Shashwat Drolia | 18NA3AI37

MLFA: Assignment 1

[Sol #1]

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)
1	1	1	A	15	0	0
2	1	2	A	15	0	0
3	2	2	A	2	9	1
4	2	1	A	3	5	0
5	1	1	B	0	10	4
6	1	2	B	0	10	1
7	2	2	B	8	2	4
8	2	1	B	7	3	1
9	1	1	C	0	6	0
10	1	2	C	0	9	0
11	2	2	C	1	0	14
12	2	1	C	0	0	20
13	1	1	D	0	2	15
14	1	2	D	1	3	14
15	2	2	D	1	0	9
16	2	1	D	0	0	5

For the given set of features X1, X2, X3 class value Y is not unique and given the training set of 200 examples we have certain probabilities for each of the class values Y=1, 2 or 3.

Now, let us classify the dataset of features to a single Y value based on the max occurrence as follows:

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y
1	1	1	A	15	0	0	1
2	1	2	A	15	0	0	1
3	2	2	A	2	9	1	2
4	2	1	A	3	5	0	2
5	1	1	B	0	10	4	2
6	1	2	B	0	10	1	2
7	2	2	B	8	2	4	1
8	2	1	B	7	3	1	1
9	1	1	C	0	6	0	2
10	1	2	C	0	9	0	2
11	2	2	C	1	0	14	3
12	2	1	C	0	0	20	3
13	1	1	D	0	2	15	3
14	1	2	D	1	3	14	3
15	2	2	D	1	0	9	3
16	2	1	D	0	0	5	3

We check for 1st relevant split on X1, X2 and X3 that segregates Y into 2 groups with least entropy

(a) X1

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y	$\widehat{Y}$
1	1	1	A	15	0	0	1	1 or 2 or 3
2	1	1	B	0	10	4	2	
5	1	1	C	0	6	0	2	
6	1	1	D	0	2	15	3	
9	1	2	A	15	0	0	1	
10	1	2	B	0	10	1	2	
13	1	2	C	0	9	0	2	
14	1	2	D	1	3	14	3	
3	2	1	A	3	5	0	2	1 or 2 or 3
4	2	1	B	7	3	1	1	
7	2	1	C	0	0	20	3	
8	2	1	D	0	0	5	3	
11	2	2	A	2	9	1	2	
12	2	2	B	8	2	4	1	
15	2	2	C	1	0	14	3	
16	2	2	D	1	0	9	3	

(b) X2

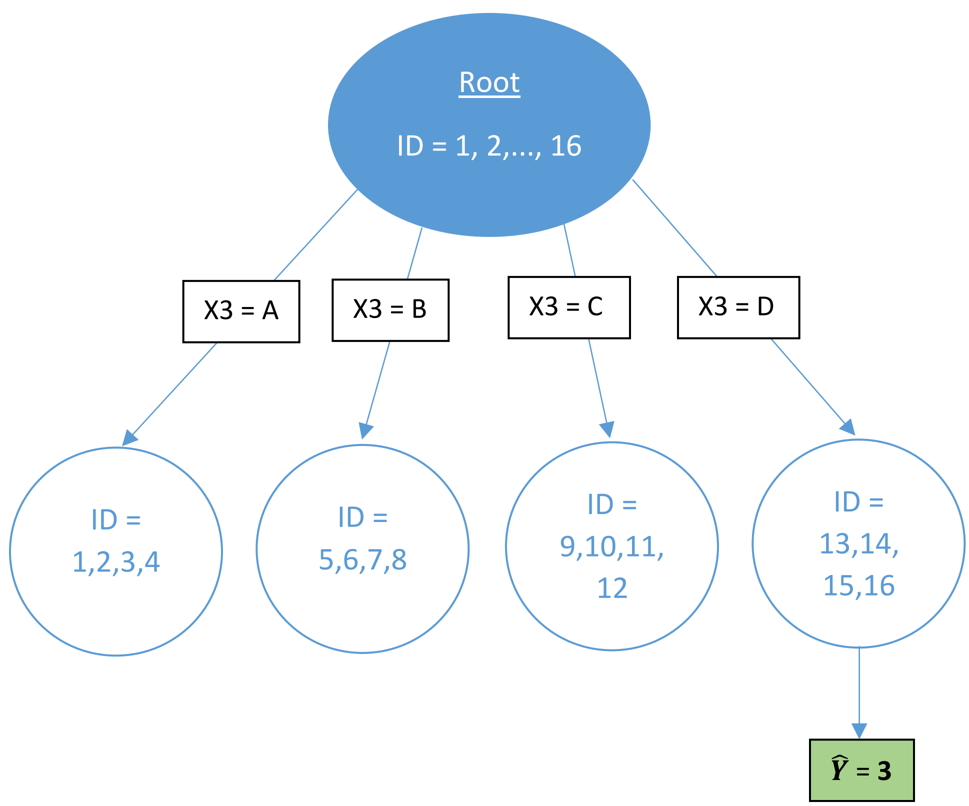
ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y	$\widehat{Y}$
1	1	1	A	15	0	0	1	1 or 2 or 3
5	2	1	A	3	5	0	2	
9	1	1	B	0	10	4	2	
13	2	1	B	7	3	1	1	
4	1	1	C	0	6	0	2	
8	2	1	C	0	0	20	3	
12	1	1	D	0	2	15	3	
16	2	1	D	0	0	5	3	
2	1	2	A	15	0	0	1	1 or 2 or 3
6	2	2	A	2	9	1	2	
10	1	2	B	0	10	1	2	
14	2	2	B	8	2	4	1	
3	1	2	C	0	9	0	2	
7	2	2	C	1	0	14	3	
11	1	2	D	1	3	14	3	
15	2	2	D	1	0	9	3	

(c) X3

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y	$\hat{Y}$
1	1	1	A	15	0	0	1	1 or 2
2	1	2	A	15	0	0	1	
3	2	2	A	2	9	1	2	
4	2	1	A	3	5	0	2	
5	1	1	B	0	10	4	2	1 or 2
6	1	2	B	0	10	1	2	
7	2	2	B	8	2	4	1	
8	2	1	B	7	3	1	1	
9	1	1	C	0	6	0	2	2 or 3
10	1	2	C	0	9	0	2	
11	2	2	C	1	0	14	3	
12	2	1	C	0	0	20	3	
13	1	1	D	0	2	15	3	3
14	1	2	D	1	3	14	3	
15	2	2	D	1	0	9	3	
16	2	1	D	0	0	5	3	

So just be observation, 1st split at X3 gives the max information gain due to the least entropy

$\Longrightarrow$

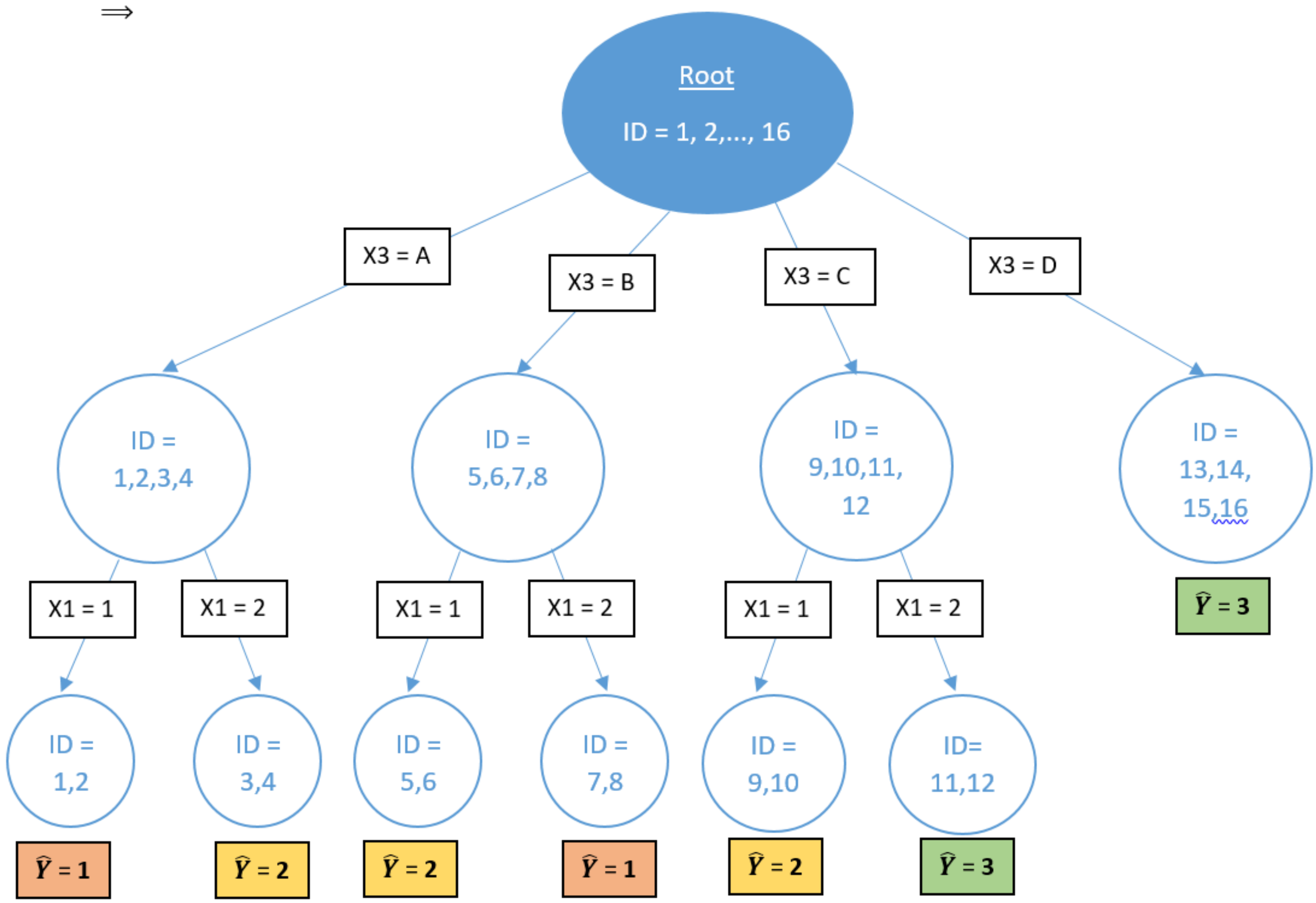


We check each of the 3 nodes for our 2^{nd} split

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y	$\widehat{Y}$
1	1	1	A	15	0	0	1	1
2	1	2	A	15	0	0	1	
3	2	2	A	2	9	1	2	2
4	2	1	A	3	5	0	2	
5	1	1	B	0	10	4	2	2
6	1	2	B	0	10	1	2	
7	2	2	B	8	2	4	1	1
8	2	1	B	7	3	1	1	
9	1	1	C	0	6	0	2	2
10	1	2	C	0	9	0	2	
11	2	2	C	1	0	14	3	3
12	2	1	C	0	0	20	3	
13	1	1	D	0	2	15	3	3
14	1	2	D	1	3	14	3	
15	2	2	D	1	0	9	3	
16	2	1	D	0	0	5	3	

2^{nd} Split is made w.r.t. X1 as it provides the maximum information gain as compared to split w.r.t X2

$\Longrightarrow$



Since we had classified dataset IDs to a particular Y based on max occurrence in training example, only those will be correctly classified and all other occurrences will be incorrectly classified.

The #correct classifications include:

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	Y
1	1	1	A	15	0	0	1
2	1	2	A	15	0	0	1
3	2	2	A	2	9	1	2
4	2	1	A	3	5	0	2
5	1	1	B	0	10	4	2
6	1	2	B	0	10	1	2
7	2	2	B	8	2	4	1
8	2	1	B	7	3	1	1
9	1	1	C	0	6	0	2
10	1	2	C	0	9	0	2
11	2	2	C	1	0	14	3
12	2	1	C	0	0	20	3
13	1	1	D	0	2	15	3
14	1	2	D	1	3	14	3
15	2	2	D	1	0	9	3
16	2	1	D	0	0	5	3

∴ #correct classifications = 15 + 15 +9 + 5 + 10 + 10 + 8 + 7 + 6 + 9 + 14 + 20 + 15 + 14 + 9 + 5

= 171

⟹ Accuracy on training set = $\frac{171}{200} = \mathbf{85.5}\%$

[Sol #2]

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)
1	1	1	A	15	0	0
2	1	2	A	15	0	0
3	2	2	A	2	9	1
4	2	1	A	**X**	**X**	**X**
5	1	1	B	0	10	4
6	1	2	B	0	10	1
7	2	2	B	8	2	4
8	2	1	B	**X**	**X**	**X**
9	1	1	C	**X**	**X**	**X**
10	1	2	C	0	9	0
11	2	2	C	1	0	14
12	2	1	C	0	0	20
13	1	1	D	0	2	15
14	1	2	D	1	3	14
15	2	2	D	1	0	9
16	2	1	D	**X**	**X**	**X**

$Posterior\ distribution$ = $P(Y \mid X) = P(Y = 1|X) + P(Y = 2|X) + P(Y = 3|X)$

$$= P(X|Y) \cdot P(Y)$$

Where, $P(X|Y) \equiv Classs\ conditional$

$P(Y) \equiv Prior\ distribution$

ID	X1	X2	X3	#(Y=1)	#(Y=2)	#(Y=3)	
1	1	1	A	15	0	0	15
2	1	2	A	15	0	0	15
3	2	2	A	2	9	1	12
5	1	1	B	0	10	4	14
6	1	2	B	0	10	1	11
7	2	2	B	8	2	4	14
10	1	2	C	0	9	0	9
11	2	2	C	1	0	14	15
12	2	1	C	0	0	20	20
13	1	1	D	0	2	15	17
14	1	2	D	1	3	14	18
15	2	2	D	1	0	9	10
				43	45	82	170

Prior distribution

$P(Y = 1) = \frac{43}{170}$ $P(Y = 2) = \frac{45}{170}$ $P(Y = 3) = \frac{82}{170}$

Class Conditional

(i) $P(X1 \mid Y)$

	$X1 = 1$	$X1 = 2$	
$Y = 1$	$\frac{31}{43}$	$\frac{12}{43}$	1
$Y = 2$	$\frac{34}{45}$	$\frac{11}{45}$	1
$Y = 3$	$\frac{34}{82}$	$\frac{48}{82}$	1

(ii) $P(X2 \mid Y)$

	$X2 = 1$	$X2 = 2$	
$Y = 1$	$\frac{15}{43}$	$\frac{28}{43}$	1
$Y = 2$	$\frac{12}{45}$	$\frac{33}{45}$	1
$Y = 3$	$\frac{39}{82}$	$\frac{43}{82}$	1

(iii) $P(X3 \mid Y)$

	$X3 = A$	$X3 = B$	$X3 = C$	$X3 = D$	
$Y = 1$	$\frac{32}{43}$	$\frac{8}{43}$	$\frac{1}{43}$	$\frac{2}{43}$	1
$Y = 2$	$\frac{9}{45}$	$\frac{22}{45}$	$\frac{9}{45}$	$\frac{5}{45}$	1
$Y = 3$	$\frac{1}{82}$	$\frac{9}{82}$	$\frac{34}{82}$	$\frac{38}{82}$	1

(a) $\hat{Y}$ for $(X1 = 2, X2 = 1, X3 = A)$

$$P(Y = 1 \mid X1 = 2, X2 = 1, X3 = A) = K \cdot P(X1 = 2, X2 = 1, X3 = A \mid Y = 1) \cdot P(Y = 1)$$

By Naives Bayes assumption:

$$P(X1 = 2, X2 = 1, X3 = A \mid Y = 1) = P(X1 = 2 \mid Y = 1) \cdot P(X2 = 1 \mid Y = 1) \cdot P(X3 = A \mid Y = 1)$$

$$\Rightarrow P(\boldsymbol{Y}=\mathbf{1} \mid X1=2, X2=1, X3=A) = K \cdot \frac{12}{43} \cdot \frac{15}{43} \cdot \frac{32}{43} \cdot \frac{43}{170} = \frac{\mathbf{576}}{\mathbf{31433}}\boldsymbol{K}$$

Similarly,

$$P(\boldsymbol{Y}=\mathbf{2} \mid X1=2, X2=1, X3=A) = K \cdot P(X1=2 \mid Y=2) \cdot P(X2=1 \mid Y=2) \cdot P(X3=A \mid Y=2) P(Y=2)$$

$$\Rightarrow P(\boldsymbol{Y}=\mathbf{2} \mid X1=2, X2=1, X3=A) = K \cdot \frac{11}{45} \cdot \frac{12}{45} \cdot \frac{9}{45} \cdot \frac{45}{170} = \frac{\mathbf{22}}{\mathbf{6375}}\boldsymbol{K}$$

And,

$$P(\boldsymbol{Y}=\mathbf{3} \mid X1=2, X2=1, X3=A) = K \cdot \frac{48}{82} \cdot \frac{39}{82} \cdot \frac{1}{82} \cdot \frac{82}{170} = \frac{\mathbf{234}}{\mathbf{142885}}\boldsymbol{K}$$

$$P(\boldsymbol{Y}=\mathbf{1} \mid X) + P(\boldsymbol{Y}=\mathbf{2} \mid X) + P(\boldsymbol{Y}=\mathbf{3} \mid X) = 1$$

$$\Rightarrow \frac{576}{31433}K + \frac{22}{6375}K + \frac{234}{142885}K = 1$$

$$\Rightarrow K = 42.7107$$

∴ The respective confidence values are as follows:

$$\Rightarrow \boxed{P(\boldsymbol{Y}=\mathbf{1} \mid X1=2, X2=1, X3=A) = \frac{576}{31433}K = \mathbf{0.7827} = \mathbf{78.27\%}}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y}=\mathbf{2} \mid X1=2, X2=1, X3=A) = \frac{22}{6375}K = \mathbf{0.1474} = \mathbf{14.74\%}}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y}=\mathbf{3} \mid X1=2, X2=1, X3=A) = \frac{234}{142885}K = \mathbf{0.0699} = \mathbf{6.99\%}}$$

∴

ID	**X1**	**X2**	**X3**	$\hat{Y}$
4	**2**	**1**	**A**	**1**

(b) $\hat{Y}$ for $(X1 = 2, X2 = 1, X3 = B)$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 2, X2 = 1, X3 = B) = K \cdot \frac{12}{43} \cdot \frac{15}{43} \cdot \frac{8}{43} \cdot \frac{43}{170} = \frac{\mathbf{144}}{\mathbf{31433}} \boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 2, X2 = 1, X3 = B) = K \cdot \frac{11}{45} \cdot \frac{12}{45} \cdot \frac{22}{45} \cdot \frac{45}{170} = \frac{\mathbf{484}}{\mathbf{57375}} \boldsymbol{K}$$

And,

$$P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 2, X2 = 1, X3 = B) = K \cdot \frac{48}{82} \cdot \frac{39}{82} \cdot \frac{9}{82} \cdot \frac{82}{170} = \frac{\mathbf{2106}}{\mathbf{142885}} \boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X) + P(\boldsymbol{Y} = \mathbf{2} \mid X) + P(\boldsymbol{Y} = \mathbf{3} \mid X) = 1$$

$$\Rightarrow \frac{144}{31433} K + \frac{484}{57375} K + \frac{2106}{142885} K = 1$$

$$\Rightarrow K = 36.0282$$

∴ The respective confidence values are as follows:

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 2, X2 = 1, X3 = B) = \frac{144}{31433} K = \mathbf{0.1651} = \mathbf{16.51}\%}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 2, X2 = 1, X3 = B) = \frac{484}{57375} K = \mathbf{0.3039} = \mathbf{30.39}\%}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 2, X2 = 1, X3 = B) = \frac{2106}{142885} K = \mathbf{0.5310} = \mathbf{53.10}\%}$$

∴

ID	X1	X2	X3	$\hat{Y}$
8	2	1	B	3

(c) $\hat{Y}$ for $(X1 = 1, X2 = 1, X3 = C)$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 1, X2 = 1, X3 = C) = K \cdot \frac{31}{43} \cdot \frac{15}{43} \cdot \frac{1}{43} \cdot \frac{43}{170} = \frac{\mathbf{93}}{\mathbf{62866}}\boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 1, X2 = 1, X3 = C) = K \cdot \frac{34}{45} \cdot \frac{12}{45} \cdot \frac{9}{45} \cdot \frac{45}{170} = \frac{\mathbf{4}}{\mathbf{375}}\boldsymbol{K}$$

And,

$$P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 1, X2 = 1, X3 = C) = K \cdot \frac{34}{82} \cdot \frac{39}{82} \cdot \frac{34}{82} \cdot \frac{82}{170} = \frac{\mathbf{663}}{\mathbf{16810}}\boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X) + P(\boldsymbol{Y} = \mathbf{2} \mid X) + P(\boldsymbol{Y} = \mathbf{3} \mid X) = 1$$

$$\Rightarrow \frac{93}{62866}K + \frac{4}{375}K + \frac{663}{16810}K = 1$$

$$\Rightarrow K = 19.3848$$

∴ The respective confidence values are as follows:

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 2, X2 = 1, X3 = B) = \frac{144}{31433}K = \mathbf{0.0287} = \mathbf{2.87}\%}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 2, X2 = 1, X3 = B) = \frac{484}{57375}K = \mathbf{0.2068} = \mathbf{20.68}\%}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 2, X2 = 1, X3 = B) = \frac{2106}{142885}K = \mathbf{0.7646} = \mathbf{76.46}\%}$$

∴

ID	X1	X2	X3	$\hat{Y}$
9	1	1	C	3

(d) $\hat{Y}$ for $(X1 = 2, X2 = 1, X3 = D)$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 2, X2 = 1, X3 = D) = K \cdot \frac{12}{43} \cdot \frac{15}{43} \cdot \frac{2}{43} \cdot \frac{43}{170} = \frac{\mathbf{36}}{\mathbf{31433}} \boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 2, X2 = 1, X3 = D) = K \cdot \frac{11}{45} \cdot \frac{12}{45} \cdot \frac{5}{45} \cdot \frac{45}{170} = \frac{\mathbf{22}}{\mathbf{11475}} \boldsymbol{K}$$

And,

$$P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 2, X2 = 1, X3 = D) = K \cdot \frac{48}{82} \cdot \frac{39}{82} \cdot \frac{38}{82} \cdot \frac{82}{170} = \frac{\mathbf{8892}}{\mathbf{142885}} \boldsymbol{K}$$

$$P(\boldsymbol{Y} = \mathbf{1} \mid X) + P(\boldsymbol{Y} = \mathbf{2} \mid X) + P(\boldsymbol{Y} = \mathbf{3} \mid X) = 1$$

$$\Rightarrow \frac{36}{31433} K + \frac{22}{11475} K + \frac{8892}{142885} K = 1$$

$$\Rightarrow K = 15.3153$$

∴ The respective confidence values are as follows:

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{1} \mid X1 = 2, X2 = 1, X3 = B) = \frac{144}{31433} K = \mathbf{0.0175} = \mathbf{1.75}\%}$$

$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{2} \mid X1 = 2, X2 = 1, X3 = B) = \frac{484}{57375} K = \mathbf{0.0294} = \mathbf{2.94}\%}$$

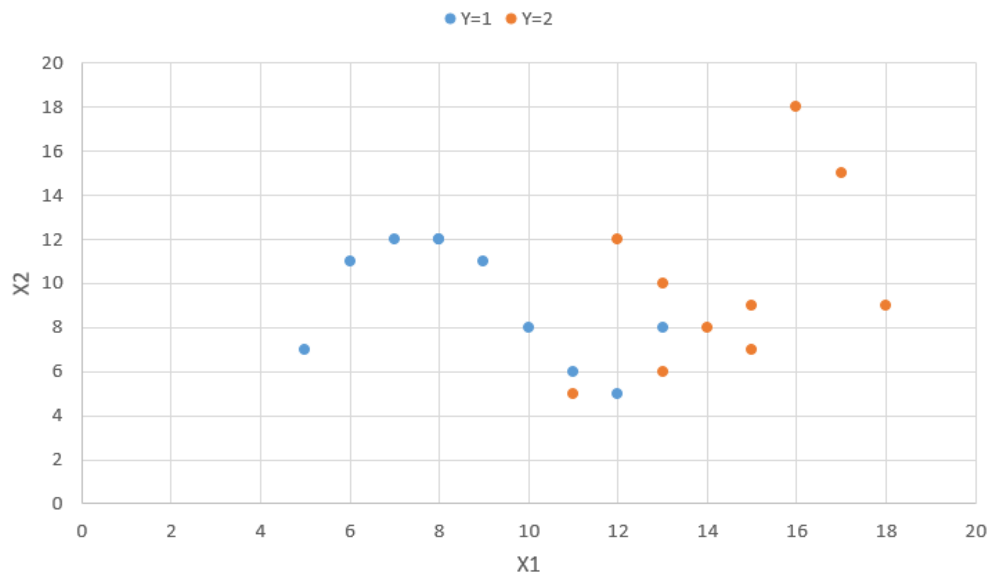
$$\Rightarrow \boxed{P(\boldsymbol{Y} = \mathbf{3} \mid X1 = 2, X2 = 1, X3 = B) = \frac{2106}{142885} K = \mathbf{0.9531} = \mathbf{95.31}\%}$$

∴

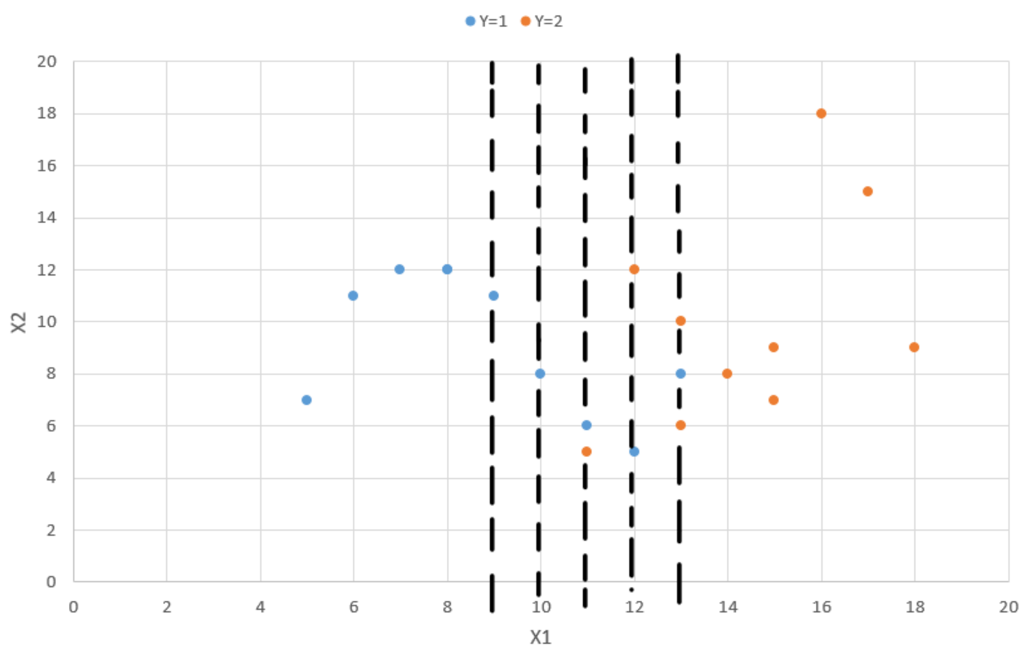
ID	X1	X2	X3	$\hat{Y}$
16	2	1	D	3

[Sol #3]

Plotting the given points on a 2D plane:



From the 2D plot it is quite evident that we should make the split using the feature $X1$ and any threshold for $X1 \in [9, 13]$ can possibly give the best decision stump.



So we check for max information gain for each of the 5 cases of X1

Entropy is defined as

$$H = -\sum p_i \log_2 p_i$$

Information gain is defined as

$$IG = H_{final} - H_{initial}$$

$$H_{initial} = -\frac{10}{20}\log_2\frac{10}{20} - \frac{10}{20}\log_2\frac{10}{20} = 1$$

For $X = 9$,

$$H_{(X1\le 9)} = 0$$

$$H_{(X1>9)} = -\frac{4}{14}\log_2\frac{4}{14} - \frac{10}{14}\log_2\frac{10}{14} = 0.8631$$

$$H_{final} = \frac{6}{20}(0) + \frac{14}{20}(0.8631) = 0.6042$$

	X1<=9	X1>9
#(Y=1)	6	4
#(Y=2)	0	10
#Total	6	14
Entropy	0	0.8631
Final entropy	0.6042	
IG	**0.3958**	

Similarly, we evaluate the information gain (IG) for other cases of split as follows:

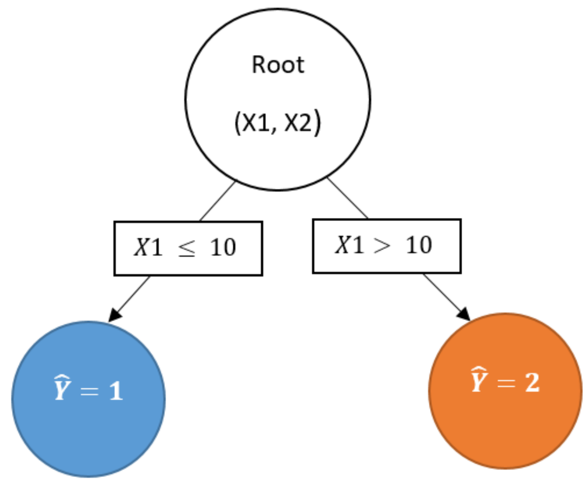
	X1<=10	X1>10
#(Y=1)	7	3
#(Y=2)	0	10
#Total	7	13
Entropy	0	0.7793
Final entropy	0.5066	
IG	**0.4934**	

	X1<=11	X1>11
#(Y=1)	8	2
#(Y=2)	1	9
#Total	9	11
Entropy	0.5033	0.6840
Final entropy	0.6027	
IG	**0.3973**	

	X1<=12	X1>12
#(Y=1)	9	1
#(Y=2)	2	8
#Total	11	9
Entropy	0.6840	0.5033
Final entropy	0.6027	
IG	**0.3973**	

	X1<=13	X1>13
#(Y=1)	10	0
#(Y=2)	4	6
#Total	14	6
Entropy	0.8631	0
Final entropy	0.6042	
IG	**0.3958**	

Hence, split across $\boldsymbol{X1 = 10}$ gives the maximum information gain and thus the best decision stump


Root
(X1, X2)
$X1 \leq 10$
$X1 > 10$
$\hat{Y} = 1$
$\hat{Y} = 2$

[Sol #4]

$Posterior\ distribution = P(Y \mid X) = P(Y = 1|X) + P(Y = 2|X)$

$$= P(X|Y) \cdot P(Y)$$

Where, $P(X|Y) \equiv Classs\ conditional\ \equiv N(\mu, \sigma)$

$P(Y) \equiv Prior\ distribution$

Prior distribution

$P(Y = 1) = P(Y = 2) = \frac{10}{20} = 0.5$

Class Conditional

ID	X1	X2	Y
1	5	7	1
2	7	12	1
3	12	5	1
4	10	8	1
5	6	11	1
6	13	8	1
7	8	12	1
8	9	11	1
9	11	6	1
10	8	12	1
$Average\ (\mu)$	8.900	9.200	
$Variance(\sigma)$	6.767	7.289	

ID	X1	X2	Y
11	13	6	2
12	14	8	2
13	17	15	2
14	15	9	2
15	13	10	2
16	11	5	2
17	16	18	2
18	15	7	2
19	12	12	2
20	18	9	2
$Average\ (\mu)$	14.400	9.900	
$Variance(\sigma)$	4.933	16.544	

Calculating $P(X|Y)$ using normal distribution

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

So,

$$P(X1 = 5 \mid Y = 1) = \frac{1}{6.767\sqrt{2\pi}} e^{-\frac{(5-8.9)^2}{2*6.767^2}} = 0.0499$$

$$P(X2 = 7 \mid Y = 1) = \frac{1}{7.289\sqrt{2\pi}} e^{-\frac{(7-9.2)^2}{2*7.289^2}} = 0.0523$$

$$P(X1 = 5 \mid Y = 2) = \frac{1}{4.933\sqrt{2\pi}} e^{-\frac{(5-14.4)^2}{2*4.933^2}} = 0.0132$$

$$P(X2 = 7 \mid Y = 2) = \frac{1}{16.544\sqrt{2\pi}} e^{-\frac{(7-9.9)^2}{2*16.544^2}} = 0.0237$$

Similarly, we calculate the other normal distribution probabilities as follows:

ID	X1	X2	Y	P(X1 | Y=1)	P(X2 | Y=1)	P(X1 | Y=2)	P(X2 | Y=2)
1	5	7	1	0.0499	0.0523	0.0132	0.0237
2	7	12	1	0.0567	0.0508	0.0263	0.0239
3	12	5	1	0.0531	0.0464	0.0718	0.0231
4	10	8	1	0.0582	0.0540	0.0543	0.0240
5	6	11	1	0.0538	0.0531	0.0190	0.0241
6	13	8	1	0.0491	0.0540	0.0777	0.0240
7	8	12	1	0.0584	0.0508	0.0349	0.0239
8	9	11	1	0.0590	0.0531	0.0444	0.0241
9	11	6	1	0.0562	0.0497	0.0638	0.0235
10	8	12	1	0.0584	0.0508	0.0349	0.0239

ID	X1	X2	Y	P(X1 | Y=2)	P(X2 | Y=2)	P(X1 | Y=1)	P(X2 | Y=1)
11	13	6	2	0.0777	0.0235	0.0491	0.0497
12	14	8	2	0.0806	0.0240	0.0444	0.0540
13	17	15	2	0.0704	0.0230	0.0288	0.0399
14	15	9	2	0.0803	0.0241	0.0393	0.0547
15	13	10	2	0.0777	0.0241	0.0491	0.0544
16	11	5	2	0.0638	0.0231	0.0562	0.0464
17	16	18	2	0.0767	0.0214	0.0340	0.0264
18	15	7	2	0.0803	0.0237	0.0393	0.0523
19	12	12	2	0.0718	0.0239	0.0531	0.0508
20	18	9	2	0.0620	0.0241	0.0239	0.0547

Now, posterior distribution is calculated as follows:

$P(Y \mid X1, X2) = K * P(X1, X2 \mid Y) * P(Y)$, where K is the proportionality constant

Using the independence assumption under Naïve Bayes

$$P(Y \mid X) = K * P(X1 \mid Y) * P(X2 \mid Y) * 0.5$$

ID	X1	X2	Y	P(Y=1|X1, X2)	P(Y=2|X1, X2)
1	5	7	1	0.00131*K1	0.00016*K1
2	7	12	1	0.00144*K2	0.00031*K2
3	12	5	1	0.00123*K3	0.00083*K3
4	10	8	1	0.00157*K4	0.00065*K4
5	6	11	1	0.00143*K5	0.00023*K5
6	13	8	1	0.00132*K6	0.00093*K6
7	8	12	1	0.00149*K7	0.00042*K7
8	9	11	1	0.00156*K8	0.00053*K8
9	11	6	1	0.0014*K9	0.00075*K9
10	8	12	1	0.00149*K10	0.00042*K10

ID	X1	X2	Y	P(Y=2|X1, X2)	P(Y=1|X1, X2)
11	13	6	2	0.00091*K11	0.00122*K11
12	14	8	2	0.00097*K12	0.0012*K12
13	17	15	2	0.00081*K13	0.00057*K13
14	15	9	2	0.00097*K14	0.00107*K14
15	13	10	2	0.00094*K15	0.00133*K15
16	11	5	2	0.00074*K16	0.0013*K16
17	16	18	2	0.00082*K17	0.00045*K17
18	15	7	2	0.00095*K18	0.00103*K18
19	12	12	2	0.00086*K19	0.00135*K19
20	18	9	2	0.00075*K20	0.00065*K20

Now, $$P(Y = 1 \mid X1 = 5, X2 = 7) + P(Y = 2 \mid X1 = 5, X2 = 7) = 1$$

$$\Rightarrow 0.00131 * K1 + 0.00016 * K1 = 1$$

$$\Rightarrow K1 = 684.0014$$

$$\therefore P(Y = 1 \mid X1 = 5, X2 = 7) = 0.00131 * 684.0014 = 0.8931$$, and

$$P(Y = 2 \mid X1 = 5, X2 = 7) = 0.00016 * 684.0014 = 0.1069$$

So similarly, we evaluate $Ki's$ for $i = 1,2, \dots 20$ and correspondingly the relative posterior probabilities

ID	X1	X2	Y	Ki	P(Y=1|X1, X2)	P(Y=2|X1, X2)	Confidence
1	5	7	1	684.0014	0.8931	0.1069	89.31%
2	7	12	1	569.8835	0.8211	0.1789	82.11%
3	12	5	1	485.5546	0.5975	0.4025	59.75%
4	10	8	1	450.134	0.7071	0.2929	70.71%
5	6	11	1	603.8805	0.8621	0.1379	86.21%
6	13	8	1	443.4314	0.5875	0.4125	58.75%
7	8	12	1	525.6524	0.7808	0.2192	78.08%
8	9	11	1	476.3681	0.7454	0.2546	74.54%
9	11	6	1	466.3892	0.6512	0.3488	65.12%
10	8	12	1	525.6524	0.7808	0.2192	78.08%

ID	X1	X2	Y	Ki	P(Y=2|X1, X2)	P(Y=1|X1, X2)	Confidence
11	13	6	2	469.4094	0.4276	0.5724	57.24%
12	14	8	2	462.2007	0.4462	0.5538	55.38%
13	17	15	2	722.8345	0.5849	0.4151	58.49%
14	15	9	2	490.0369	0.4736	0.5264	52.64%
15	13	10	2	440.2783	0.4123	0.5877	58.77%
16	11	5	2	490.6117	0.3610	0.6390	63.90%
17	16	18	2	787.7283	0.6464	0.3536	64.64%
18	15	7	2	505.0788	0.4814	0.5186	**51.86%**
19	12	12	2	452.7742	0.3890	0.6110	61.10%
20	18	9	2	714.8463	0.5333	0.4667	53.33%

For the feature values **(X1 = 15, X2 = 7)** our NBC is the **least confident** with confidence of **51.86%**

[Sol #5]

Part (i)

Given N inputs and outputs of the form:

$$(x_i, y_i, w_i)$$

We define the line of best fit as

$$\hat{y} = a'x + b$$

Now, given the dependency on weights w_i we define our loss function as

$$L = \sum_{i=1}^{N} w_i (y_i - \hat{y_i})^2$$

$$\Rightarrow L = \sum_{i=1}^{N} w_i (y_i - a'x_i - b)^2$$

To minimize L first derivative of L w.r.t. a' and b should be equal to 0

$$\Rightarrow \frac{\delta L}{\delta a'} = 0 \; and \; \frac{\delta L}{\delta b} = 0$$

∴ (a)

$$\frac{\delta}{\delta a'} \sum_{i=1}^{N} w_i (y_i - a'x_i - b)^2 = 0$$

$$\Rightarrow \sum_{i=1}^{N} -2x_i w_i \, (y_i - a'x_i - b) = 0$$

Eq. 1 -

$$\Rightarrow \sum_{i=1}^{N} (w_i \, y_i x_i - a' w_i x_i^2 - b w_i x_i) = 0$$

(b)

$$\frac{\delta}{\delta b} \sum_{i=1}^{N} w_i (y_i - a'x_i - b)^2 = 0$$

$$\Rightarrow \sum_{i=1}^{N} -2w_i \, (y_i - a'x_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^{N} w_i\, y_i - a' \sum_{i=1}^{N} w_i\, x_i - b \sum_{i=1}^{N} w_i = 0$$

$$\Rightarrow b = \frac{\sum_{i=1}^{N} w_i\, y_i - a' \sum_{i=1}^{N} w_i\, x_i}{\sum_{i=1}^{N} w_i}$$

We can observe weighted means as

$$\overline{y_w} = \frac{\sum_{i=1}^{N} w_i\, y_i}{\sum_{i=1}^{N} w_i}$$

$$\overline{x_w} = \frac{\sum_{i=1}^{N} w_i\, x_i}{\sum_{i=1}^{N} w_i}$$

∴

Eq. 2 -

$$\boxed{b = \overline{y} - a'\overline{x}}$$

Now putting Eq. 2 in Eq. 1

$$\sum_{i=1}^{N} (w_i\, y_i x_i - a' w_i x_i^2 - w_i x_i (\overline{y_w} - a'\overline{x_w})) = 0$$

Rearranging the terms,

$$\Rightarrow \sum_{i=1}^{N} (w_i\, y_i x_i - w_i x_i \overline{y_w}) - a' \sum_{i=1}^{N} w_i (x_i^2 - x_i \overline{x_w}) \; = 0$$

Eq. 3 -

$$\Rightarrow \boxed{a' = \frac{\sum_{i=1}^{N} w_i (\, y_i x_i - x_i \overline{y_w})}{\sum_{i=1}^{N} w_i (x_i^2 - x_i \overline{x_w})}}$$

Hence, we have derived our linear regression model

$$\hat{y} = a' x + b$$

$$\Rightarrow \boxed{\widehat{y_i} = \overline{y} + \frac{\sum_{i=1}^{N} w_i (\, y_i x_i - x_i \overline{y_w})}{\sum_{i=1}^{N} w_i (x_i^2 - x_i \overline{x_w})} (x_i - \overline{x_w})}$$

Part (ii)

Given N inputs and outputs of the form:

$$(x_i, y_i)$$

We define the line of best fit as

$$\hat{\boldsymbol{y}} = \boldsymbol{a}\boldsymbol{x} + \boldsymbol{b}$$

We need to fit a linear model such that a_i is close to the given vector v in terms of Euclidean distance

Euclidean distance D of vector between $\boldsymbol{a}$ and $\boldsymbol{v}$ is the L2 norm:

$$D(a, v) = \left|\left|\boldsymbol{a} - \boldsymbol{v}\right|\right|_{\boldsymbol{2}}$$

$$D = \sqrt{\sum_{i=1}^{N} (a_i - v_i)^2}$$

For simplicity of calculation we square both sides:

$$D^2 = \sum_{i=1}^{N} (a_i - v_i)^2$$

$$D^2 = \left|\left|\boldsymbol{a} - \boldsymbol{v}\right|\right|_{\boldsymbol{2}}^{\boldsymbol{2}}$$

The L2-norm of the vector $\boldsymbol{a} - \boldsymbol{v}$ can be represented as

$$\left|\left|\boldsymbol{a} - \boldsymbol{v}\right|\right|_{\boldsymbol{2}}^{\boldsymbol{2}} = (\boldsymbol{a} - \boldsymbol{v})^T(\boldsymbol{a} - \boldsymbol{v})$$

This seems similar to Ridge regression where instead of limiting the distance from origin we are doing it from a given vector v

Our original loss function is defined as

$$= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\Rightarrow L = \sum_{i=1}^{N} (y_i - \boldsymbol{a}^{\boldsymbol{T}} x_i - b)^2$$

The regularization function is defined as

$$f(a) = \frac{1}{2}\left|\left|\boldsymbol{a} - \boldsymbol{v}\right|\right|_{\boldsymbol{2}}^{\boldsymbol{2}} = (\boldsymbol{a} - \boldsymbol{v})^T(\boldsymbol{a} - \boldsymbol{v})$$

Our objective function becomes

$$\phi(a,b) = L(a,b) + \lambda f(a)$$

To minimize L first derivative of L w.r.t. a and b should be equal to 0

$$\Rightarrow \frac{\delta\phi}{\delta a} = 0 \; and \; \frac{\delta\phi}{\delta b} = 0$$

∴ (a)

$$\frac{\delta}{\delta a}\left(\sum_{i=1}^{N}(y_i - \boldsymbol{a}^{\boldsymbol{T}}x_i - b)^2 + \lambda(\boldsymbol{a}-\boldsymbol{v})^T(\boldsymbol{a}-\boldsymbol{v})\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} x_i\,(y_i - \boldsymbol{a}^{\boldsymbol{T}}x_i - b) + \lambda(\boldsymbol{a}-\boldsymbol{v}) = 0$$

Eq. 1 -

$$\Rightarrow \sum_{i=1}^{N} x_i\,y_i - \boldsymbol{a}^{\boldsymbol{T}}\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i + \lambda(\boldsymbol{a}-\boldsymbol{v}) = 0$$

(b)

$$\frac{\delta}{\delta b}\left(\sum_{i=1}^{N}(y_i - \boldsymbol{a}^{\boldsymbol{T}}x_i - b)^2 + \lambda(\boldsymbol{a}-\boldsymbol{v})^T(\boldsymbol{a}-\boldsymbol{v})\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N}(y_i - \boldsymbol{a}^{\boldsymbol{T}}x_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i - \boldsymbol{a}^{\boldsymbol{T}}\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} b = 0$$

$$\Rightarrow Nb = \sum_{i=1}^{N} y_i - \boldsymbol{a}^{\boldsymbol{T}}\sum_{i=1}^{N} x_i$$

$$\Rightarrow b = \frac{\sum_{i=1}^{N} y_i}{N} - \boldsymbol{a}^{\boldsymbol{T}}\frac{\sum_{i=1}^{N} x_i}{N}$$

Eq. 2 -

$$\boxed{b = \bar{\mathrm{y}} - \boldsymbol{a}^{\boldsymbol{T}}\bar{\mathrm{x}}}$$

Putting Eq. 2 in Eq.1

$$\Rightarrow \sum_{i=1}^{N} x_i\, y_i - \boldsymbol{a^T} \sum_{i=1}^{N} x_i^2 - (\bar{\mathrm{y}} - \boldsymbol{a^T}\bar{\mathrm{x}}) \sum_{i=1}^{N} x_i + \lambda(\boldsymbol{a} - \boldsymbol{v}) = 0$$

$$\Rightarrow \left(\bar{\mathrm{x}} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2 + \lambda\right) \boldsymbol{a} = \bar{\mathrm{y}} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i\, y_i + \lambda(\boldsymbol{v})$$

$$\Rightarrow \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{x}}_{\boldsymbol{i}} (\widetilde{\boldsymbol{x}}_{\boldsymbol{\imath}})^{\boldsymbol{T}} + \lambda \boldsymbol{I}\right) \boldsymbol{a} = \sum_{i=1}^{N} \widetilde{\boldsymbol{x}}_{\boldsymbol{i}} \widetilde{\boldsymbol{y}}_{\boldsymbol{\imath}} + \lambda(\boldsymbol{v})$$

Where, $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} = x_i - \bar{x}$

$\widetilde{\boldsymbol{y}}_{\boldsymbol{i}} = y_i - \bar{y}$

and I is the $D * D$ identity matrix

$$\boxed{\boldsymbol{a} = \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{x}}_{\boldsymbol{i}} (\widetilde{\boldsymbol{x}}_{\boldsymbol{\imath}})^{\boldsymbol{T}} + \lambda \boldsymbol{I}\right)^{\boldsymbol{-1}} \left(\sum_{i=1}^{N} \widetilde{\boldsymbol{x}}_{i} \widetilde{\boldsymbol{y}}_{\boldsymbol{\imath}} + \lambda(\boldsymbol{v})\right)}$$