

Statistical Machine Learning 2018  
Assignment 2  
Deadline: 28th of October 2018

Christoph Schmidl s4226887 c.schmidl@student.ru.nl	Mark Beijer s4354834 mbeijer@science.ru.nl
--	--

October 16, 2018

## Exercise 1 - weight 3

The financial services department of an insurance company receives numerous phone calls each day from people who want to make a claim against their policy. Most claims are genuine, however about 1 out of every 6 are thought to be fraudulent. To tackle this problem the company has installed a trial version of a software voice-analysis system that monitors each conversation and gives a numerical score  $z$  between 0 and 1, depending on allegedly suspicious vocal intonations of the customer. Unfortunately, nobody seems to know anymore how to interpret the score in this particular version of the system...

Tests revealed that the conditional probability density of  $z$ , given that a claim was valid ( $c = 1$ ) or false ( $c = 0$ ) are

$$\begin{aligned} p(z|c=0) &= \alpha_0(1-z^2), \\ p(z|c=1) &= \alpha_1 z(z+1). \end{aligned}$$

### 1.1

Compute the normalization constants  $\alpha_0$  and  $\alpha_1$ . How does the  $z$  score relate to the validity of the claim? What values for  $z$  would you expect when the claim is valid / false?

**Answer:**

### 1.2

Use the sum and product rule to show that the probability distribution function  $p(z)$  can be written as

$$p(z) = \frac{(3z+1)(z+1)}{4} \tag{1}$$

**Answer:**

### 1.3

Use Bayes' rule to compute the posterior probability distribution function  $p(c|z)$ . Plot these distributions in MATLAB as a function of  $z$ . How can these posterior probabilities help in making a decision regarding the validity of the claim?

**Answer:**

## 1.4

Compute the optimal decision boundary (based on our numerical score  $z$ ) that minimizes the misclassification rate. For which  $z$  should we classify  $c = 0$  (false) and for which  $z$  should we classify  $c = 1$  (valid)? Explain your decision.

**Answer:**

## 1.5

Compute the misclassification rate, given the optimal decision boundary determined previously. Interpret the result you have obtained. Is the  $z$  score useful in determining the validity of the claim? Compare this with your prior guess from 1.

**Answer:**

## Exercise 2 - weight 2

The government of the United Kingdom has decided to call a referendum regarding the country's European Union membership. The citizens of the UK will be asked the following question at the referendum: "Should the United Kingdom remain a member of the European Union or leave the European Union?". The European Commission (EC) is interested in the potential outcome of this referendum and has contracted a polling agency to study this issue.

Suppose that a person's vote follows a Bernoulli distribution with parameter  $\theta$  and suppose the EC's prior distribution for  $\theta$ , the proportion of British citizens that would be in favor of leaving the EU, is beta distribution with  $\alpha = 90$  and  $\beta = 110$ .

### 2.1

Determine the mean and variance of the prior distribution. Plot the prior density function.

**Answer:**

### 2.2

A random sample of 1000 British citizens is taken, and 60% of the people polled support leaving the European Union. What are the posterior mean and variance for  $\theta$ ? Plot the posterior density function together with the prior density. Explain how the data from the sample changed the prior belief.

**Answer:**

### 2.3

Examine the effect of changing the prior hyperparameters  $(\alpha, \beta)$  on the posterior by looking at several other hyperparameter configurations. Which values for  $\alpha$  and  $\beta$  correspond to a non-informative prior? What is the interpretation of  $\alpha$  and  $\beta$  for the beta prior? What does the choice of  $\alpha$  and  $\beta$  in Question 1 tell you about the strength of the prior belief?

**Answer:**

### 2.4

Imagine you are now a reporter for the polling agency and you have been sent on field duty to gather more data. Your mission is to go out on the streets and randomly survey people on their thoughts regarding the upcoming referendum. Given all the available information you have acquired, what is the probability that

the first person you talk to will vote 'Leave'?

**Hint:** Derive the predictive distribution for the next vote using the posterior distribution for  $\theta$  computed in Question 2. For a reminder on predictive distribution, see subsection 1.2.6 in Bishop, in particular Equation (1.68).

**Answer:**

## Exercise 3 - Sequential learning (weight 5)

### Part 1 - Obtaining the prior

Consider a four dimensional variable  $[x_1, x_2, x_3, x_4]^T$ , distributed according to a multivariate Gaussian with mean  $\tilde{\mu} = [1, 0, 1, 2]^T$  and covariance matrix  $\tilde{\Sigma}$  given as

$$\tilde{\Sigma} = \left( \begin{array}{cc|cc} 0.14 & -0.3 & 0.0 & 0.2 \\ -0.3 & 1.16 & 0.2 & -0.8 \\ \hline 0.0 & 0.2 & 1.0 & 1.0 \\ 0.2 & -0.8 & 1.0 & 2.0 \end{array} \right)$$

We are interested in the conditional distribution over  $[x_1, x_2]^T$ , given that  $x_3 = x_4 = 0$ . We know this conditional distribution will also take the form of a Gaussian:

$$p([x_1, x_2]^T \mid x_3 = x_4 = 0) = \mathcal{N}([x_1, x_2]^T \mid \mu_p, \Sigma_p) \quad (2)$$

for which the mean and covariance matrix are most easily expressed in terms of the (partitioned) precision matrix (see Bishop, 2.3.1).

#### 3.1.1

Use the partitioned precision matrix  $\tilde{\Lambda} = \tilde{\Sigma}^{-1}$  to give an explicit expression for the mean  $\mu_p$  and covariance matrix  $\Sigma_p$  of this distribution and calculate their values. (This distribution will be taken as the prior information for the rest of this exercise, hence the subscript p). You may use the MATLAB command `inv` to calculate the matrix inverses.

**Answer:**

#### 3.1.2

[MATLAB] - Create a function that can generate random number pairs, distributed according to the distribution in 2. Initialize your random generator and the draw a single pair

$$\mu_t = [\mu_{t1}, \mu_{t2}]^T \quad (3)$$

from this distribution. (These will be the 'true' means, hence the subscript t).

**Hint:** you can use the MATLAB function `mvnrnd` (which resides the Statistics toolbox).

**Answer:**

#### 3.1.3

[MATLAB] - Make a plot of the probability density of the distribution 2.

**Hint:** Use the MATLAB function `mvnpdf` (which resides in the Statistics toolbox) to calculate the probability density of a multivariate Gaussian random variable. The MATLAB functions `meshgrid` and `surf` may also prove useful.

**Answer:**

Part 2 - Generating the data  
Part 3 - Sequential learning algorithms

## Hints