IS - 733

Classwork Date : 03/03/2025

1.a) We know that

H → Hypothesis that class is spam or Ham

For spam :	For Ham
E_1 - containg Link = Yes	E_4 - containg Link = Yes
E_2 - Contains money = No	E_5 = contains money = No
E_3 - Word length = Long	E_6 = word length = Short

$P(H | E_1, E_2, E_3)$

$P(H) = \frac{\text{spam}}{\text{Total}} = \frac{6}{10} = 0.6$

$P(H) = \frac{\text{Ham}}{\text{Total}} = \frac{4}{10} = 0.4$

For spam

$P(E_1) \rightarrow$ YES → containg link

$P(E_1) = \frac{4}{6} = 0.67$

$P(E_2) \rightarrow$ containg money → No

$P(E_2) = \frac{2}{6} = 0.33$

$P(E_3) \rightarrow$ word length → Long

$P(E_3) = \frac{4}{6} = 0.67$

Combined Probability $= P(E_1) \times P(E_2) \times P(E_3)$

$= 0.67 \times 0.33 \times 0.67$

$= 0.148$

For Ham

$P(E_4)$ → containing link → Yes

$P(E_4) = \frac{2}{4} = 0.5$

$P(E_5)$ → containing money → No

$P(E_5) = \frac{3}{4} = 0.75$

$P(E_6)$ → Word length

$P(E_6) = \frac{2}{4} = 0.5$

Combined probability $= P(E_4) \times P(E_5) \times P(E_6)$

$= 0.5 \times 0.75 \times 0.5$

$= 0.187$

Using Bayes theorem

$$P(\text{Spam}|x) = \frac{P(x|\text{spam}) \times P(\text{spam})}{P(x)}$$

$P(\text{spam}|x) = 0.6 \times 0.67 \times 0.33 \times 0.67$

$= 0.6 \times 0.148$

$= 0.0885$

Similarly

$P(\text{Ham}|x) = 0.4 \times 0.5 \times 0.75 \times 0.5$

$= 0.4 \times 0.187$

$= 0.075$

We can clearly see that $P(\text{spam}|x) > p(\text{Ham}|x)$

Therefore, it is classified as SPAM.

K-Nearest Neighbor (KNN)

K = 2

Test example : containing Link = Yes → 1
containing Money = Yes → 1
word length = short → 0

(1, 1, 0)

Euclidean distance

$$d(A, B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

ID	contain link	contain money	word length	class	Distance
1	1	1	1	spam	$\sqrt{(1-1)^2+(1-1)^2+(0-1)^2} = 1$
2	0	0	0	spam	$\sqrt{(0-1)^2+(0-1)^2+(0-0)^2} = 1.41$
3	1	0	1	Ham	$\sqrt{(1-1)^2+(0-1)^2+(1-0)^2} = 1.41$
4	0	1	0	spam	$\sqrt{(1-0)^2+(1-1)^2+(0-0)^2} = 1$
5	1	1	0	spam	$\sqrt{(1-1)^2+(1-1)^2+(0-0)^2} = 0$
6	0	0	1	Ham	$\sqrt{(1-0)^2+(0-1)^2+(1-0)^2} = 1.73$
7	1	0	0	Ham	$\sqrt{(1-1)^2+(1-0)^2+(0-0)^2} = 1$
8	0	1	1	spam	$\sqrt{(1-0)^2+(1-1)^2+(0-1)^2} = 1.41$
9	1	1	1	spam	$\sqrt{(1-1)^2+(1-1)^2+(0-1)^2} = 1$
10	0	0	0	Ham	$\sqrt{(1-0)^2+(1-0)^2+(0-0)^2} = 1.41$

Since K=2, if we see the nearest neighbor, they are ID: 3, 1, 9. And all are labelled as SPAM.

Therefore, the test sample is classified as SPAM

1.b) Please the find the solution in below link :

https://github.com/Sumanth457/is7332025/blob/main/data-mining-project-repo/03032025_CW/CW_03032025.ipynb

2.a) We know that ,

True Positive Rate(TPR) = TP/(TP+FN)

False Positive Rate(FPR) = FP/(FP+TN)

According the formulas, values are computed in the table below

	Threshold	TP	FP	TN	FN	TPR	FPR
0	0.95	39	4	74	33	0.541667	0.051282
1	0.90	46	5	73	26	0.638889	0.064103
2	0.85	51	5	73	21	0.708333	0.064103
3	0.80	54	5	73	18	0.750000	0.064103
4	0.75	55	6	72	17	0.763889	0.076923
5	0.70	58	6	72	14	0.805556	0.076923

The ROC plot is plotted in the below link:

https://github.com/Sumanth457/is7332025/blob/main/data-mining-project-repo/03032025_CW/CW_03032025.ipynb

2.b) Please the find the solution in below link :

https://github.com/Sumanth457/is7332025/blob/main/data-mining-project-repo/03032025_CW/CW_03032025.ipynb