# Datamining Assignment 4

Elkhan Ismayilzada

June 5, 2021

## Exercise 1

### 1.1

Probabilities are the areas under a fixed distribution. To put it as a syntax we can write it as $pr(data|distribution)$ which means given a distribution what is the probability of some data which will be the area under the curve. On the other hand, likelihoods are the y-axis values for fixed data points with distributions that can be moved. In a syntax notation we can write it as $L(distribution|data)$ which means that what is the likelihood of given fixed data under some distributions. In probability case, the distribution does not change and we can change the data as we wish. However, in likelihood case, we do not change the data, we change the distribution.

### 1.2

In statistics, sampling is to use the subset of examples from the population to estimate the population parameter. This process is done only once. In resampling, however, we do this process several times in order to effectively find out the estimation of population parameter. This clearly shows the advantage of resampling over sampling. However, resampling is computationally expensive since we estimate the population parameter several times. Sampling includes random sampling and stratified sampling whereas resampling includes k-fold cross-validation and boostrapping.

## Exercise 2

### 2.1

We can put it into the following equation

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = 42.6 - 0.0003 \cdot lsep + 6.7 \cdot wsep - 9.4 \cdot lpet - 18.3 \cdot wpet \tag{2.1}$$

Where $lsep, wsep, lpet, wpet$ are length, width of sepals, length, width of petals respectively If we increase $lsep$ then the right hand side decreases which implies that in left hand side the denominator overcomes the numerator. And if we decrease it then numerator overcomes the denominator. To make things more simple, lets write the equation as follows

$$\begin{gathered}\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -0.0003 \cdot lsep + C \\ \frac{P(Y=1)}{1-P(Y=1)} = (e^{-0.0003})^{lsep} \cdot e^{C} = 0.9997^{lsep} \cdot e^{C}\end{gathered} \tag{2.2}$$

As we do not change the other values, we can consider them as constant. Since we are dealing with P(Y=1) $lsep$ is the length of sepals Y=1 which is Versicolor. So in conclusion

**Species Versicolor is 0.9997 times more probable than species Virginica when the length of sepals increases.**

**2.2**

For $x_1$

$$\begin{aligned}
\ln\left(\frac{P(Y=1|x_1)}{1-P(Y=1|x_1)}\right) &= 42.6 - 0.0003\cdot 6.4 + 6.7\cdot 3.1 - 9.4\cdot 4.3 - 18.3\cdot 1.3 = -0.842 \\
\frac{P(Y=1|x_1)}{1-P(Y=1|x_1)} &= e^{-0.842} = 0.431 \\
P(Y=1|x_1) &= 0.431 - 0.431\cdot P(Y=1|x_1) \\
P(Y=1|x_1) &= \frac{0.431}{1.431} = 0.301 \\
P(Y=0|x_1) = 1 - P(Y=1|x_1) &= 1 - 0.301 = 0.699
\end{aligned} \tag{2.3}$$

Now for $x_2$

$$\begin{aligned}
\ln\left(\frac{P(Y=1|x_2)}{1-P(Y=1|x_2)}\right) &= 42.6 - 0.0003\cdot 6.9 + 6.7\cdot 3.0 - 9.4\cdot 3.9 - 18.3\cdot 1.4 = 0.418 \\
\frac{P(Y=1|x_2)}{1-P(Y=1|x_2)} &= e^{0.418} = 1.519 \\
P(Y=1|x_2) &= 1.519 - 1.519\cdot P(Y=1|x_2) \\
P(Y=1|x_2) &= \frac{1.519}{2.519} = 0.603 \\
P(Y=0|x_2) = 1 - P(Y=1|x_2) &= 1 - 0.603 = 0.397
\end{aligned} \tag{2.4}$$

# Exercise 3

### 3.1

$$\begin{aligned} K(x,z) = (x^T \cdot z)^2 = (x^T \cdot z)(x^T \cdot z) \\ = (x_1 \cdot z_1 + x_2 \cdot z_2)(x_1 \cdot z_1 + x_2 \cdot z_2) \\ = x_1^2 \cdot z_1^2 + 2 \cdot x_1 \cdot z_1 \cdot x_2 \cdot z_2 + x_2^2 \cdot z_2^2 \\ = [x_1^2, \sqrt{2} \cdot x_1 \cdot x_2, x_2^2]^T \cdot [z_1^2, \sqrt{2} \cdot z_1 \cdot z_2, z_2^2] = \phi(x)^T \cdot \phi(z) \end{aligned} \tag{3.1}$$

### 3.2

$$\begin{aligned} \phi(x_1) = [1^2, \sqrt{2} \cdot 1 \cdot 2, 2^2] = [1, 2.828, 4] \\ \phi(x_2) = [4^2, \sqrt{2} \cdot 4 \cdot 3, 3^2] = [16, 16.971, 9] \\ \phi(x_3) = [2^2, \sqrt{2} \cdot 2 \cdot 0, 0] = [4, 0, 0] \end{aligned} \tag{3.2}$$

### 3.3

We have

$$x_1 = (1,2), x_2 = (4,3), x_3 = (2,0) \tag{3.3}$$

Therefore,

| | | |
|---|---|---|
| $K(x_1, x_1)$ | $K(x_1, x_2)$ | $K(x_1, x_3)$ |
| $K(x_2, x_1)$ | $K(x_2, x_2)$ | $K(x_2, x_3)$ |
| $K(x_3, x_1)$ | $K(x_3, x_2)$ | $K(x_3, x_3)$ |

| | | |
|---|---|---|
| 25 | 100 | 4 |
| 100 | 625 | 64 |
| 4 | 64 | 16 |

# Exercise 4

### 4.1

The equation for naive bayes classifier is given as

$$P(y|X_1, X_2, \dots X_n) = \frac{P(y) \cdot \prod_{i=1}^{n} P(X_i|y)}{P(X_1) \cdot P(X_2) \dots P(X_n)} \tag{4.1}$$

We can remove the denominator since it remains constant for a given input

$$P(y|X_1, X_2, \dots X_n) \propto P(y) \cdot \prod_{i=1}^{n} P(X_i|y) \tag{4.2}$$

Now we are going to find all $P(X_i|y)$ and $P(y)$

$$\begin{aligned}
X_1 = a &=> [A, B, A], P(X_1 = a|Y = A) = \frac{2}{4}, P(X_1 = a|Y = B) = \frac{1}{3} \\
X_1 = b &=> [B, B, A, A], P(X_1 = b|Y = A) = \frac{2}{4}, P(X_1 = b|Y = B) = \frac{2}{3} \\
X_2 = a &=> [B, A, A], P(X_2 = a|Y = A) = \frac{2}{4}, P(X_2 = a|Y = B) = \frac{1}{3} \\
X_2 = b &=> [A, B, B, A], P(X_2 = b|Y = A) = \frac{2}{4}, P(X_2 = b|Y = B) = \frac{2}{3} \\
X_3 = a &=> [B, B, A, A], P(X_3 = a|Y = A) = \frac{2}{4}, P(X_3 = a|Y = B) = \frac{2}{3} \\
X_3 = b &=> [A, B, A], P(X_3 = b|Y = A) = \frac{2}{4}, P(X_1 = b|Y = B) = \frac{1}{3} \\
y = A => [A, A, A, A], P(Y = A) = \frac{4}{7}, y = B &=> [B, B, B], P(Y = B) = \frac{3}{7}
\end{aligned} \tag{4.3}$$

Now we can calculate the given probability question

$$\begin{aligned}
P(y = A|X_1 = a, X_2 = b) \propto P(y = A) \cdot P(X_1 = a|Y = A) \cdot P(X_2 = b|y = A) &= \frac{4}{7} \cdot \frac{2}{4} \cdot \frac{2}{4} = 0.143 \\
P(y = A|X_1 = a, X_2 = b) = \frac{P(y = A) \cdot P(X_1 = a|Y = A) \cdot P(X_2 = b|y = A)}{P(X_1 = a) \cdot P(X_2 = b)} &= \frac{4}{7} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{7}{3} \cdot \frac{7}{4} = 0.583
\end{aligned} \tag{4.4}$$

### 4.2

$$\begin{aligned}
P(Y = A|X_1 = a, X_2 = b, X_3 = b) \propto P(y = A) \cdot P(X_1 = a|Y = A) \cdot P(X_2 = b|y = A) \cdot P(X_3 = b|y = A) \\
= \frac{4}{7} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} = 0.071 \\
P(Y = B|X_1 = a, X_2 = b, X_3 = b) \propto P(y = B) \cdot P(X_1 = a|Y = B) \cdot P(X_2 = b|y = B) \cdot P(X_3 = b|y = B) \\
= \frac{3}{7} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = 0.032 \\
P(Y = A|X_1 = a, X_2 = b, X_3 = b) + P(Y = B|X_1 = a, X_2 = b, X_3 = b) \\
= 0.071 + 0.032 = 0.103 \\
P(Y = A|X_1 = a, X_2 = b, X_3 = b) = \frac{0.071}{0.103} = 0.689 \\
P(Y = B|X_1 = a, X_2 = b, X_3 = b) = \frac{0.032}{0.103} = 0.311
\end{aligned} \tag{4.5}$$

# Exercise 5

**5.1** What is the significance of $C$ value in SVM?[10pts] If we are given a data in which we predict whether the patient has breast cancer or not, should the $C$ value be large or small? Justify your answer[5pts].

**Solution**

The equation of soft margin is given by

$$\phi(w) = \frac{1}{2} w^T \cdot w + C \sum \xi_i \tag{5.1}$$

$C$ is a hyperparameter that decides the trade-off between maximizing the margin and minimizing the mistakes. If we choose $C$ value large, it means we focus on classification error more, hence we have very few classification errors and the margin will be small. On the other hand, if we have small $C$ value , it means we focus on margin error more, hence we will have large margin but we may make more classification errors. The given data is about patient's health which implies that we should make classification errors as small as possible. Therefore, C parameter should be a large value.