Exercise 5: Bayes rule – Pablo Castro Ilundain

Question 1)

When I had to use the load() function on MATLAB, I realized that the data is a struct with 2 different fields, in this case, arrays: x1 and x2. Their length is 100x1, I studied the covariance between the arrays and the result is:

```
c =

    1.2565    0.0076

    0.0076    0.6610
```

Because of this, I considered each one as a totally different dataset and I tested both with different testing methods, in my case I used kstest() and lillietest().

```
% Exercise 5 - Question 1
clear all; close all;
load('t096');
c = cov(x1,x2);

%% Testing for the first array
m1 = mean(x1);
s1 = std(x1);
normdata1 = ((x1-m1)/s1);
[hK1,pK1] = kstest(normdata1);
[hL1,pL1] = lillietest(x1);

%% Testing for the second array
m2 = mean(x2);
s2 = std(x2);
normdata2 = ((x2-m2)/s2);
[hK2,pK2] = kstest(normdata2);
[hL2,pL2] = lillietest(x2);
```

I had to normalize the data before using kstest() because the input data for testing must be scaled and centered (by default it tests for a standard normal distribution), so I normalized it by doing

$$normdata = \frac{x - mean(x)}{std(x)}$$

After running this script and using this test methods (kstest() and lillietest()), we can extract from our results:

```
hK1 = 0 (logical)

pK1 = 0.7731

hL1 = 0

pL1 = 0.3697
```

This means that the first array fails to reject the null hypothesis at the default 5% significance level on both tests and the value of p is the probability of observing a test statistic with the same value as the observed value or even more than the one under the null hypothesis.

```
hK2 = 1 (logical)

pK2 = 0.0097

hL2 = 1

pL2 = 1.0000e-03
```

This means that the second array rejects the null hypothesis at the default 5% significance level. In this case there is a warning when we use the lillietest() function and it's because the returned p-value is the smallest value in the table of precomputed values (is not very accurate).

The conclusion is that, we can say that the first array rejects the null hypothesis and the second doesn't, this means that the first array does not come from a normal distribution and the second does.

Question 2)

Bayes rule is the following equation:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

The exercise gives us the following information:

- Probability that the test has succeeded revealing cancer: $P(Pos|C) = 0.98$
- Probability that the test is positive (patient without cancer): $P(Pos|NC) = 0.03$
- Probability of having cancer $P(C) = 0.008$

To respond to "What is the probability that Mr. Onymous has cancer when the test is positive?" as if we must calculate $P(C|Pos)$:

$$P(C|Pos) = \frac{P(Pos|C) * P(C)}{P(Pos)}$$

We must calculate $P(Pos)$ and it's going to be:

$$P(Pos) = P(C) * P(Pos|C) + P(NC) * P(Pos|NC) \rightarrow$$

$$\rightarrow 0.008 * 0.98 + (1 - 0.008) * 0.03 \ = \ 0.0376$$

Then, the last step is:

$$P(C|Pos) = \frac{0.98 * 0.008}{0.0376} = \ 0{,}2085$$

So, the conclusion is that the probability that Mr. Onymous has cancer when the test is positive is 20,85%.

Question 3 & 4)

The answers to both questions are answered on the MATLAB code.