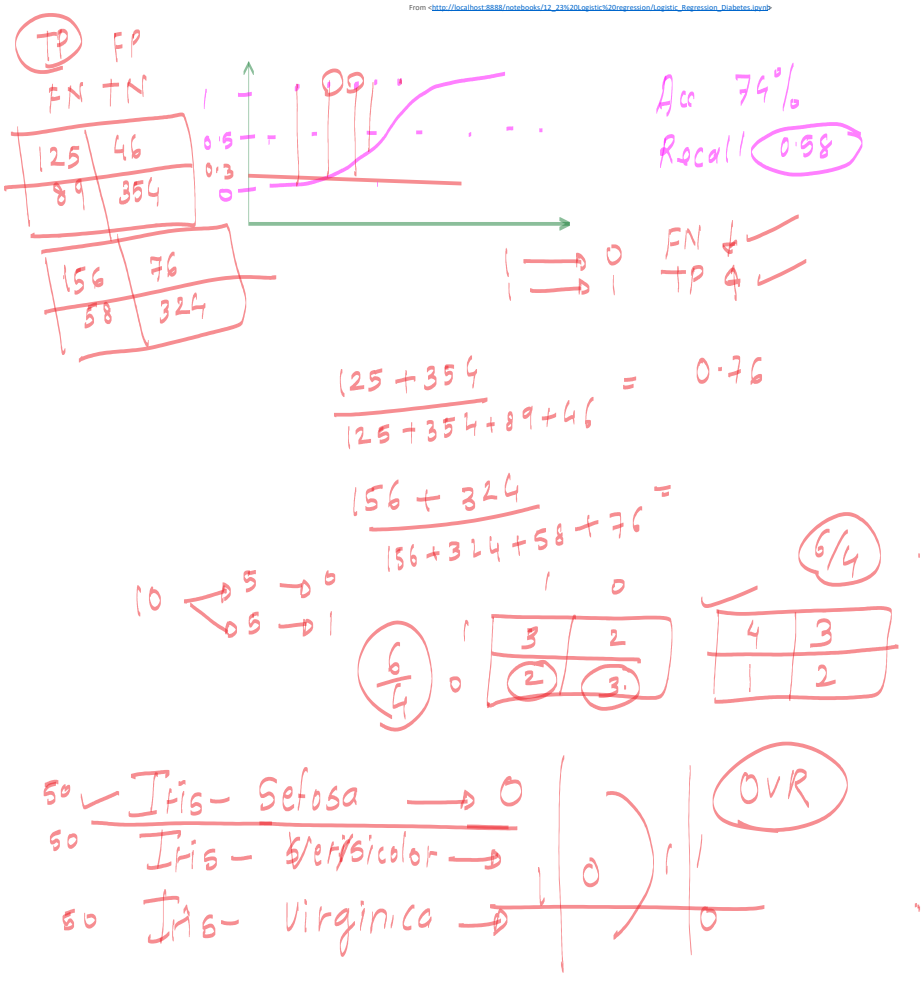
Non linear

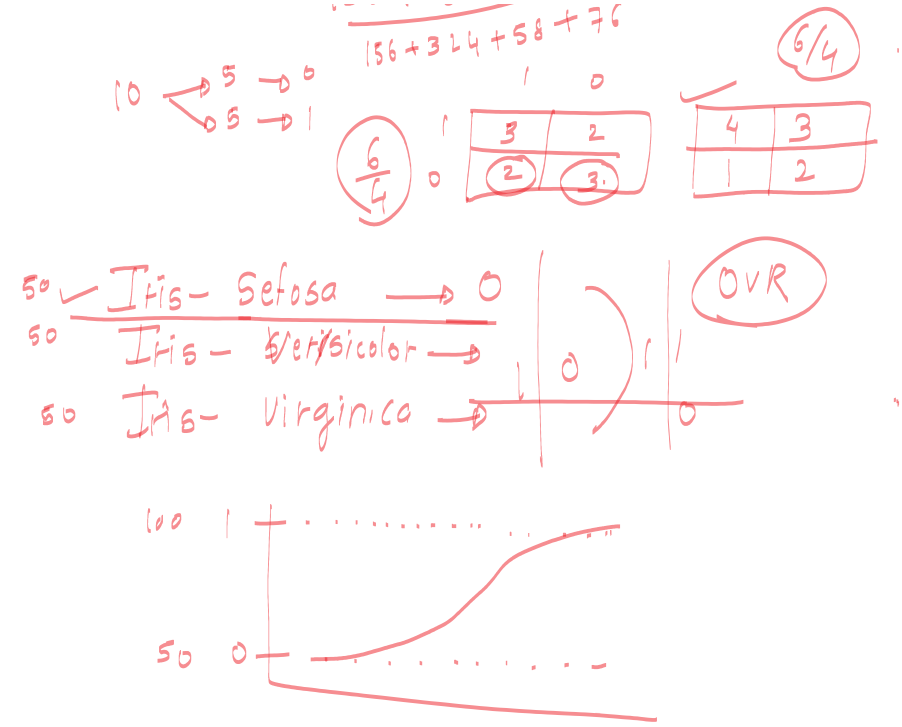
SVM

```
[[354  46]     TN  FP
 [ 89 125]]    FN  TP
```

```
[[324  76]
 [ 58 156]]
```

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Logistic_Regression_Diabetes.ipynb>



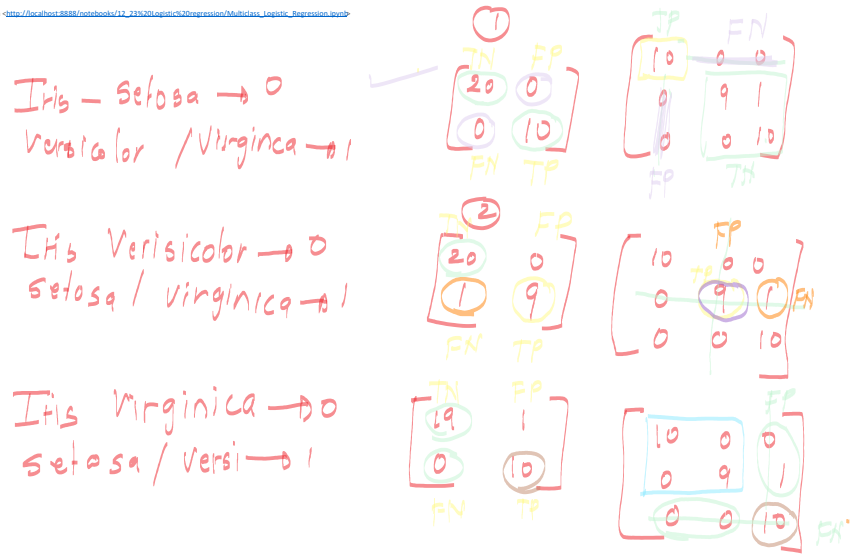


Confusion Matrix :
[[10 0 0]
[0 9 1]
[0 0 10]]
Multilabel Confusion Matrix :
[[[20 0]
[0 10]]
[[20 0]
[1 9]]
[[19 1]
[0 10]]]

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Multiclass_Logistic_Regression.ipynb>

Iris-setosa	1.00	1.00	1.00	10
Iris-versicolor	1.00	0.90	0.95	10
Iris-virginica	0.91	1.00	0.95	10

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Multiclass_Logistic_Regression.ipynb>

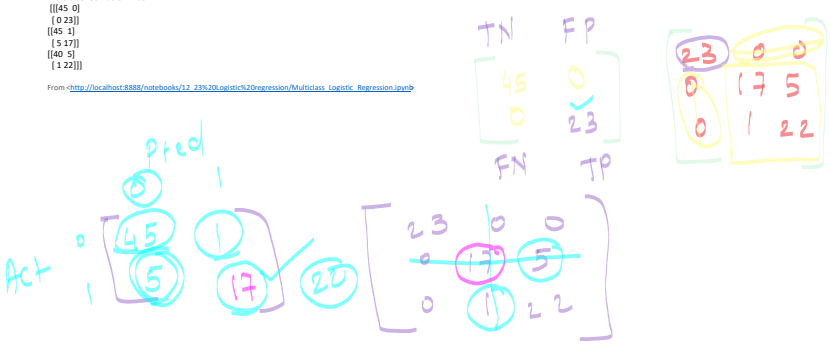


Confusion Matrix :
[[23 0 0]
[0 17 5]
[0 1 22]]
Multilabel Confusion Matrix :
[[[45 0]
[0 23]]
[[45 1]
[5 17]]
[[40 5]
[1 22]]]

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Multiclass_Logistic_Regression.ipynb>

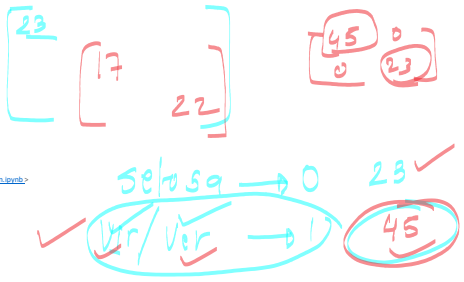
Iris-setosa	1.00	1.00	1.00	23
Iris-versicolor	0.94	0.77	0.85	22
Iris-virginica	0.81	0.96	0.88	23

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Multiclass_Logistic_Regression.ipynb>



```
Confusion Matrix :
 [[23  0  0]
 [ 0 17  5]
 [ 0  1 22]]
Multilabel Confusion Matrix :
 [[[45  0]
  [ 0 23]]
[[45  1]
  [ 5 17]]
[[40  5]
  [ 1 22]]]
```

From <http://localhost:8888/notebooks/12_23%20Logistic%20regression/Multiclass_Logistic_Regression.ipynb>



Logistic Regression

17 December 2022 19:40

$y = mx + c$

(Logistic) Sigmoid Curve

Age (x)	Cancer (y)
10	0
30	1
40	1
20	0
80	0
45	1
25	1

$$P = \frac{1}{1+e^{-(mx+c)}}$$

$P = \frac{1}{1+e^{-y}}$, $y = mx+c$

$P \rightarrow 0$ to 1

Best sigmoid curve.

$p(y=1) = 0.8$, $p(y=0) = 0.2$

$P = \frac{1}{1+e^{-(mx+c)}}$, $e^{-(mx+c)} \rightarrow 2.78$

$m = m - \alpha \frac{\partial CF}{\partial m}$

$c = c - \alpha \frac{\partial CF}{\partial c}$

$P \geq 0.5 \rightarrow 1$

$P < 0.5 \rightarrow 0$

$y_a = 1, y_p = 0$

$y_a = 0, y_p = 1$

y_a, y_p error = 0

C.F → Linear Regression = MSE $\frac{\sum (y_a - y_p)^2}{N}$

C.F → Logistic Regression → Log Loss

↳ LOR → linear mode

↳ Sigmoid curve → $\frac{1}{1+e^{-(mx+c)}}$

log odds → logit Function → $\log\left(\frac{p}{1-p}\right) \rightarrow \frac{x}{1-x} \rightarrow 0$ to ∞

$$\text{Log Loss} = -\frac{1}{N}\left[\sum y_a \log(p) + (1-y_a)\log(1-p)\right]$$

$\frac{1}{1+e^{-y}} \rightarrow 0$ to 1

$P \geq 0.5 \rightarrow 1$

$P < 0.5 \rightarrow 0$

Blue

Red 0

$y_a = 1$ correct

$y_a = 1, y_p = 0$ misclassified

$y_a = 0$ correct

$y_a = 0, y_p = 1$ misclassified

Pred class 1 if $P \geq 0.5$

Pred class 0 if $P < 0.5$

$P = \frac{1}{1+e^{-y}}$

$$L.L = -\frac{1}{N}\left[\sum \underbrace{y_a \log(p)}_{0} + \underbrace{(1-y_a)\log(1-p)}_{0}\right]$$

$$L.L = -\frac{1}{N}\left[\sum \underbrace{y_a \log(p)}_{0} + \underbrace{(1-y_a)\log(1-p)}_{0}\right]$$

$y_a = 1 \rightarrow L.L = -y_a \log(p) = -\log p$

$$L.L = -\log p$$

$y_a = 0 \rightarrow L.L = -[(1-y_a)\log(1-p)$

$$L.L = -\log(1-p)$$

65% = 13/20

Total = 6
correct = 4

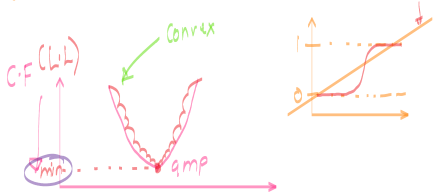
	y_a	p	y_p	
10	0 ✓	0.2 ✓	0 ✓	C
30	1 ✓	0.8 ✓	1 ✓	C
40	1 ✓	0.3	0 ✗	I (Mis)
80	0	0.9	1 ✗	I (Mis)
90	0 ✓	0.4	0 ✓	
70	1 ✓	0.7	1 ✓	

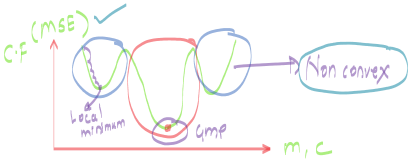
$p \geq 0.5 \rightarrow 1$
$p < 0.5 \rightarrow 0$

$$P(y) = \frac{1}{1+e^{-y}} \rightarrow \text{eqn of linear} \rightarrow mx+c$$

(linear model)
LR = $y \rightarrow mx+c$ range $-\infty$ to $+\infty$
0 to 1

Logistic (Sig curve) $\rightarrow P = \frac{1}{1+e^{-z}}$





✳ Confusion Matrix

Pred \ Actual	1	0
1	TP	FP
0	FN	TN

Positive → 1
Negative → 0

Pred \ Actual	1	0
1	TP	FP
0	FN	TN

Act \ Pred	0	1
0	TN	FP
1	FN	TP

sklearn

✳ Confusion Matrix

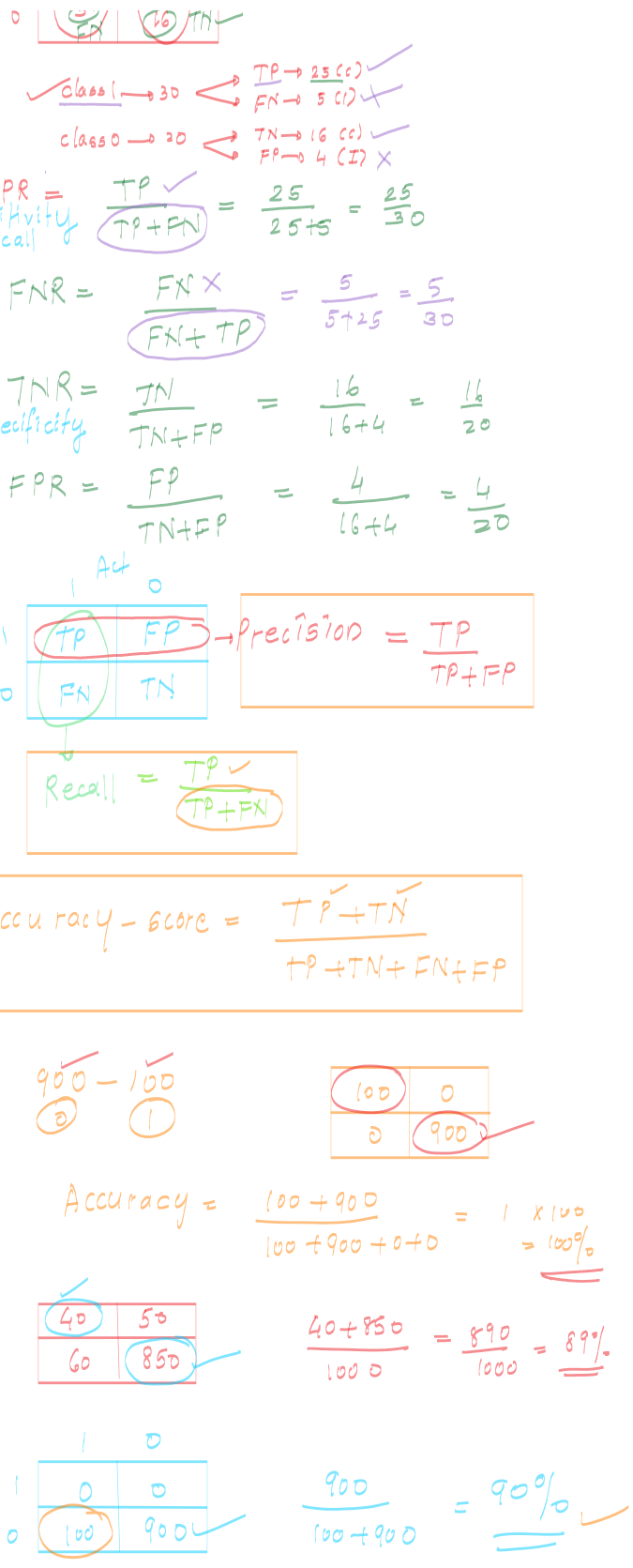
Pred \ Act	1	0
1	TP 25	FP 4
0	FN 5	TN 16

class 1 → 30 < TP → 25 (C) ✓ ; FN → 5 (I) ✗
class 0 → 20 < TN → 16 (C) ✓ ; FP → 4 (I) ✗

① TPR = Sensitivity / Recall $= \frac{TP}{TP+FN} = \frac{25}{25+5} = \frac{25}{30}$

② FNR $= \frac{FN}{FN+TP} = \frac{5}{5+25} = \frac{5}{30}$

③ TNR = TN 16


0
FN
16 TN
class 1 → 30
TP → 25 (C)
FN → 5 (I)
class 0 → 20
TN → 16 (C)
FP → 4 (I)
1. TPR = Sensitivity Recall = TP / (TP+FN) = 25/25+5 = 25/30
2. FNR = FN / (FN+TP) = 5/5+25 = 5/30
3. TNR = Specificity = TN / (TN+FP) = 16/16+4 = 16/20
4. FPR = FP / (TN+FP) = 4/16+4 = 4/20
Act
1
0
Pred
1
0
TP
FP
FN
TN
Precision = TP / (TP+FP)
Recall = TP / (TP+FN)
Accuracy-score = (TP+TN) / (TP+TN+FN+FP)
1. 900 - 100
2
1
100
0
0
900
Accuracy = 100+900 / 100+900+0+0 = 1 x 100 = 100%
2.
40
50
60
850
40+850 / 1000 = 890/1000 = 89%
3.
1
0
1
0
0
0
100
900
900 / 100+900 = 90%

(4)

	1	0
1	900	100
0	0	0

$\frac{900}{100+900} = 90\%$

50 - 50% Ideal Balanced

70 - 30
60 - 40
65 - 35

90 - 10
80 - 20 } Imbalanced

- SMOTE
- Over Sampling
- Under Sampling

80 - 80
20 - 20

* Classification Report :

Precision Recall F1-score

Class 0
Class 1

(1) Recall (↑) → Cancer / No Cancer

$\frac{TP}{TP+FN}$ ↓

	1	0	
1	40	10	FP
0	20	30	
	FN		

(2) Precision (↑)

$\frac{TP}{TP+FP}$ (↓)

Spam / Not Spam

	1	0	
1	40	15	FP
0	10	35	
	FN		

Offer letter → Spam → Not Spam

(3) f_1 - score.

f_β - score $= (1+\beta^2)\frac{P \times R}{(\beta^2 \times P)+R}$

$\beta = 1$

f_1 - score $= \frac{2PR}{P+R}$

$\frac{2xy}{x+y}$ Harmonic Mean

$f_1\text{-score} = \frac{2PR}{P+R}$

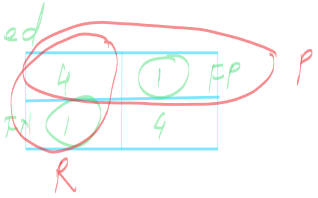
$\frac{x+y}{}$ Harmonic Mean

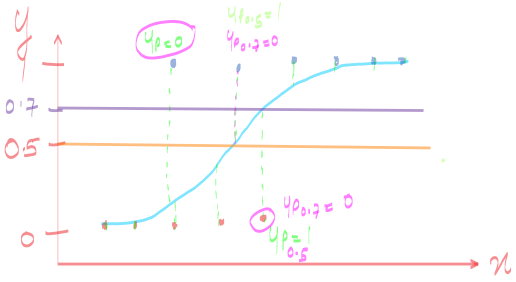
$\beta = 0.5$

$f_{0.5}\text{-score} = \frac{1.25\,PR}{0.25P + R}$

$\beta = 2, \quad f_2\text{-score} = \frac{5PR}{4P+R}$

Approved / Declined



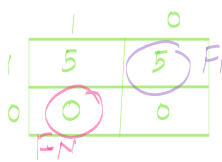


y_a	p	$y_p(0)$	$y_p(0.1)$	$y_p(0.2)$	$y_p(0.4)$	$y(0.5)$	$y(1)$
1	0.8	1	1	1	1	1	0
0	0.95	1	1	1	1	1	0
1	0.42	1	1	0	0	0	0
0	0.33	1	1	0	0	0	0
1	0.72	1	1	1	1	1	0
1	0.65	1	1	1	1	1	0
0	0.45	1	1	1	1	0	0
1	0.56	1	1	1	1	1	0
0	0.15	1	1	0	0	0	0
0	0.26	1	1	0	0	0	0

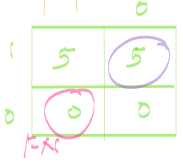
$P \geq 0.5$

$P < 0.5$

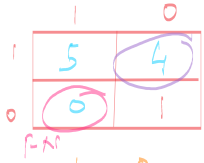
Th = 0



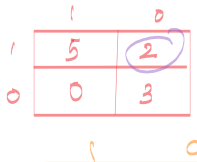
Th = 0.1



Th = 0.2



Th = 0.4



Th = 0.5

Th = 1

Precision

Th ↑ FP ↓ FN ↑

Th ↓ FP ↑, FN ↓

Recall

✳ ROC – AUC Curve

TPR (Sensitivity) ↑

Sensi ← TP | FP

FN | TN → Specificity

FN	TN

→ Specificity

TPR = TP / (TP + FN)

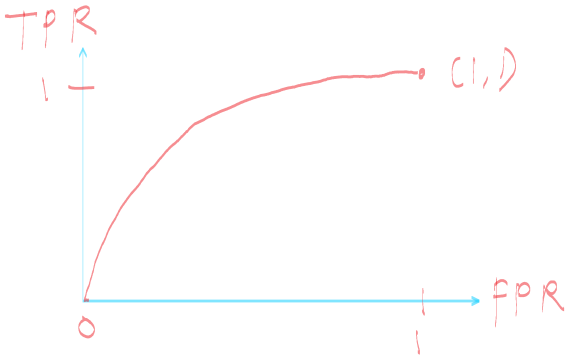
TNR = TN / (TN + FP)

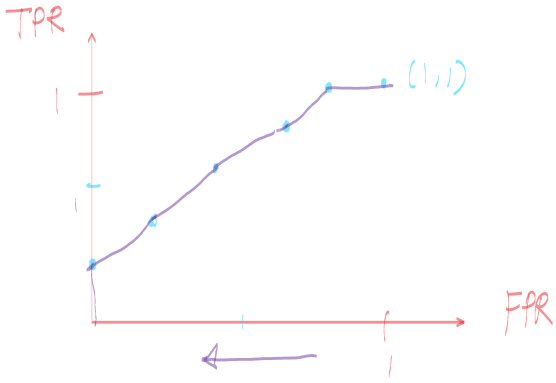
FPR = FP / (TN + FP)

FPR = 1 − TNR

TNR → Specificity

FPR = 1 − Specificity





Th = 0

5	5
0	0

TPR = 1

FPR = 1

Th = 0.2

5	4
0	1

TPR = 1

FPR = 0.1

Th = 0.4

4	3
1	2

TPR = 0.8

FPR = 0.6

Th = 0.5

3	2
2	3

TPR = 0.6

FPR = 0.4

Th = 0.7

2	1
3	4

TPR = 0.4

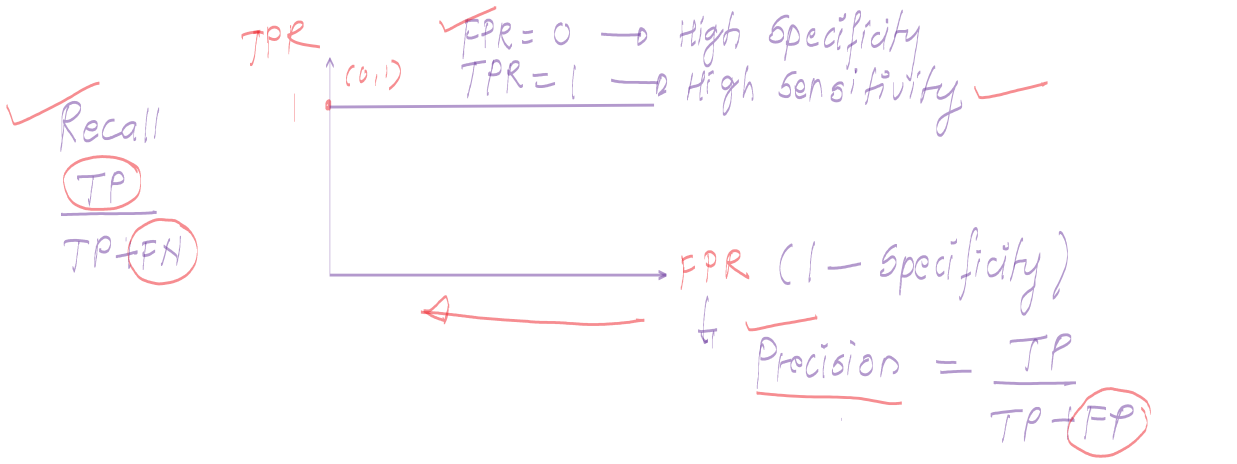
FPR = 0.2

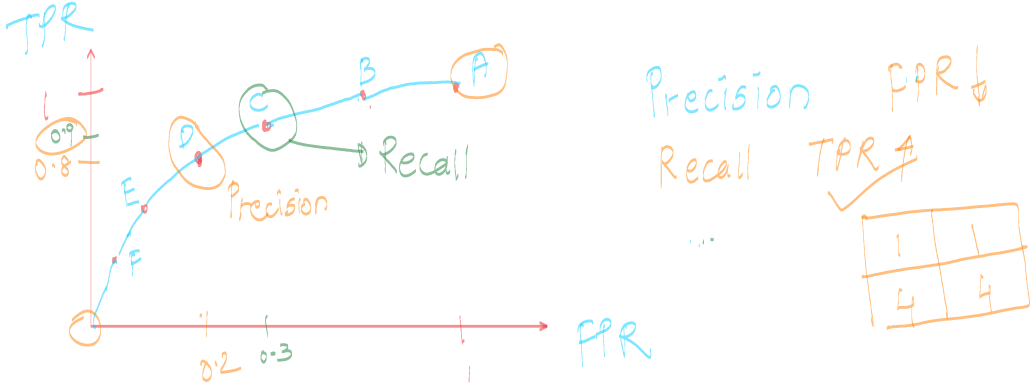
Th = 0.9

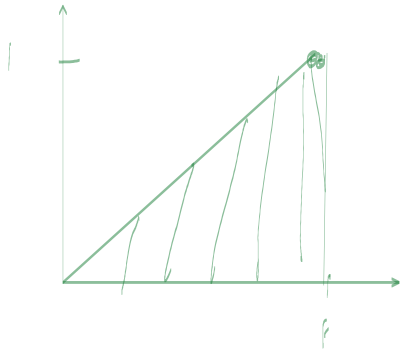
1	0
4	5

TPR =

FPR =

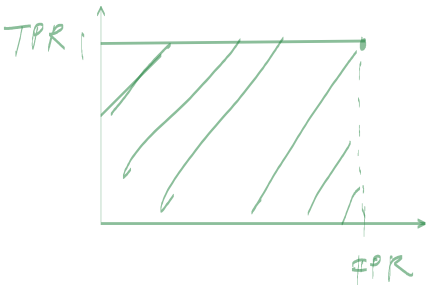




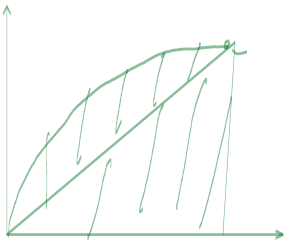


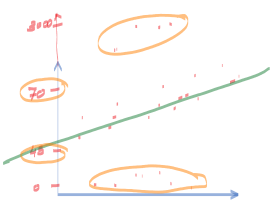
$\frac{1}{2} \times 1 \times 1 = 0.5$

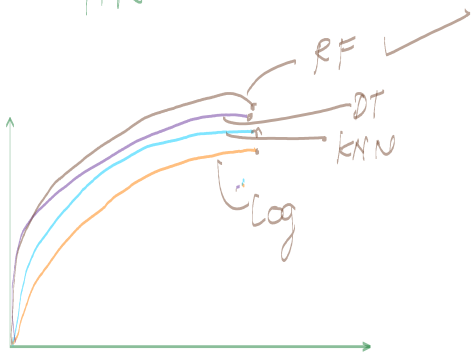
0.5 < AUC < 1



$1 \times 1 = 1$







* Classification → Target
  Imbalanced → 0 → 90, 80
  1 → 10, 20

* Classification → Target

Imbalanced → 0 → 90, 80 ✓

1 → 10, 20 ✓

Balanced { 50-50 Ideal Balanced
60-40
65-35

80-20 (Imbalanced)

SMOTE

✓ OverSampling (↑) 80-80 Balanced

✗ UnderSampling (↓) 20-20 Balancing

✓ 1,00,000

100 → 70-0, 30-1

Train → 70 → y_train → 0 70%, 1 30%

Test → 30 → y_test → 0 70%, 1 30%

70 × 70/100 = 49 → 0

70 × 30/100 = 21 → 1

80 0, 20 1

0 - 65, 1 - 35 → y → y_train → 0 - 65, 1 - 35%

y_test → 0 - 15%, 1 - 35%

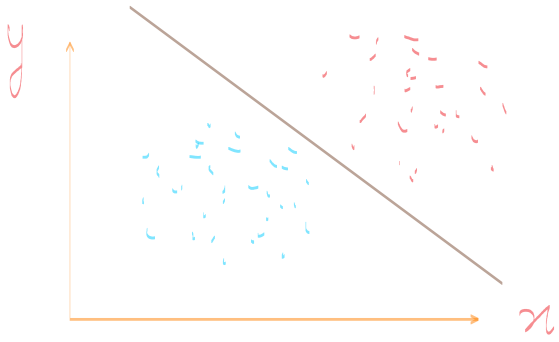
Recall = TP/(TP+FN) = 125/(125+89) =

[[354 46]
[89 125]]

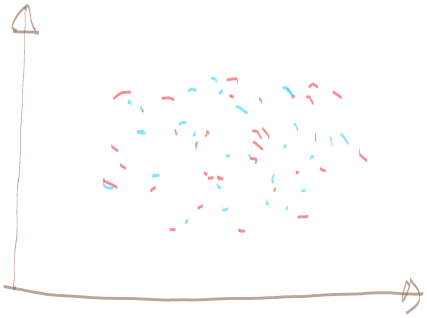
From <http://localhost:8888/notebooks/12_21%20Logistic%20Regression/Logistic%20Regression.ipynb#>

=

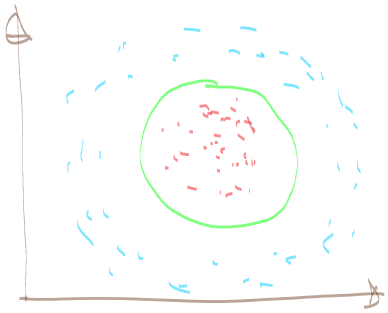
Act \ Pred	0	1	
0	TN	FP	354 46
1	FN	TP	89 125



linearly seperable

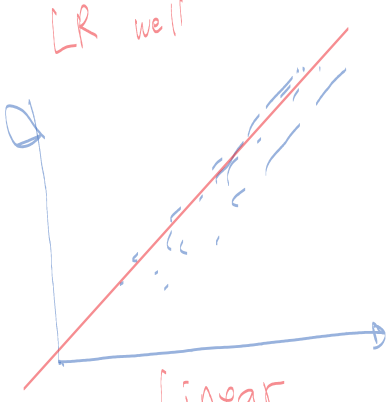


Non linearly seperable



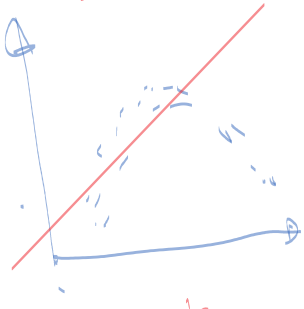
SVM

LR well



Linear

LR worst



Non linear

* Assumption

Train → 10 Features → Train 0.78 / test 0.75

→ Linearly 5 Corr., Non Corr → 1 drop, 2 drop, 3 drop } ciBil_score

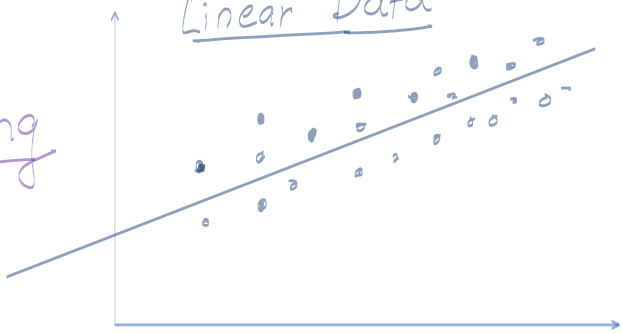
* Variance → Difference bet^n Accuracies

High Var → High Train and low Test
Low Train and High Test

Low Var → Low Train and low Test
High Train and High Test

* Bias → Training Accuracy

Linear Data



Overfitting

Train 90%

Test 70%

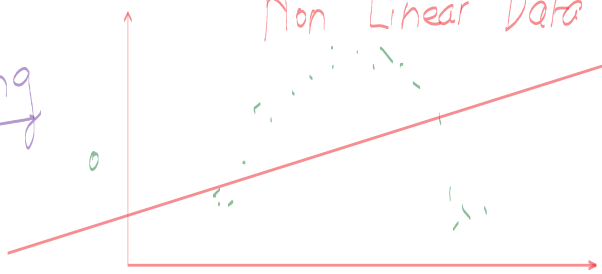
High Training Accuracy → Low Bias

High Variance → High Train, Low Test

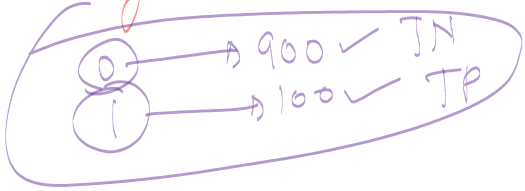
Non Linear Data - Train 60%

Test 58%

Underfitting



High Bias, Low Var → Low Train, Low Test

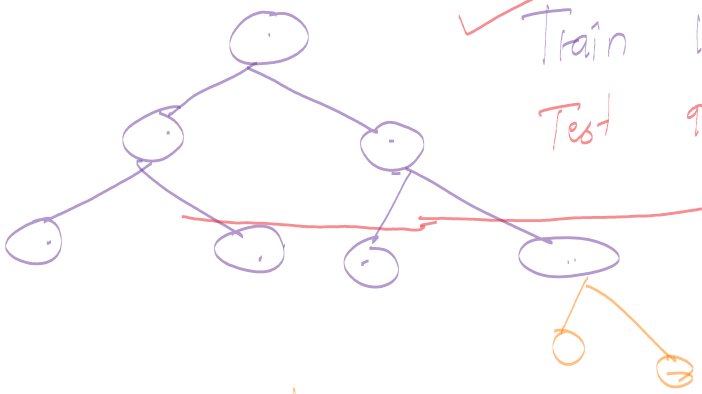


DT

Train 100%

Test 90%

Overfitting

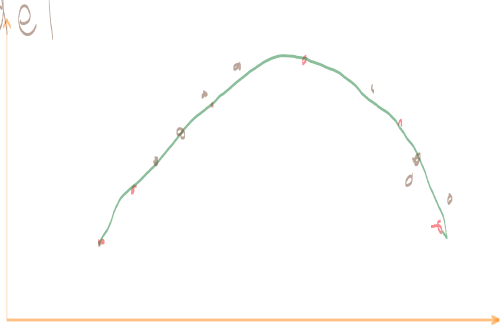


Pruning (Cut

Hyper Parameter Tuning

Train 100%

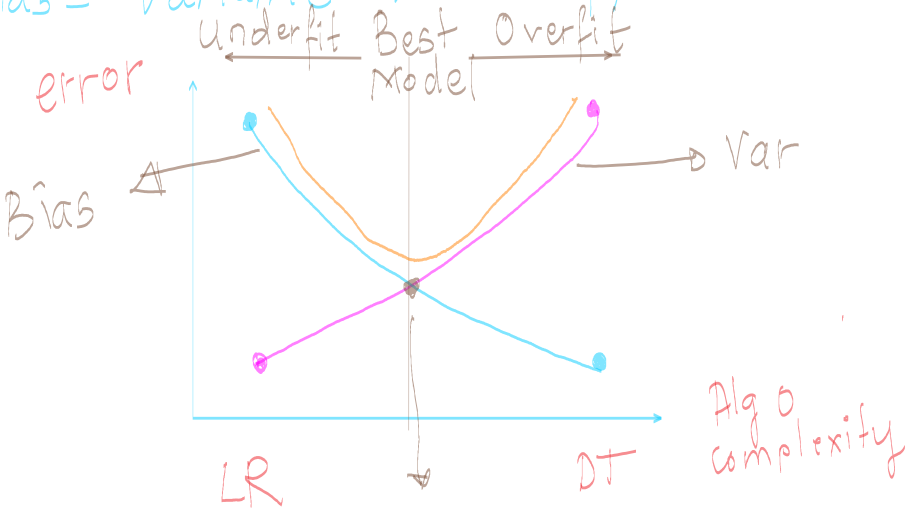
* Best Model



Train 98%
Test 95%

High Training → Low Bias ✓
→ Low Variance ✓

* Bias - Variance Trade-off



Low Bias, Low Var

Non - linear

LR → Und

High Bias, L

DT → Ove

Low Bias, H

$TE = Bias^2 + Var + Irreducible\ Error$

↓

Noise

* Cross Validation

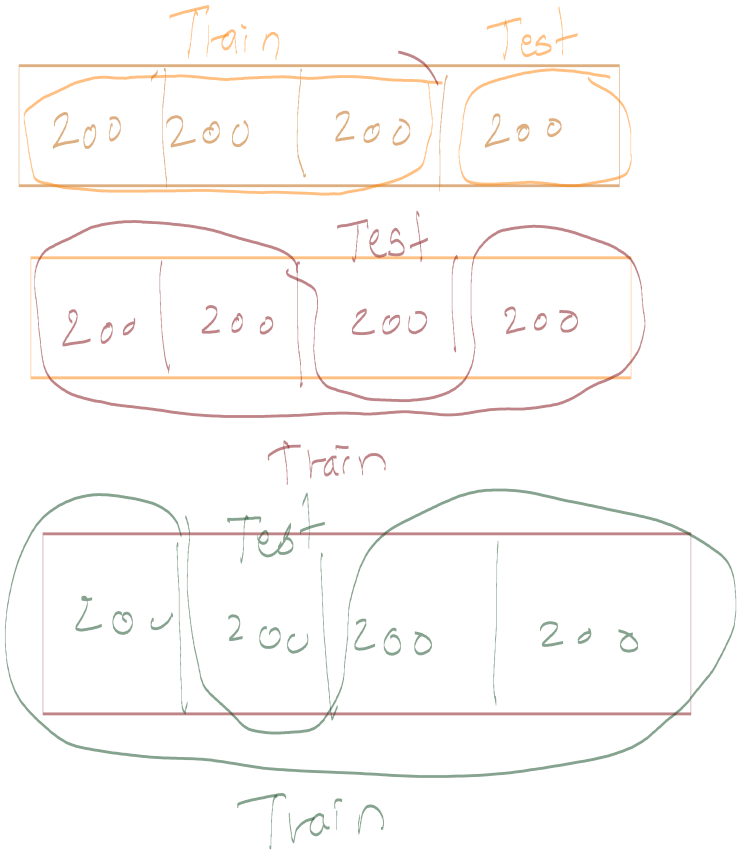
1000 Rows


er fit
ow var
r fit
igh var

cv = 4

800 Train

200 Test



Decision_Tree(min-sample_leaf = 2)

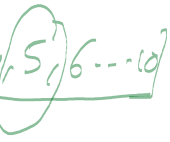
= min-sample-leaf = [2, 3, 4

Logistic_Train

0.5
0.36

* Binary Classification

Diabetes → 0
→ 1



Diabetes → 0, 1

* Multiclass Classification

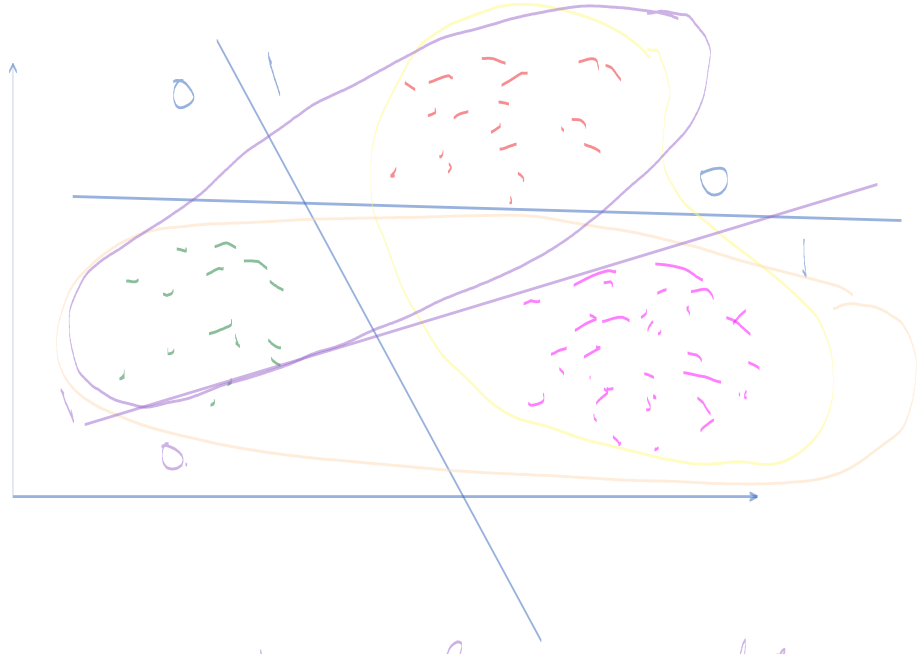
One Vs Rest

50 Iris - Setosa — 0 (50)

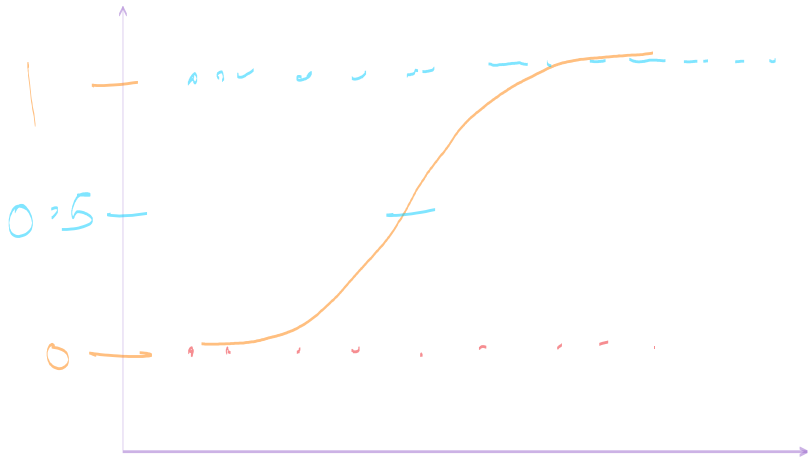
50 Iris - Versicolor } 1 (100)

50 Iris - Virginica

0 (50) | 1 (100)



Multilabel confusion matrix



From -http://localhost:8888/notebooks/12_23%20Logistic%20Regression%20MultiClass/Logistic%20Regression%20MultiClass%20Classification.ipynb#

Multilable Confusion Matrix :
[[[20 0]
[0 10]]
[[20 0]
[1 9]]
[[19 1]
[0 10]]]

From -http://localhost:8888/notebooks/12_23%20Logistic%20Regression%20MultiClass/Logistic%20Regression%20MultiClass%20Classification.ipynb#

Iris-setosa 1.00 1.00 1.00 10
Iris-versicolor 1.00 0.90 0.95 10
Iris-virginica 0.91 1.00 0.95 10

From -http://localhost:8888/notebooks/12_23%20Logistic%20Regression%20MultiClass/Logistic%20Regression%20MultiClass%20Classification.ipynb#

Confusion Matrix :
[[10 0 0]
[0 9 1]
[0 0 10]]

10 Iris-Setosa (0)
20 Iris-Versi/Virginica (1)

10 Iris-Versi (0)
20 Iris-Setosa/Virginia (1)

10 Iris-Virginica (0)
20 Iris-Setosa/Versi (1)

Confusion Matrix :
[[23 0 0]
[0 17 5]
[0 1 22]]

From -http://localhost:8888/notebooks/12_23%20Logistic%20Regression%20MultiClass/Logistic%20Regression%20MultiClass%20Classification.ipynb#

[[[45 0]
[0 23]]
[[45 1]
[5 17]]
[[40 5]
[1 22]]]

From -http://localhost:8888/notebooks/12_23%20Logistic%20Regression%20MultiClass/Logistic%20Regression%20MultiClass%20Classification.ipynb#

0 → 23
1 → 45

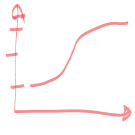
Linear Data
Linear Regression

Linearly Seperable

Non-linearly Data

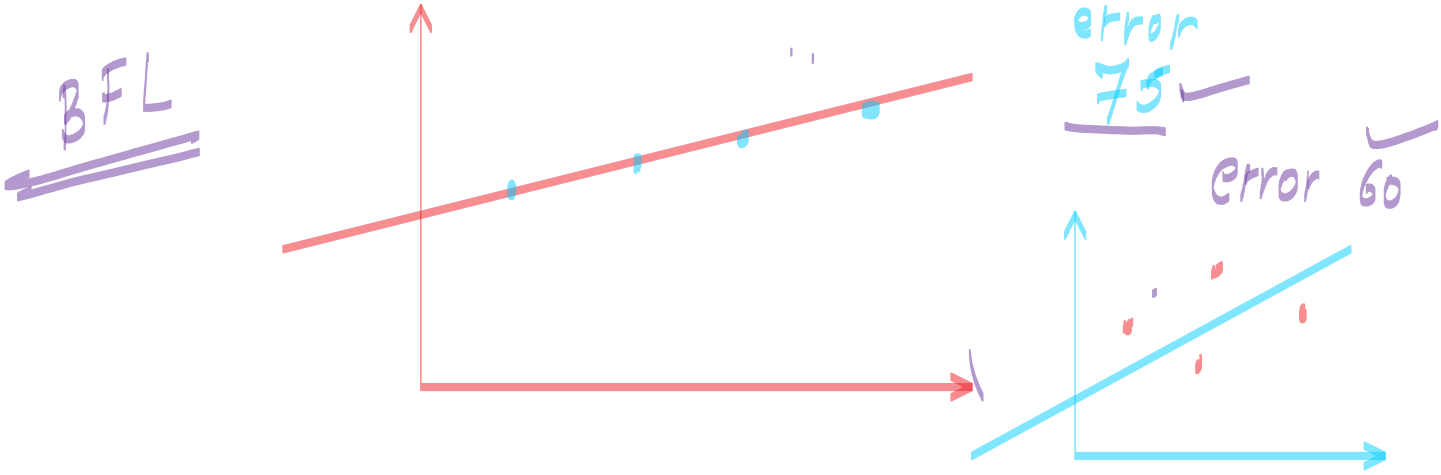
(50) Iris-Setosa → 0

(100) (Iris-Versi / virginica) → 1



Regularization

26 December 2022 07:30



* Regularization → Linear Models

Linear Regression

Logistic Regression

Overfitting → Low Bias, High Var

High Train

Low Test

Var

90%

70%

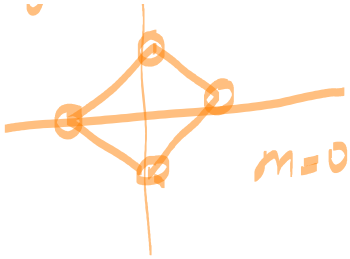
80% 80%

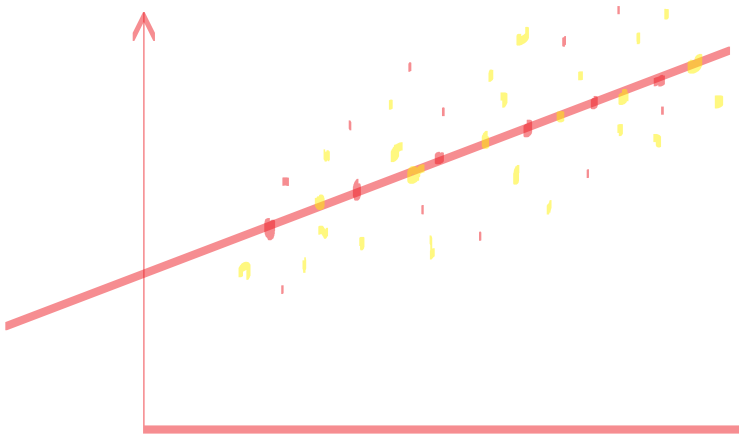
85% 76%

1. Lasso Regression (L1 Regularization)
2. Ridge Regression (L2 Regularization)

m≠0

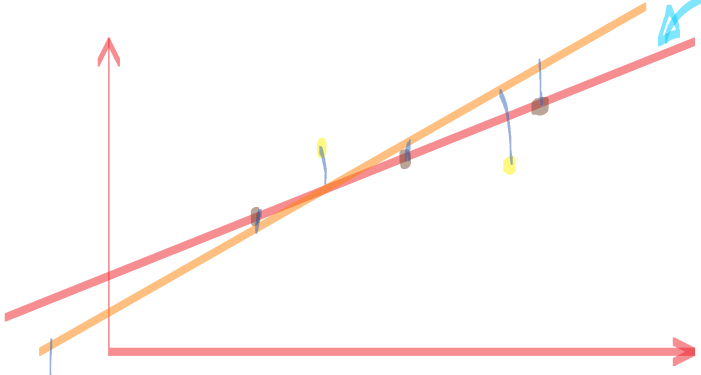
* Lasso Regression





0.8 Train
0.75 Test

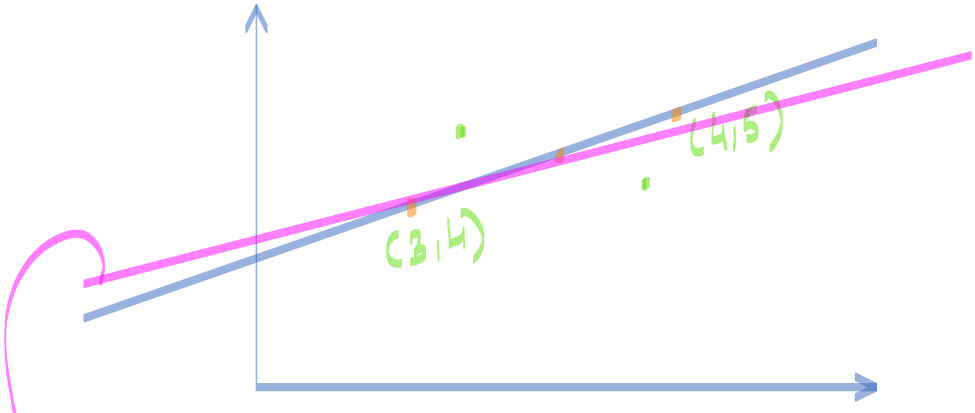
LOR → Linear.



Train error = 0
↳ 100%

Test error = 10
↳ 80%

Train → 8 = error → 91% } var ↓
Test → 6 = error → 85% }



LOR

$m = \frac{5-4}{4-3} = 1$

$= \frac{5.5-5}{4-3} = 0.8$

CF = MSE → Linear Reg

CF = SSE + λ|m| → slope

λ → 1 → m^2 → Ridge

LR

① CF = SSE + λ|m| → Linear model

= 0 + λ|1.4|

= 1.4

reducing slope

② CF = 0.2 + (λ|0.8|

= 1.0

③ CF = 0.3 + (λ|0.6|

= 0.9

④ CF = 0.4 + (λ|0.4|

= 0.8 → optimum

⑤ CF = 0.6 + (λ|0.3|

= 0.9

$m_2 \rightarrow 1$ → reduce → 0

$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + c$

= 0

Lasso Regression

Feature Selection Technique

→ Feature Selection Technique

Embedded

Correlation → Before Training

* Ridge Regression.

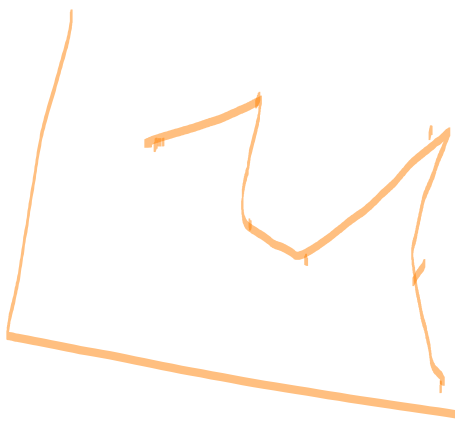
→ L2 Reg.

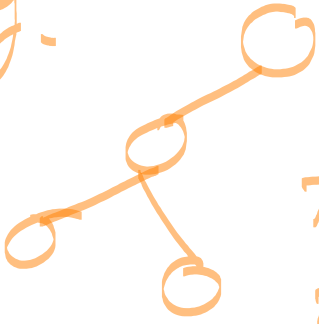
CF = SSE + λ m^2 → Ridge

CF = SSE + λ |m| → Lasso
→ L1 Reg

$y = m_1x_1 + m_2x_2 + m_3x_3$ $m \neq 0$

Overfitting.



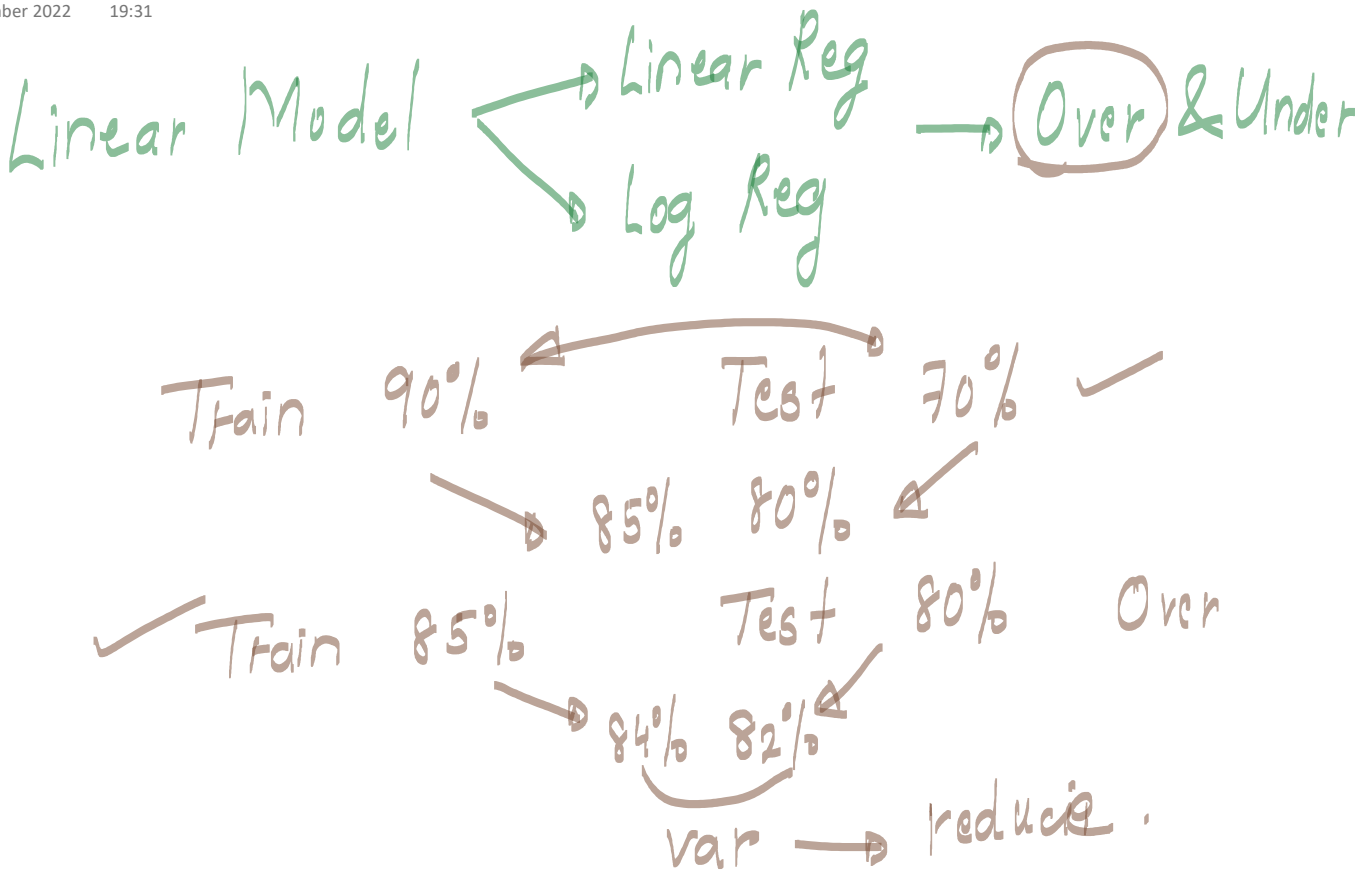


Train 100%

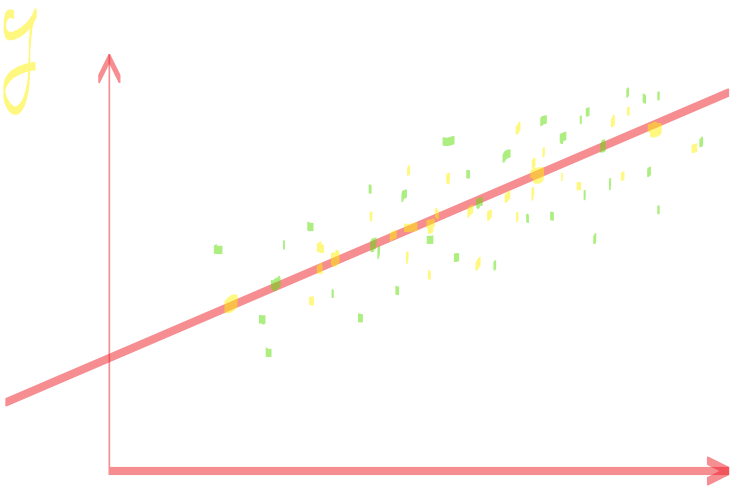
Test

Regularization

26 December 2022 19:31



* Lasso Regression (L1 Regularization)
  ↳ Least Absolute shrinkage and selection Operator.



Train → 90%
Test → 88%

Linear Regression

Train, error = 0
100%

Test, error = 10
89%

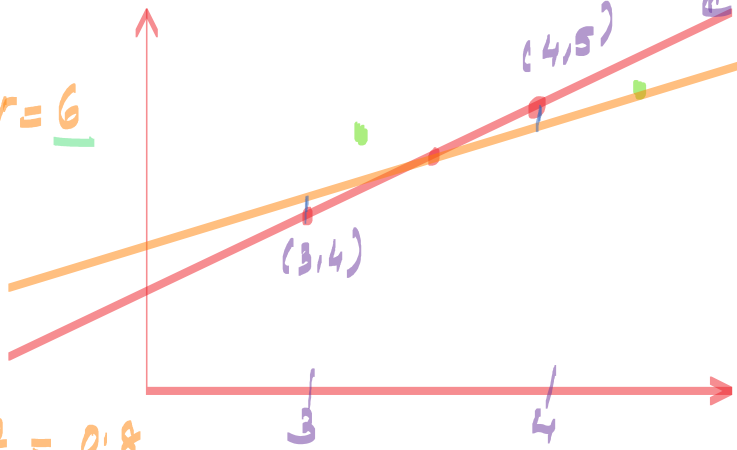
$m = \frac{5-4}{4-3} = 1$

SSE
$(y_a - y_p)^2$

5 pts

Train, error = 6
95%

Test, error = 8
92%



$m = \frac{5-4.2}{4-3} = 0.8$

CF → MSE, error

penalty

Regularization Paramter ← penalty factor → lambda / alpha

slope → LOR

* Lasso Regression

$$CF = SSE + \lambda |m|$$

$$C.F = SSE + |m|$$

① $C.F = 0 + 1 \times |1.4|$ → error (LOR)
$= 1.4$ ✓ — $m = 1.4$

② $C.F = 0.2 + 1 \times |0.8|$
$= 1$ ✓ — $m = 0.8$

③ $C.F = 0.3 + 1 \times |0.6|$
$= 0.9$ ✓ — $m = 0.6$

④ $C.F = 0.4 + 1 \times |0.4|$
$= 0.8$ ✓ — $m = 0.4$

⑤ $C.F = 0.6 + 1 \times |0.3|$
$= 0.9$ ✓ — $m = 0.3$

$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + c$.
$m_2 = 0$

→ Lasso Reg → Feature Selection
Ensembled → Random Forest

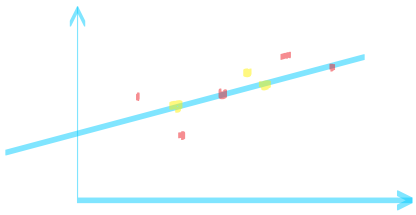
* Ridge Regression

$$C.F = SSE + \lambda m^2$$

← L2 Regularization

non-zero

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3$$

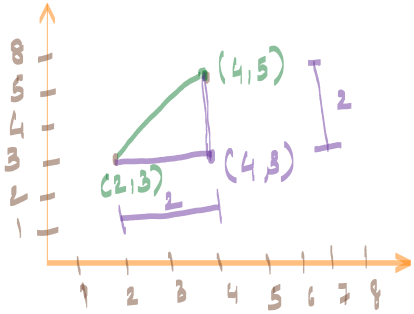


80% ✓
85% ✓

* Distance Based Algorithm

KNN → Regression
→ Classification

① Euclidean Distance (p=2) → Default.
② Manhattan Distance (p=1)



$P_1(x_1,y_1) = (2,3)$
$P_2(x_2,y_2) = (4,5)$

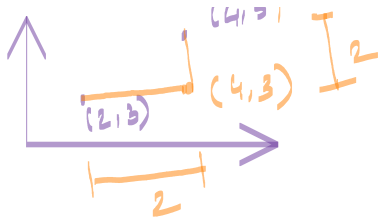
① Euclidean distance

$$ED = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$
$$= \sqrt{(4-2)^2 + (5-3)^2} = \sqrt{4+4} = \sqrt{8} = 2.8$$

$P_1(5,1)$, $P(4,5)$

$$ED = \sqrt{(4-5)^2 + (5-1)^2} = \sqrt{1+16} = \sqrt{17} = 4.1$$

② Manhattan distance



$2+2=4$

MD > ED

$P_1(2,4,3)$, $P_2(6,7,8)$

$x_1\ y_1\ z_1$ $x_2\ y_2\ z_2$

$ED = \sqrt{16+9+25} = \sqrt{50} = 5\sqrt{2}$

$ED = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$

$MD = |x_1-x_2| + |y_1-y_2| + |z_1-z_1|$

$= |6-2| + |7-4| + |8-3|$

$= 4+3+5 = 12$

MD > ED

* Minkowski Distance

$D_1 \quad D_2$

$\left(\sum (D_1-D_2)^p\right)^{1/p}$

ED $\Rightarrow p=2$

$\left(\sum (D_{x_1y_1}-D_{x_2y_2})^2\right)^{1/2}$

$$ED = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

MH $\Rightarrow p=1$

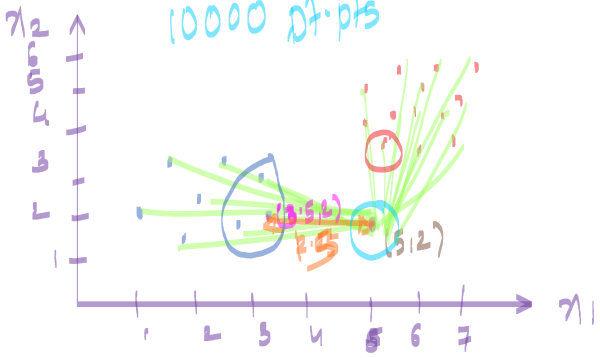
$\left(\sum (D_1-D_2|^1)\right)^{1/1}$

MD = $|D_1-D_2|$

$MD = |x_1-x_2| + |y_1-y_2|$

$$MD = |x_2 - x_1| + |y_1 - y_2|$$

KNN



$$ED = \sqrt{(5-3.5)^2 + (2-2)^2}$$
$$= \sqrt{1.25} = 2.25$$
$$= 1.5$$

x_1	x_2	y
–	–	B
–	–	R
–	–	B
–	–	B
–	–	R
–	–	B

	E.D	y_a	
1	2.3	B	5
2	6	B	
3	2.5	B	
4	4	B	
5	6.2	B	
6	8.5	R	3
7	1.2	R	
8	1.8	R	

K=5 ✓ odd.

No. of neighbours

B → 3 ✓ up

R → 2

B → 4 ✓ up

R → 2

K=6

B = 3 → No. of Dt·pt = 5 ✓

R = 3 → No of Dt·pt = 3

Tens x_1	Units x_2	Thous. x_3
10	2	15000
20	4	20000
30	6	30000

P_1 (10, 2, 15000)
(x_1 x_2 y_1)

P_2 (20, 4, 20000)

P_3 (30, 6, 30000)

20	4	20000
30	6	30000
25	5	25000

P_3 (30, 6, 30000)

4 (25, 5, 25000)
x_2 y_2 z_2

$$ED = \sqrt{(25-10)^2 + (5-2)^2 + (25000-15000)^2}$$

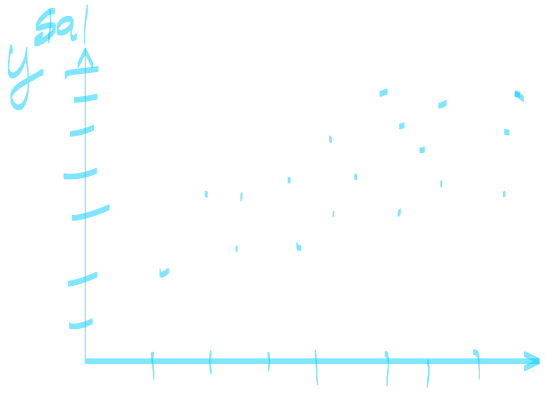
$$= \sqrt{225 + 9 + 100000000}$$

$$= \sqrt{10000234} = \text{_____}$$

Feature Scaling.

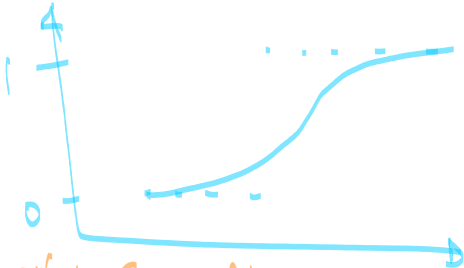
[4, 6, 10, 20, 25]

0 10 1



$y = mx + c$

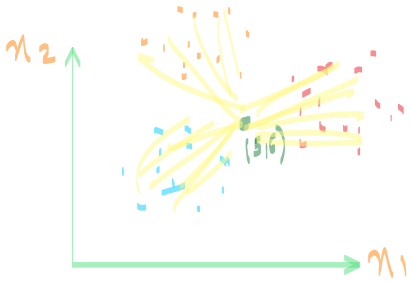
$MSE = \frac{\sum (y_a - y_p)^2}{N}$

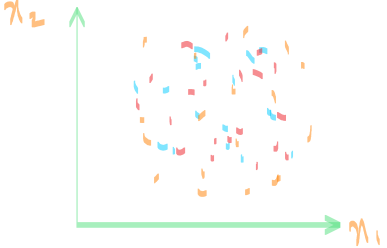


$p = \frac{1}{1+e^{-y}}$

Linearly Seperable

Non - Linearly seperable





① It will save the data (training)

(1) It will save the [illegible]

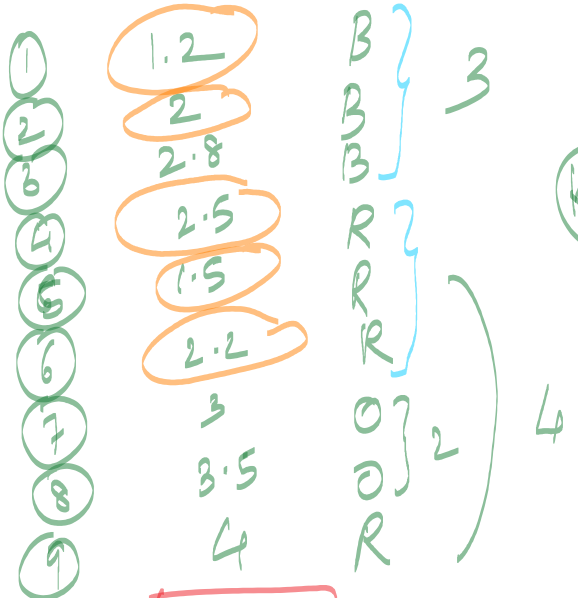
x_1	x_2	y	
2	10	B	Training
3	4	R	
4	8	O	
5	6	?	Test

(2) It will find distance between Test pt. & all the training dt pts.

(3) Arrange distances in ascending

(4) nearest neighbours, $K=5$

(5) Majority classification → yp



$K=5$

B → 2
R → 3 ✓ yp
O → O

$K=8$

B = 3. } No. of Dt. pts = 3
yp ~~R = 3.~~ } No. of Dt. pts = 4
O = 2

$K=8$

~~R = 3~~ No. of pts = 4.

K = 8

B = 3 No. of pts = 4

R = 3 No. of pts = 4

O = 2

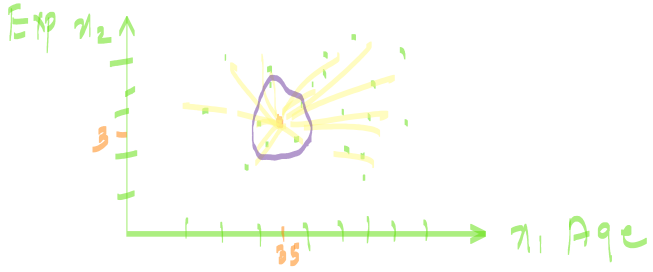
y_p 1.2 + 1.8 + 2 = 5.0

2 + 1.2 + 2.5 = 5.7

km	m
2	2000
3	3000
4	4000

cm	ft
180	6
150	5

* KNN Regression



Distance	n_1 (Age)	n_2 (Exp)	y (Salk)
1.2	30	4	30
2	40	5	40
3	50	6	50
1.8	20	2	25
4	25	2.5	30
3.2	45	5	55
4.1	35	4	50
35	3	?	

K = 5

Training

y_p = 40

$y_{p\,mean} = \frac{30+40+50+25+55}{5}$

= 40

	Xi	Xnew
	148	0.743
1	85	0.427
2	183	0.91
3	150	0.753
4	150	0.66

Xmin = 85
xmax = 183

0.056

Norm → $\frac{X_i - X_{min}}{X_{max} - X_{min}}$

$= \frac{148 - 85}{1830 - 85}$

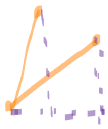
$= 0.036$

Sensitive to Outliers

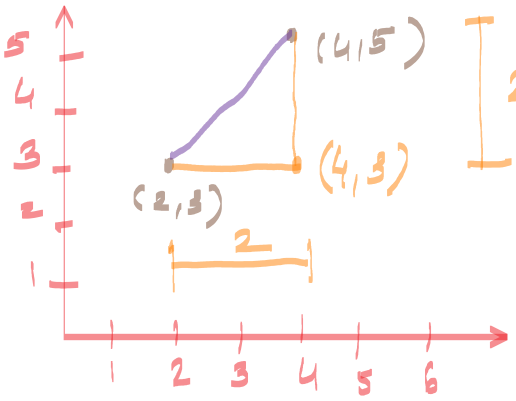
K = 3

ED =

1.2	2
2	3
3.1	3.8
3.5	3.9
5.8	4



* KNN → Regression — Continous
→ Classification — Categorical



$2+2=4$

* Euclidean Distance

$P_1(x_1, y_1)$ (2, 3) , $P(x_2, y_2)$ (4, 5)

$$ED = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

$$= \sqrt{(4-2)^2 + (5-3)^2} = \sqrt{4+4}$$

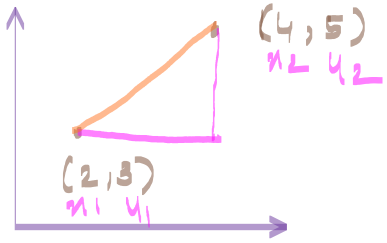
$$= \sqrt{8} = 2.8$$

$P_1(5, 1)$, $P_2(4, 5)$

$$ED = \sqrt{(4-5)^2 + (5-1)^2} = \sqrt{1+4^2} = \sqrt{17}$$

* Manhattan Distance

$$MD = |x_1 - x_2| + |y_1 - y_2|$$



$$MD = |2-4| + |3-5|$$
$$= 2 + 2$$
$$= 4$$

MD > ED

$P_1(2, 4, 3)$, $P_2(6, 7, 8)$
$x_1\ y_1\ z_1$ $x_2\ y_2\ z_2$

$P_1(2, 4, 3)$, $P_2(6, 7, 8)$

$x_1\ y_1\ z_1$ $x_2\ y_2\ z_2$

$$ED = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

$$= \sqrt{(6-2)^2+(7-4)^2+(8-3)^2}$$

$$= \sqrt{16+9+25} = \sqrt{50} = 5\sqrt{2}$$

$$MD = |6-2| + |7-4| + |8-3|$$

$$= 4+3+5 = 12$$

MD > ED

(1) Minkowski Distance.

$$\left(\Sigma |D_1 - D_2|^p\right)^{1/p}$$

$p=2 \longrightarrow$ Euclidean Distance

$$ED = \left(\Sigma |D_1 - D_2|^2\right)^{1/2}$$

$$= \sqrt{\Sigma |D_1 - D_2|^2}$$

(D_1: $x_1 y_1$, D_2: $x_2 y_2$)

$$ED = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

$p=1 \longrightarrow$ Manhattan Distance

$$MD = \left(\Sigma |D_1 - D_2|^1\right)^{1/1}$$

(D_1: $x_1 y_1$, D_2: $x_2 y_2$)

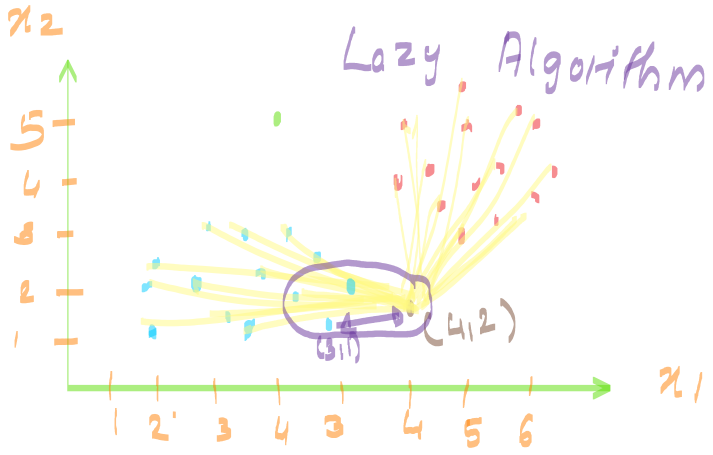
$$MD = |x_1-x_2| + |y_1-y_2|$$

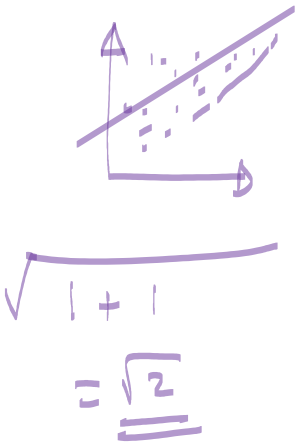
$$MD = |x_1 - x_2| + |y_1 - y_2|$$

MD ≥ ED

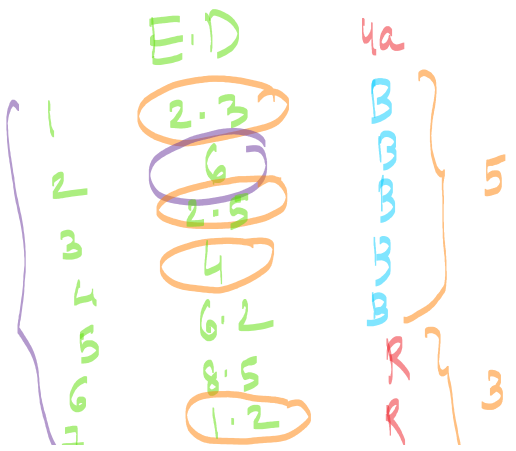
P = 2 → Default

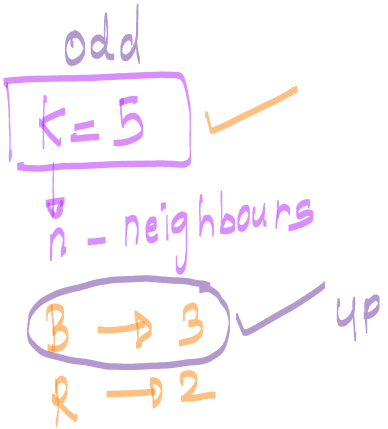
KNN



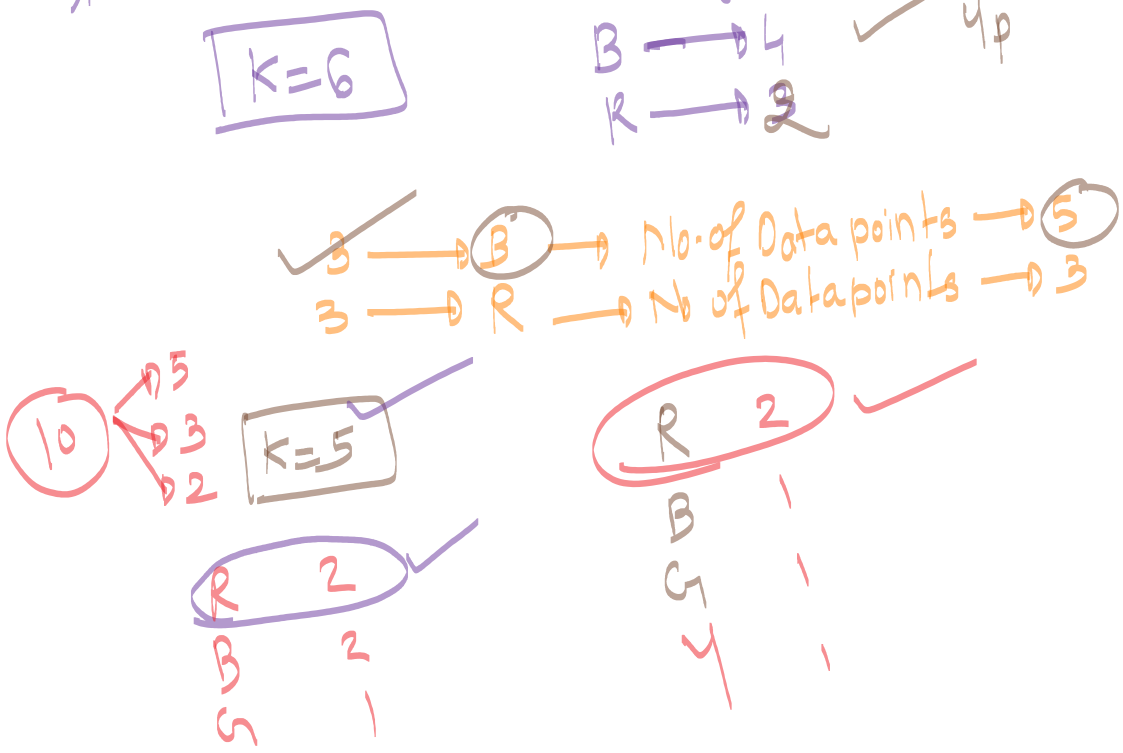


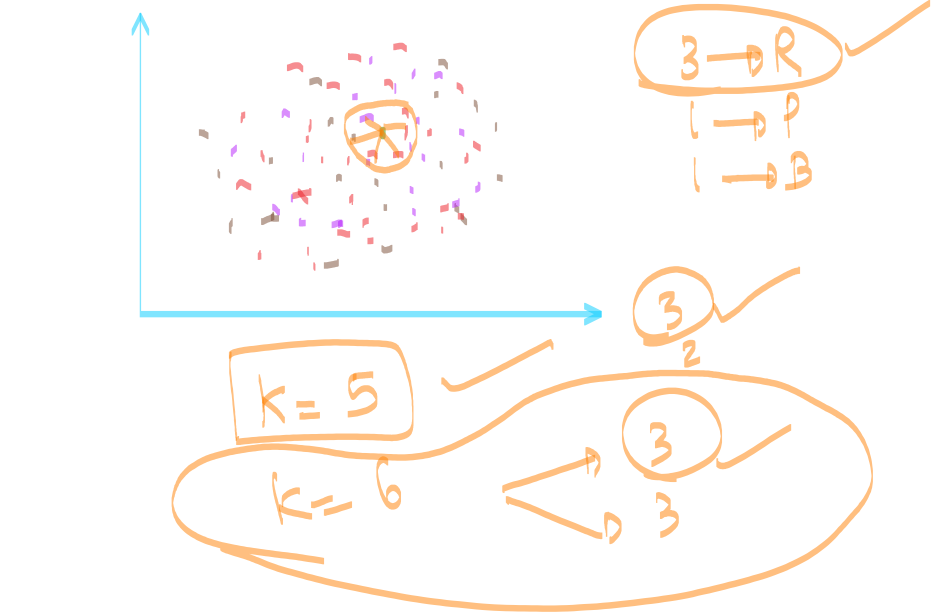
x_1	x_2	y
—	—	B
—	—	R
—	—	B
—	—	B





* Distance Based Algorithm.





* Feature Scaling.

x_1	x_2	x_3
10	1	15000
20	2	20000
30	3	30000

P_1 $(\underset{x_1}{10}, \underset{x_2}{1}, \underset{x_3}{15000})$

P_2

P_3

30 3 30000 p_3

25 4 25000 $T(25,4,25000)$ $u_1\ u_2\ u_3$

$\sqrt{(25-10)^2+(4-1)^2+(25000-15000)^2}$

$= \sqrt{15^2+3^2+10000^2}$

$= \sqrt{225+9+100000000}$

$= \sqrt{100000234}$

np.log

[1, 2, 3 .... 4, 9, 15, 20]

✓ [0, 0.1, 0.3 .... 0.5, 0.8, 0.10, 0.1]

* Feature Scaling

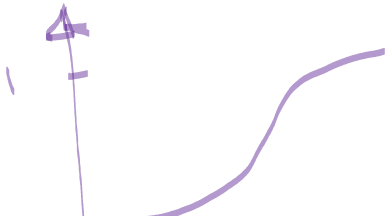
1. Normalization → MinMax Scalar
2. Standardization → Standard Scaler.

* Gradient Descent Algorithm

$y = mx + c$

$\frac{\partial c \cdot f}{\partial m}$

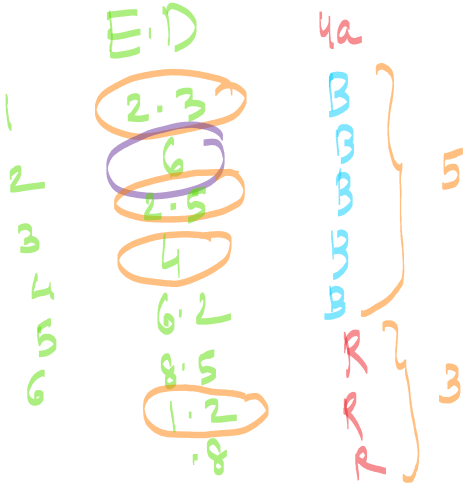
$MSE = (y_a - y_p)^2$

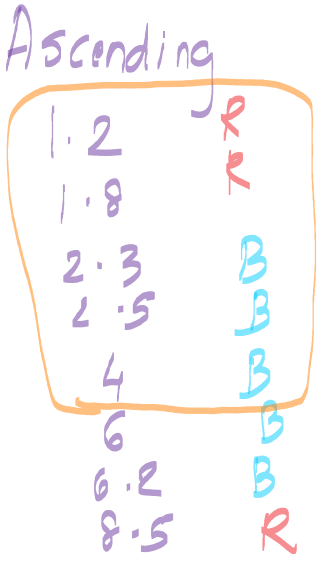


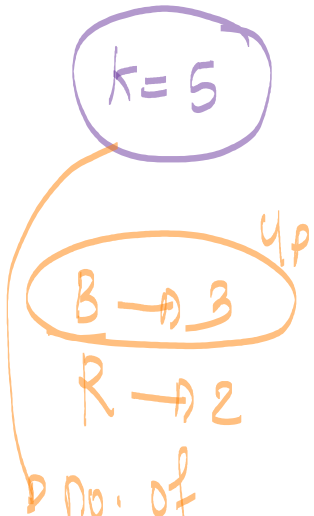
$p = \frac{1}{1+e^{-y}}$

$y = mx + c$

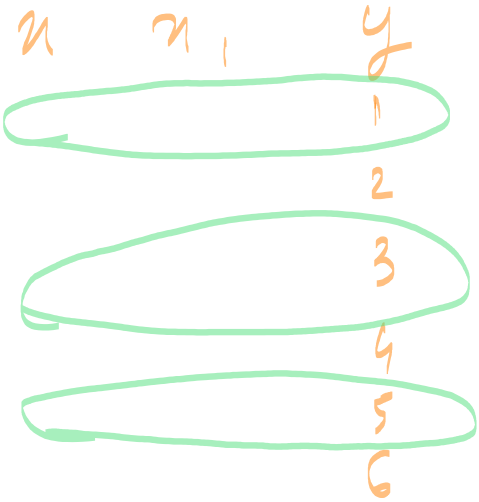
✓ $LL = -\log P$ }
✓ $LL = -\log(1-P)$







n_neighbours

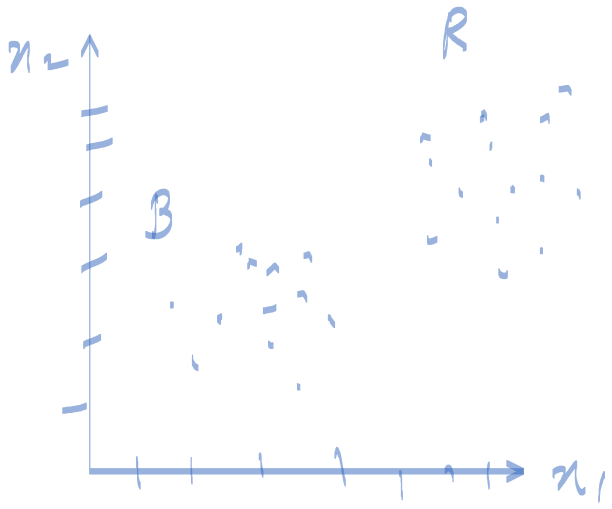


$k = 3$

$\frac{1+3+5}{3} = \frac{9}{3} = 3$

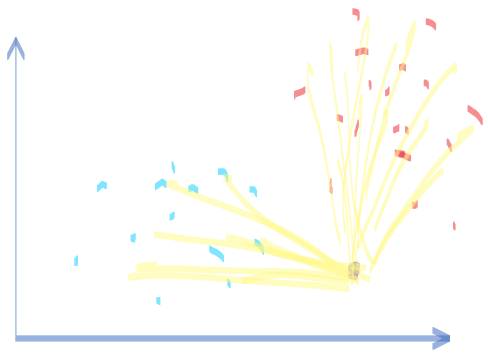
① It will scatter plot and save training data.

x_1	x_2	y
—	—	B
—	—	R
—	—	B
⋮		
—	—	R



② Testing pt. for prediction

Training data



x_1	x_2	y
2	20	R
3	30	B
4	50	B
5	60	R
6	80	R
7	40	?

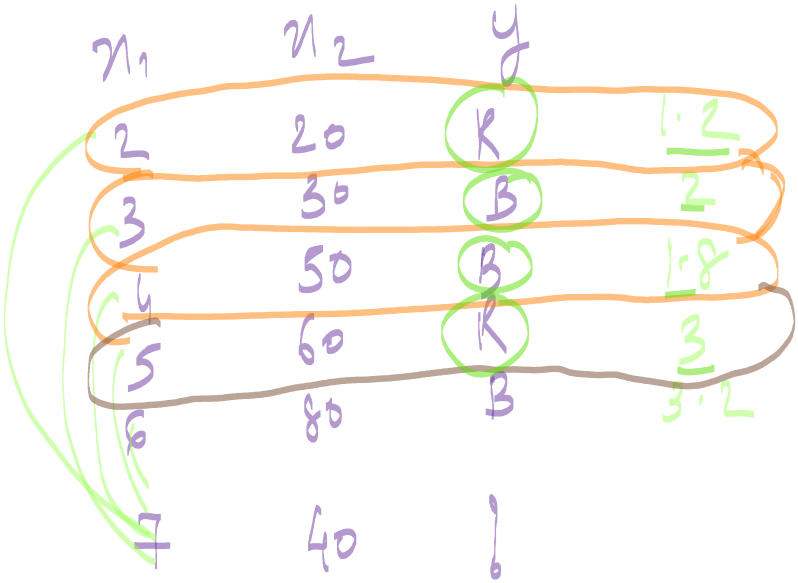
Odd

K=3

n_neighbors

B → 2 ✓ yp

R → 1



K=4

Training data

Training data

y_p ✓ R = 2 } No. of dt. pts 80

B = 2 } No of dt. pts 60

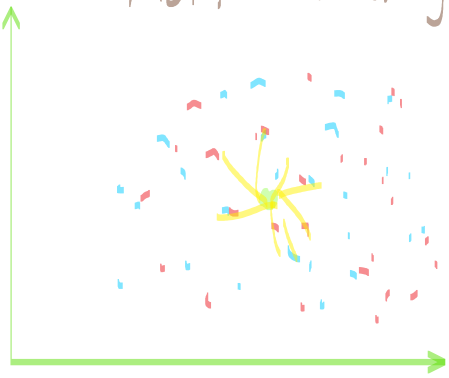
Setosa 50 } R → 2 } → 1.2 + 3 = 4.2

Versicolor 50 } B → 2 } → 2 + 1.8 = 3.8

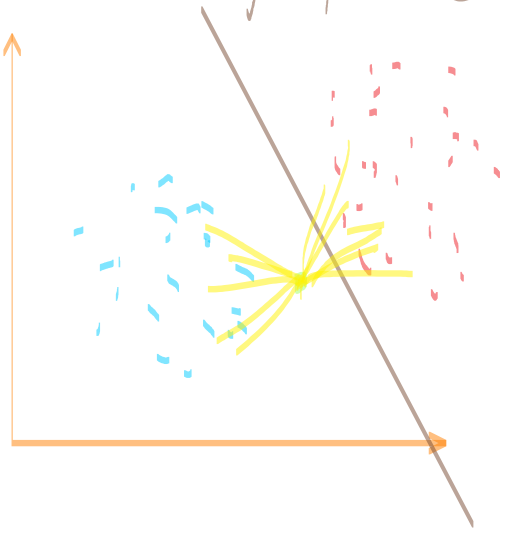
y_p

* Regression

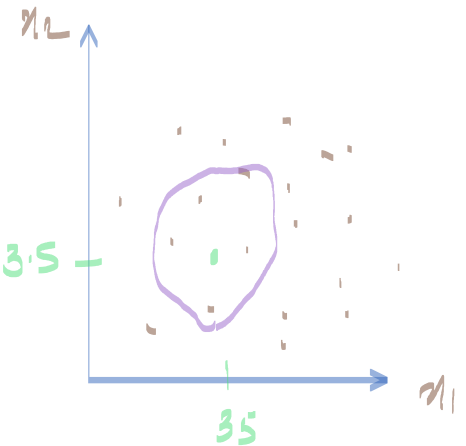
Non-linearly



Linearly seperable



Age x_1	Exp x_2	Sal y	
20	2	20k	2
30	3	30k	2.3
40	4	40k	3
50	5	45k	5
60	6	50k	5.4
70	7	60k	4
80	8	86k	4.5
35	3.5	?	



K = 5

y_p → mean of y value of n neighbors

y_p = 20 + 30 + 40 + 60 + 80

$$y_p = \frac{20 + 30 + 40 + 60 + 80}{5}$$

$$= 45k$$

* Feature Scaling

	x_1	x_2	x_3
	1	100	15000
	2	200	20000
	3	300	30000

$$\sqrt{(2-1)^2 + (200-100)^2 + (20000-15000)^2}$$

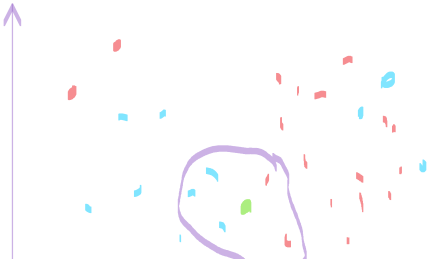
impact of higher scale values.

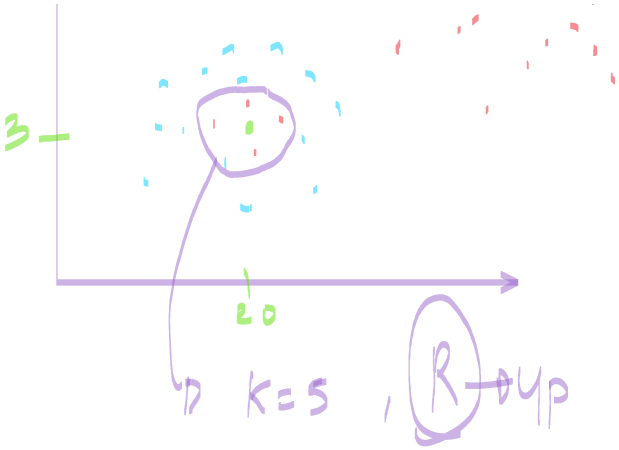
km	m	cm	ft
2	2000	180	6
3	3000	150	5
4	4000		

* Outliers

Sensitive to outliers

Not Sensitive





	Xi	Xnew
0	50	0.409
1	660	0.540
2	64	0.524
3	66	0.540
4	40	0.327

0.016

Xmin = 40
Xmax = 660

$\frac{50 - 40}{660 - 40} = 0.016$

From <http://localhost:8888/notebooks/12_29%20KNN/K%20-%20Nearest%20Neighbour.ipynb>

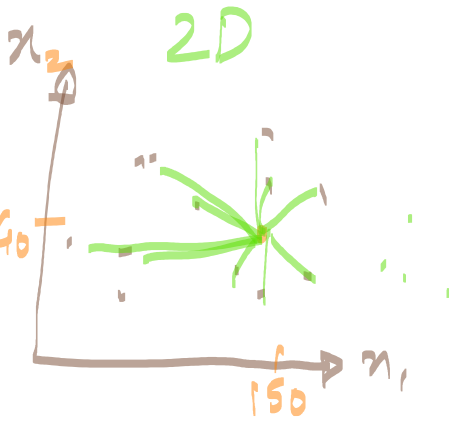
KNN IMPUTER

29 December 2022 20:40

148	50	50	35	0	33.6	0.627	50	1
85	66	66	29	0	26.6	0.351	31	0
183	64	64	0	0	23.3	0.672	52	1
150	66	66	23	94	28.1	0.167	21	0
150	40	? 63	35	168	43.1	2.288	33	1

Training

Test



$\frac{50+66}{2}$ $K = 2$

→ missing value = 63

Target

$P_1(20, 30)$, $P_2(30, 50)$

$ED = \sqrt{10^2 + 20^2}$

$=$

$* \text{nan\_ED} = \sqrt{\text{wt.} \times ED}$

$= \sqrt{wt \times (x_2 - x_1)^2 + (y_2 - y_1)^2}$

$= \sqrt{\text{Total}}$

$$= \sqrt{\frac{\text{Total of cordinates}}{\text{Present cordin}} \times ED}$$

30 December 2022 08:29

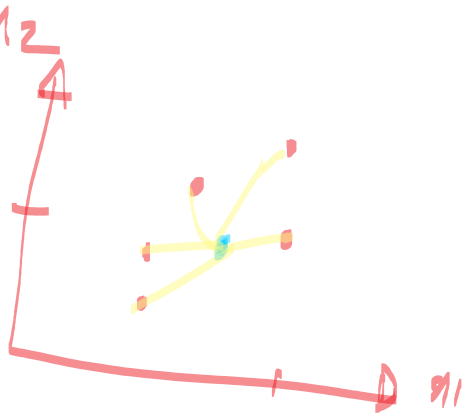
$k=3$

$P_1(148, 50)$, $P_2(150, 74)$

Glucose	BloodPressure	Target
148	50	50
85	66	66
183	64	64
150	66	66
150	40	40
150	74	?

Training

P_2, Testing



$$\frac{50+66+64}{3} = 60$$

y	n_1	n_2	n_3	
1	np.nan	10	Training
2	7	20	
3	8	np.nan	
np.nan	10	40	→ Test pt.

P_2

$P_2(10, 40)$, $P_1(np.nan, 10)$

$P_2(10, 40)$, $P_1(7, 20)$

$$ED = \sqrt{(10-7)^2 + (40-20)^2}$$

$$= \sqrt{\quad}$$

$P_1(8, np.nan)$ $P_2(10, 40)$

Nan - Euclidean → $\sqrt{wt. \times ED}$

$$= \sqrt{wt. \times (x_2-x_1)^2 + (y_2-y_1)^2}$$

$$= \sqrt{\omega t \cdot \times (x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$= \sqrt{\frac{\text{Total cordinates}}{\text{present cordinates}} \times \_\_\_\_\_}$$

Outliers

02 January 2023 08:06

Z-score

+3 >

-3 <