<u>Practice Exercise 2 – Bayesian Classification (solution)</u>

Question 1: A patient is tested by a lab test for a disease that has prevalence of 1 in 1000 in the population. The lab test has a false-positive rate of 1% and a false negative rate of 1%.

a) If the lab test result is positive, what is the probability that the patient actually has the disease?

P(disease) = 1/1000 = 0.001; P(not-disease) = 1 - 0.001 = 0.999

P(test=positive|not-disease) = 1/100 = 0.01; P(test=negative|disease) = 1/100 = 0.01

P(test=negative|not-disease) = 1 - 0.01 = 0.99; P(test=positive|disease) = 1 - 0.01 = 0.99

Using Bayes rule:

P(disease|test=positive) =

P(test=positive|disease)P(disease)/P(test=positive)

Finding individual values:

P(test=positive|disease) = 0.99

P(disease) = 0.001

P(test=positive) = P(test=positive|disease)P(disease)+P(test=positive|not-disease)P(not-disease)

P(test=positive) = 0.99*0.001 + 0.01*0.999 = 0.00099 + 0.00999 = 0.01098

So

P(disease|test=positive) = 0.99*0.001/0.01098 = 0.0901

b) Is it more probable that the patient has the disease or not?

P(not-disease|test=positive) = 1-0.0901 = 0.9099

Thus, it's more probable that the patient does not have the disease.

c) Would the answers to a) and b) differ if a maximum likelihood versus maximum a posteriori hypothesis estimation method is used? Comment on your answer.

Since the disease prevalence (0.001) is very low, it has a high impact on probability and would change the decision. The ML vs MAP estimate can be calculated to observe this.

For example, the MAP estimate for disease case will be:

P(disease|test=positive) = P(test=positive|disease)P(disease) = 0.99*0.001 = 0.00099

while the ML estimate for disease case will be:

P(disease|test=positive) = P(test=positive|disease) = 0.99

Question 3: Assume we have a data set described by the following three variables (i.e. attributes):

Hair = {B,D}, where B=blonde, D=dark.

Height = {T,S}, where T=tall, S=short.

Country = {G,P}, where G=Greenland, P=Poland.

You are given the following training data set (Hair, Height, Country):

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

We want to answer the following question: If you observe a new individual who is tall with blonde hair, what is the most likely country of origin?

a) Find the maximum a posteriori (MAP) estimate to the above question, using the Naïve Bayes assumption. Show all of your working.

P(P|B,T) = P(B,T|P)P(P) = P(B|P)B(T|P)P(P) = 4/8 * 5/8 * 8/24 = 0.5 * 0.625 * 0.33 = 0.1031

P(G|B,T) = P(B,T|G)P(G) = P(B|G)B(T|G)P(G) = 8/16 * 10/16 * 16/24 = 0.5 * 0.625 * 0.67 = 0.2094

Hence, an individual with blonder hair and tall height is more likely to belong to G.

b) Find the maximum a posteriori (MAP) estimate to the above question, without using the Naïve Bayes assumption. Show all of your working.

P(P|B,T) = P(B,T|P)P(P) = 3/8 * 8/24 = 0.375 * 0.625 = 0.2344

P(G|B,T) = P(B,T|G)P(G) = 4/16 * 16/24 = 0.25 * 0.625 = 0.1562

Hence, an individual with blonder hair and tall height is more likely to belong to P.

c) Find the Maximum Likelihood (ML) estimate to the above question, using the Naïve Bayes assumption.

P(P|B,T) = P(B,T|P) = P(B|P)B(T|P) = 0.5 * 0.625 = 0.3125

P(G|B,T) = P(B,T|G) = P(B|G)B(T|G) = 0.5 * 0.625 = 0.3125

Hence, an individual with blonder hair and tall height is equally likely to belong to G or P.

d) Find the Maximum Likelihood (ML) estimate to the above question, without using the Naïve Bayes assumption.

P(P|B,T) = P(B,T|P) = 3/8 = 0.375

P(G|B,T) = P(B,T|G) = 4/16 = 0.25

Hence, an individual with blonder hair and tall height is more likely to belong to P.