Naive Bayes classifier

p(A/B) = p(B/A) p(A)/ p(B)

p(disease/+ve) where p(disease) = 1% , p(test+Ve/disease) = 99%
p(disease/+ve) = 0.5

p(disease/+ve) where p(disease) = 10% , p(test+Ve/disease) = 99%
p(disease/+ve) = ??

p(disease/+ve) = p(+ve/disease) p(disease) / p(+Ve)
= 0.99 * 0.1 / p(test+ve/disease) p(disease) + p(test+ve/nodisease) * p(nodisease)
= 0.99 * 0.1 / ((0.99 * 0.1) + (0.01 * 0.9))
= 0.99* 0.1 / (0.099+0.099)
= 0.099 / 0.108
= 0.916

Problem#1: Email classifier using a naive Bayes classifier

training set:

E1 : “good”.	not spam (NS)
E2 : “ very good”	not spam
E3: “Bad”	spam (S)
E4 : “ very bad”	spam
E5: “ very bad, very bad”	spam

p(NS) = n(NS)/ n(NS) + n(S) = 2/ 2+3 = 2/5 = 0.4
p(S) = n(S) / n(NS) + n(S) = 3/5 = 0.6

unique words are seen in email = {“good”, “very”, “bad”}

p(good) = 2/5 = 0.4
p(very) = 3/5 = 0.6
p(bad) = 3/5 = 0.6

p(NS/very) = 1/ 3 = 0.33
p(S/good) = 0 = 0.05
p(NS/bad) = 0 = 0.05

p(good/NS) = p(NS/good) P(good) /p(NS) = 0.4 * 0.4 / 0.4 = 0.4
p(good/S) = p(S/good) p(good) / p(S) / = 0.05 * 0.4 / 0.6 = 0.033

p(very/NS) = p(NS/very) p(very) / p(NS) = 0.33 * 0.6 / 0.4 = 0.495
p(very/S) = p(S/very) p(very) / p(S) = 0.66 * 0.6 / 0.6 = 0.66

p(bad/NS) = p(NS/bad) * p(bad) / p(NS) = 0.05 * 0.6 / 0.4 = 0.075
p(bad/S) = p(S/bad) * p(bad) / p(S) = 1 * 0.6 / 0.6 = 1

test data
E6 : good bad very bad

p(E6/S) = P(good/S) * p(bad/S) * p(very/S) = 0.033 * 1 * 0.66 = 0.02178
p(E6/NS) = p(good/NS) * p(bad/NS) * p(very/NS) = 0.4 * 0.075 * 0.495 = 0.01485

p(S/E6) = P(E6/S) * P(S) / p(E6) = 0.6 * p(E6/S) / p(E6) = 0.6 * 0.02178 / p(E6) = 0.0130/p(E6)

p(NS/E6) = p(E6/NS) p(NS) / p(E6) = 0.4 * p(E6/NS) / p(E6) = 0.4 * 0.01485 / p(E6) = 0.0059/p(E6)

say p(E6) = 1% = 0.01

p(S/E6) = 0.0013
p(NS/E6) = 0.00059

since p(S/E6) > p(NS/E6) , E6 will be classified as “S” aka SPAM

Problem#2: Classify fruits using a naive Bayes classifier

Let’s say the fruit basket consists of fruits as shown in below table

Fruit	Total count
Bananas (B)	50
Orange (O)	30
Other fruit (OF)	100

Attributes	Long (L)	Not long (NL)	Sweet(S)	Not Sweet(NS)	Yellow(Y)	Not yellow (NY)
Bananas	40	10	35	15	45	5
Orange	0	30	15	15	30	0
Other fruit	50	50	65	35	80	20
Total	90	90	115	65	155	25

p(B) = n(B) / n(Total fruits in the basket) = 50 / 50+30+100 = 50/180 = 0.2778
p(O) = 30/180 = 1/6 = 0.1667
p(OF) = 100/ 180 = 10/18 = 5/9 = 0.5556

p(L) = n(L)/n(L) + n(NL) = 90 / 90+90 = 0.5
p(S) = n(S) / n(S) + n(NS) = 115 / 115+65 = 115/180
p(Y) = n(Y) / n(Y) + n(NY) = 155 / 155+25 = 155/ 180

p(L, S, Y) = p(L) * p(S) * p(Y) = 0.5 * 115/180 * 155/180 =

p(L, S, Y/ B) = p(L/B) * p(S/B) * p(Y/B) = 40/50 * 35/50 * 45/50 =

p(B/ L, S, Y) = p(L, S, Y/ B) p(B)/ p(L, S, Y) = 40/50 * 35/50 * 45/50 * 0.2778 / (0.5 * 115/180 * 155/180) =
p(B/ L, S, NY) =
p(B/ L, NS, Y) =
p(B/ L, NS, NY)
p(B/ NL, S, Y) =
p(B/ NL, S, NY) =
p(B/ NL, NS, Y) =
p(B/ NL, NS, NY) =

p(O/ L, S, Y) =
p(O/ L, S, NY) =
p(O/ L, NS, Y) =
p(O/ L, NS, NY)
p(O/ NL, S, Y) =
p(O/ NL, S, NY) =
p(O/ NL, NS, Y) =
p(O/ NL, NS, NY) =

p(OF/ L, S, Y) =
p(OF/ L, S, NY) =
p(OF/ L, NS, Y) =
p(OF/ L, NS, NY)
p(OF/ NL, S, Y) =
p(OF/ NL, S, NY) =
p(OF/ NL, NS, Y) =
p(OF/ NL, NS, NY) =

Given it is long, sweet & yellow, which fruit it is likely to be ?