SML Assignment 2

Daniele Maccari 4711262, Lisa Tostrams s4386167

October 29, 2017

Exercise 1

1. To compute the normalization constants, we need to solve the following equations for α_0 and α_1, respectively:

$$\int_0^1 \alpha_0(1-z^2)\,dz = 1 \tag{1}$$
$$\int_0^1 \alpha_1(z^2+z)\,dz = 1 \tag{2}$$

Let's start with 1:

$$\begin{aligned}
\int_0^1 \alpha_0(1-z^2)\,dz &= 1 \\
\alpha_0\left(\int_0^1 1\,dz - \int_0^1 z^2\,dz\right) &= 1 \\
\alpha_0\left(z\,\Big|_0^1 - \frac{z^3}{3}\,\Big|_0^1\right) &= 1 \\
\alpha_0\left(1-\frac{1}{3}\right) &= 1 \\
\alpha_0 &= \frac{3}{2}
\end{aligned}$$

Similarly, for 2 we have

$$\begin{aligned}
\int_0^1 \alpha_1(z^2+z)\,dz &= 1 \\
\alpha_1\left(\int_0^1 z^2\,dz + \int_0^1 z\,dz\right) &= 1 \\
\alpha_1\left(\frac{z^3}{3}\,\Big|_0^1 + \frac{z^2}{2}\,\Big|_0^1\right) &= 1 \\
\alpha_1\left(\frac{1}{3}+\frac{1}{2}\right) &= 1 \\
\alpha_1 &= \frac{6}{5}
\end{aligned}$$

By plotting the two likelihoods together as in Figure 1, we can observe how for extreme values of z, i.e. very close to 0 and 1, the score is quite more likely to have been caused by a fraud or a

valid claim, respectively. For intermediate values of the score, however, the difference between the two diminishes, making z less useful. (An alternative is disregarding z completely - assuming every claim is valid will result in a misclassification rate of $\frac{1}{6} \approx 0.16$.) Of course, we are not interested in the likelihood alone, but in the posterior probability $p(c|z)$, which we will compute later.

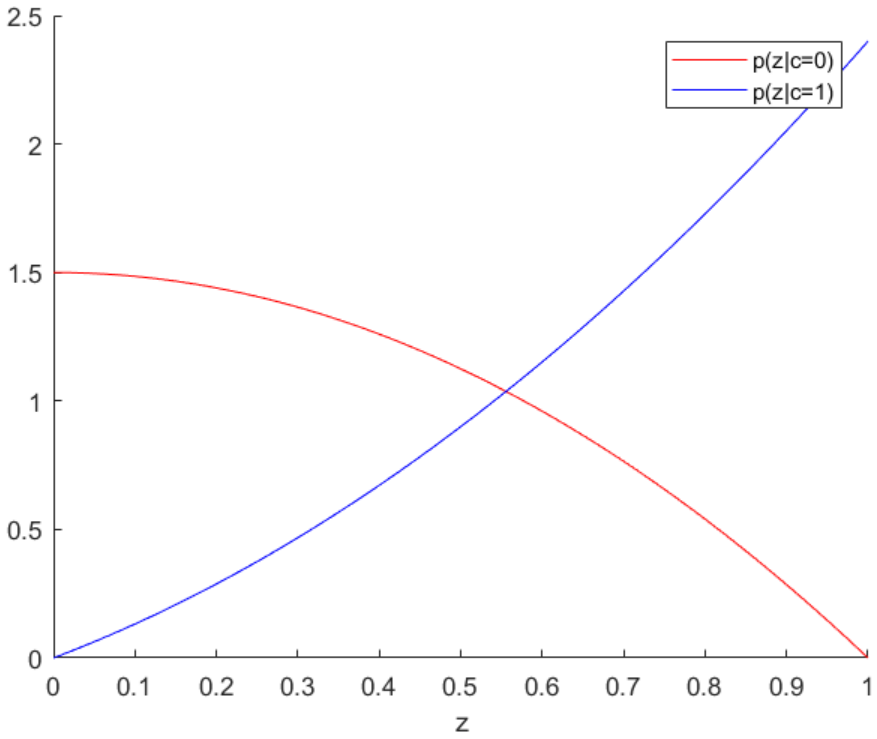

Figure 1: [Exercise 1.1] The likelihoods $p(z|c)$

2. By using the sum and product rule we have

$$\begin{aligned} p(z) &= p(z|c=0)p(c=0) + p(z|c=1)p(c=1) \\ &= \frac{3}{2}(1-z^2)\frac{1}{6} + \frac{6}{5}(z^2+z)\frac{5}{6} \\ &= \frac{1}{4}(1-z^2) + z^2 + z \\ &= \frac{1 - z^2 + 4z^2 + 4z}{4} \\ &= \frac{3z^2 + 4z + 1}{4} \\ &= \frac{(3z+1)(z+1)}{4} \end{aligned}$$

3. Posterior probability $p(c=0|z)$ (see also Figure 2):

$$
\begin{aligned}
p(c=0|z) &= \frac{p(z|c=0)p(c=0)}{p(z)} && \text{(using } \textit{Bayes rule}\text{)} \\
&= \frac{\frac{3}{2}(1-z^2)\frac{1}{6}}{(3z+1)(z+1)\frac{1}{4}} \\
&= \frac{1-z^2}{(3z+1)(z+1)}
\end{aligned}
$$

Posterior probability $p(c=1|z)$ (see also Figure 2):

$$
\begin{aligned}
p(c=1|z) &= \frac{p(z|c=1)p(c=1)}{p(z)} && \text{(using } \textit{Bayes rule}\text{)} \\
&= \frac{z(z+1)}{(3z+1)(z+1)\frac{1}{4}} \\
&= \frac{z}{(3z+1)\frac{1}{4}} \\
&= \frac{4z}{3z+1}
\end{aligned}
$$

Observing a z-score gives you information about the likeliness of having a valid of false claim. This

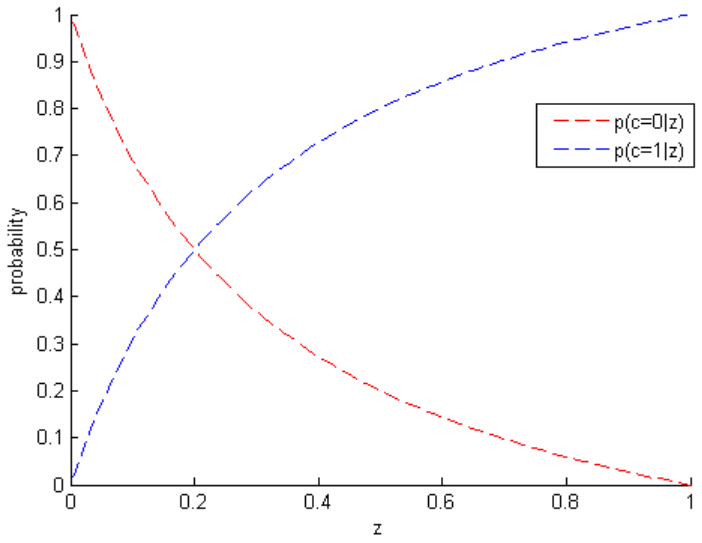

Figure 2: [Exercise 1.3] The posterior distributions $p(c=0|z)$ and $p(c=1|z)$

information is encoded in the posteriors given above. Observing a certain z-score, you can compute the probabilities of a valid and a false claim, and decide on which is most probable.

4. The misclassification rate is defined by the area under $p(c=1|z)$ to the left of the decision boundary, plus the area under $p(c=0|z)$ to the right of the decision boundary. Therefore, the optimal decision boundary that minimizes misclassification is given by the value of z for which $p(c=0|z) = p(c=$

$1|z$):

$$\begin{aligned}
\frac{1-z^2}{(3z+1)(z+1)} &= \frac{4z}{3z+1} \\
\frac{1-z^2}{z+1} &= 4z \\
\frac{-z^2+1}{z^2+z} &= 4 \\
-1+\frac{1}{z} &= 4 \\
\frac{1}{z} &= 5 \\
z &= \frac{1}{5}
\end{aligned}$$

5. The misclassification rate given decision boundary 0.2 is given by $\int_0^{0.2} p(z|c=1)p(c=1)dz + \int_{0.2}^1 p(c=0|z)p(c=0)dz$:

$$\begin{aligned}
\int_0^{0.2} p(z|c=1)p(c=1)dz + \int_{0.2}^1 p(c=0|z)p(c=0)dz &= \int_0^{0.2} z(z+1)dz + \int_{0.2}^1 (1-z^2)\frac{1}{4}dz \\
&= \int_0^{0.2} z^2+zdz + \frac{1}{4}\int_{0.2}^1 (1-z^2)dz \\
&= \left(\frac{z^3}{3}\Big|_0^{0.2} + \frac{z^2}{2}\Big|_0^{0.2}\right) + \frac{1}{4}\left(z\Big|_{0.2}^1 - \frac{z^3}{3}\Big|_{0.2}^1\right) \\
&= \left(\frac{1}{375}+\frac{1}{50}\right) + \frac{1}{4}\left(\left(1-\frac{1}{5}\right) - \left(\frac{1}{3}-\frac{1}{375}\right)\right) \\
&= \frac{1}{375}+\frac{1}{50}+\frac{1}{4}\left(\frac{4}{5}-\frac{1}{3}+\frac{1}{375}\right) \\
&= \frac{1}{375}+\frac{1}{50}+\frac{1}{5}-\frac{1}{12}+\frac{1}{1500} \\
&= \frac{7}{50} = 0.14
\end{aligned}$$

Using the z score to decide whether a claim is valid or false results in a misclassification rate of 0.14. As mentioned in 1.1, not using z will result in a misclassification rate of $\frac{1}{6} \approx 0.16$. Looking at these number alone, using z results in a slight improvement in correct decisions of valid and false claims. However, this comparison does not take into account that not using z results in *false positive* (accepting a false claim as valid) errors only - assuming every claim is valid, means that no valid claims are rejected accidentally. Using z results in both *false positive* and *false negative* (rejecting a valid claim) errors. Depending on the cost of granting money to false claims and fighting lawsuits from rejecting valid claims, these types of misclassification might influence the decision to use the z score to determine the validity of the claim.

1 Exercise 2

1. The given problem can be modeled by a multinomial distribution. The likelihood for the total number of members of parliament elected for each party is thus:

$$\frac{\Gamma(\sum_{k=1}^4 m_k)}{\prod_{k=1}^4 \Gamma(m_k)} \prod_{k=1}^4 \mu_k^{m_k}$$

where k is the number of parties, μ_k is the probability of the k-th party to get a seat, and m_k is the number of seats received by the k-th party, with $\sum_k m_k = 533$ when all seats have been distributed.

2. A convenient prior distribution over the vector μ of probabilities is the Dirichlet distribution, which is the conjugate prior for the multinomial distribution. Its pdf is defined as

$$\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \mu_k^{\alpha_k - 1}$$

In particular, this probability is parameterized by a vector $\boldsymbol{\alpha}$ of *pseudocounts*, which act as weights for the probabilities $\boldsymbol{\mu_k}$. For example, if $\alpha^T = (1, 1, 1, 1)$ we have a uniform distribution over values of $\boldsymbol{\mu}$. In our case, we set $\boldsymbol{\alpha}^T = (38, 34, 22, 6)$, i.e. the percentage of people voting for each party according to the survey. This constitutes our prior belief about the number of seats each party is likely to get.

3. The next seat being assigned to the UKIP given the following observations:

$$\boldsymbol{m}^T = (15, 11, 4, 0).$$

Hence, we need to compute the probability of $m_4 = 1$ given the rest of the observations. By applying the rules of probability we know that

$$p(m_4 = 1|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{m}_{1:3}) = p(m_4 = 1|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{m}_{1:3}) \tag{3}$$

And we know that the posterior distribution of a Dirichlet takes the form

$$p(\mu|\boldsymbol{\alpha}, \boldsymbol{m}) = \frac{\Gamma(\alpha_0 + \sum_{k=1}^4 m_k)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \prod_{k=1}^4 \mu_k^{\alpha_k + m_k - 1} \tag{4}$$

By marginalizing over $\boldsymbol{\mu}$ we then have

$$\begin{aligned}
p(m_4 = 1) &= \int_{\boldsymbol{\mu}} p(m_4 = 1|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{m}_{1:3}) \mathrm{d}\boldsymbol{\mu} \\
&= \int_{\boldsymbol{\mu}} \mu_4 \cdot p(\boldsymbol{\mu}|\boldsymbol{\alpha}, \boldsymbol{m}_{1:3}) \mathrm{d}\boldsymbol{\mu} \\
&= \int_{\boldsymbol{\mu}} \mu_4 \frac{\Gamma(\alpha_0 + \sum_{k=1}^4 m_k)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \prod_{k=1}^4 \mu_k^{\alpha_k + m_k - 1} \mathrm{d}\boldsymbol{\mu} \\
&= \int_{\boldsymbol{\mu}} \mu_4 \frac{\Gamma(100 + 30)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \prod_{k=1}^4 \mu_k^{\alpha_k + m_k - 1} \mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(130)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \int_{\boldsymbol{\mu}} \mu_4 \mu_1^{\alpha_1 + m_1 - 1} \mu_2^{\alpha_2 + m_2 - 1} \mu_3^{\alpha_3 + m_3 - 1} \mu_4^{\alpha_4 + m_4 - 1} \mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(130)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \int_{\boldsymbol{\mu}} \mu_1^{\alpha_1 + m_1 - 1} \mu_2^{\alpha_2 + m_2 - 1} \mu_3^{\alpha_3 + m_3 - 1} \mu_4^{\alpha_4 + m_4} \mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(130)}{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4)} \frac{\Gamma(\alpha_1 + m_1) \dots \Gamma(\alpha_4 + m_4 + 1)}{\Gamma(130 + 1)} \mathrm{d}\boldsymbol{\mu} \\
&= \frac{\alpha_4 + m_4}{130} = \frac{6}{130}
\end{aligned} \tag{5}$$

2 Exercise 3

Part 1

1. Since the beacon's light cannot reach the coast if it is facing opposite the coast, the support of the pdf for the observations in term of the angle is the semicircle $[-\pi/2, \pi/2]$. As we don't have any further information to suppose any angle to be more likely than the rest, we assume a uniform distribution over the total angle π.

2. By performing a change of variable, we can express θ_k in terms of the position x_k as

$$\theta_k = \tan^{-1}(\frac{x_k - \alpha}{\beta}).$$

We also have to ensure that

$$p(\theta|\alpha, \beta)\mathrm{d}\theta = p(x|\alpha, \beta)\mathrm{d}x,$$

that is

$$p(x|\alpha, \beta) = p(\theta|\alpha, \beta) \left|\frac{\mathrm{d}\theta}{\mathrm{d}x}\right| \tag{6}$$

We thus compute the Jacobian:

$$\begin{aligned}
\left|\frac{\mathrm{d}\theta}{\mathrm{d}x}\right| &= \left|\frac{\mathrm{d}}{\mathrm{d}x}\tan^{-1}(\frac{x-\alpha}{\beta})\right| \\
&= \frac{1}{1+\frac{(x-\alpha)^2}{\beta^2}}\frac{1}{\beta} \\
&= \frac{1}{\frac{\beta^2+(x-\alpha)^2}{\beta^2}}\frac{1}{\beta} \\
&= \frac{\beta^2}{\beta^2+(x-\alpha)^2}\frac{1}{\beta} \\
&= \frac{\beta}{\beta^2+(x-\alpha)^2}
\end{aligned}$$

By plugging it in in 6 we obtain

$$\begin{aligned}
p(x|\alpha, \beta) &= \frac{1}{\pi}\frac{\beta}{\beta^2+(x-\alpha)^2} \\
&= \frac{\beta}{\pi(\beta^2+(x-\alpha)^2)}
\end{aligned} \tag{7}$$

3. As we know that the position along the coast does not depend on β, we have $p(\alpha|\beta) = p(\alpha)$. Moreover, without any further constraint, we assume this to have a uniform distribution, i.e. the probability of any position for the lighthouse is the same and only depends on the size of the interval we are considering. By computing the log of the posterior we have

$$\begin{aligned}
L &= \ln\left(\frac{p(\mathcal{D}|\alpha, \beta)p(\alpha)}{p(\mathcal{D}|\beta)}\right) \\
&= \ln(p(\mathcal{D}|\alpha, \beta)) + \ln(p(\alpha)) - \ln(p(\mathcal{D}|\beta))
\end{aligned}$$

As a function of α, this is independent of the terms $p(\alpha)$ (see the previous remark) and $p(\mathcal{D}|\beta)$, i.e. they can be considered constants. We then have

$$\begin{aligned} L &= constant + \ln(p(\mathcal{D}|\alpha,\beta)) \\ &= constant + \ln\left(\prod_{k=1}^{n} \frac{\beta}{\pi(\beta^2 + (x-\alpha)^2)}\right) \\ &= constant + \sum_{k=1}^{n} \ln\left(\frac{\beta}{\pi(\beta^2 + (x-1ha)^2)}\right) \\ &= constant + \sum_{k=1}^{n} \ln\frac{\beta}{\pi} - \sum_{k=1}^{n} \ln\left(\beta^2 + (x-\alpha)^2\right) \\ &= constant - \sum_{k=1}^{n} \ln\left(\beta^2 + (x-\alpha)^2\right) \end{aligned} \tag{8}$$

To find the value $\hat{\alpha}$ which maximizes it, we need to compute its derivative w.r.t. α and set it to zero. We first proceed with the former:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha} L &= -\sum_{k=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln\left(\beta^2 + (x-\alpha)^2\right) \\ &= 2\sum_{k=1}^{n} \frac{x-\alpha}{\beta^2 + (x-\alpha)^2} \end{aligned}$$

Setting it to zero finally gives us the required expression

$$\sum_{k=1}^{n} \frac{x-\alpha}{\beta^2 + (x-\alpha)^2} = 0$$

4. By plotting the posterior, we can observe the mode to be around 4 (the mean being ca. 1.02). The difference can be explained by the following reasoning. In the mean we are computing the "center of gravity" of the observations, but this need not correspond with the location of the lighthouse. In fact we know that, albeit the mass of the pdf is concentrated around α, observations farther-away have non-negligible probabilities to occur. For example, an observation at -7.3 km could have been generated by a beacon located at 4 km along the coast with probability $\approx 0.96\%$ (cf. with $\alpha = 1$, where the probability is $\approx 1.74\%$). In this sense, the posterior is dominated by points very close to α because of the square term $(x-\alpha)^2$. This information is lost in the mean, where very large values can act as "attractors" and move the mean away from α.

Part 2

1.

2.

3.

Part 3

1. To find the likelihood for the dataset, we use equation 7, assuming i.i.d. datapoints. We thus have

$$p(\mathcal{D}|\alpha,\beta) = \prod_{k=1}^{n} \frac{\beta}{\pi(\beta^2 + (x-\alpha)^2)} = \left(\frac{\beta}{\pi}\right)^n \prod_{k=1}^{n} \frac{1}{(\beta^2 + (x-\alpha)^2)}$$

2.

3.