Course: **Natural Language Processing [A]**   03-August-2023   **(Spring 2023)**

Resource Person: **Muhammad Shakeel**   **QUIZ – 3 (Text Classification using Naïve Bayes)**

Total Points: 10

SOLUTION

Consider the following corpus with counts:

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Let's say you are given a fruit that is: **Long**, **Sweet** and **Yellow**, can you predict what fruit it is using Naïve Bayes?

Step 1: Compute the 'Prior' probabilities for each of the class of fruits:

P(Y=Banana) = 500 / 1000 = 0.50

P(Y=Orange) = 300 / 1000 = 0.30

P(Y=Other) = 200 / 1000 = 0.20

Step 2: Compute the probability of evidence:

P(x_1=Long) = 500 / 1000 = 0.50

P(x_2=Sweet) = 650 / 1000 = 0.65

P(x_3=Yellow) = 800 / 1000 = 0.80

Step 3: Compute the likelihood probabilities:

P(x_1=Long | Y=**Banana**) = 400 / 500 = 0.80

P(x_2=Sweet | Y=**Banana**) = 350 / 500 = 0.70

P(x_3=Yellow | Y=**Banana**) = 450 / 500 = 0.90

P(x_1=Long | Y=**Orange**) = 0 / 300 = 0

P(x_2=Sweet | Y=**Orange**) = 150 / 300 = 0.50

P(x_3=Yellow | Y=**Orange**) = 300 / 300 = 1.0

P(x_1=Long | Y=**Other**) = 100 / 200 = 0.5

P(x2=Sweet | Y= **Other**) = 150 / 200 = 0.75

P(x3=Yellow | Y= **Other**) = 50 / 200 = 0.25

Step 4: Compute the final probabilities:

P(**Banana** | Long, Sweet, Yellow) = P(Long | Banana)*P(Sweet | Banana)*P(Yellow | Banana)*P(Banana) /

P(Long)*P(Sweet)*P(Yellow)

= (0.8 * 0.7 * 0.9 * 0.5) / (0.5 * 0.65 * 0.80) = **0.97**

Or, ignoring the denominator as it is same for all probabilities, we get:

= (0.8 * 0.7 * 0.9 * 0.5) = **0.252**

P(**Orange** | Long, Sweet, Yellow) = P(Long | Orange)*P(Sweet | Orange)*P(Yellow | Orange)*P(Orange) /

P(Long)*P(Sweet)*P(Yellow)

= 0 / (0.5 * 0.65 * 0.80) = **0**

P(**Other** | Long, Sweet, Yellow) = P(Long | Other)*P(Sweet | Other)*P(Yellow | Other)*P(Other) /

P(Long)*P(Sweet)*P(Yellow)

= (0.5 * 0.75 * 0.25) / (0.5 * 0.65 * 0.80) = **0.36**

Or, ignoring the denominator, we get:

= (0.5 * 0.75 * 0.25) = **0.094**

As **Banana** has the highest probability, so this will be the predicted class.

END OF QUIZ SOLUTION