Modelling

Bittor Alana [1], Alessandro Scibetta[2] and Stefan Losty[3]

[1] *candidate number 97016*
[2] *candidate number 38352*
[3] *candidate number 21521*

October 16, 2019

1 The Prior

1.1 Theory

Question 1

1. The Gaussian likelihood function is chosen as it is used to encode a normally distributed noise function, as well as encoding the assumption that all observations are conditionally independent.

2. Choosing a spherical co-variance matrix for the likelihood implies that the likelihood has circular probability distribution. This comes as a result of the co-variance matrix being a scalar multiple of the identity matrix, indicating that there is 0 co-variance across each dimension thus each dimension is independent and all dimensions have equal variance. In the non-spherical case, there are two further possibilities; firstly, the co-variance matrix may be still be a diagonal matrix, in this instance there would still be 0 co-variance across each dimension, indicating independence however the likelihood would not have a circular probability distribution as each dimension would have a different variance. In the case where the co-variance matrix is not diagonal, this would show some degree of co-variance across dimensions, indicating a lack of independence.

Question 2

$$p(\mathbf{Y}|f,\mathbf{X}) = p(\mathbf{y}_1,...,\mathbf{y}_n|f,\mathbf{X})$$

We are now going to apply the product rule successively. As we know that $p(A,B|C) = p(A|C)p(B|A,C)$, we can use this recursively and get:

$$p(\mathbf{y}_1,...,\mathbf{y}_n|f,\mathbf{X}) = p(\mathbf{y}_1|f,\mathbf{X})p(\mathbf{y}_2,...,\mathbf{y}_n|f,\mathbf{X},\mathbf{y}_1) =$$

$$= p(\mathbf{y}_1|f,\mathbf{X})p(\mathbf{y}_2|f,\mathbf{X},\mathbf{y}_1)p(\mathbf{y}_3,...,\mathbf{y}_n|f,\mathbf{X},\mathbf{y}_1,\mathbf{y}_2) =$$

$$... = p(\mathbf{y}_1|f,\mathbf{X})\prod_{j=2}^{n} p(\mathbf{y}_j|f,\mathbf{X},\mathbf{y}_1,...,\mathbf{y}_{j-1})$$

1.1.1 Linear Regression

Question 3

$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i,\sigma^2\mathbf{I})$$

Question 4

From Bayes Rule:

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$$

This implies:

$$Posterior \propto Likelihood * Prior$$

As a result of the above relationship, conjugate distributions dictate that the posterior distribution will have the same functional form as the product of the likelihood and prior distributions. Conjugacy therefore, allows one to formulate the posterior distribution by determining the functional form of the product of the likelihood and prior. Finding the posterior distribution in this way is often far easier than mechanically calculating it via Bayes' rule, which will often involve integration when calculating the evidence term.

Question 5

The distance between x and μ in a Gaussian distribution is called the Mahalanobis distance (Δ), and is given by:

$$\Delta = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Therefore, if $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$, we have that $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbf{I}$, and thus we can write this distance as $\frac{1}{\sigma^2}(x - \mu)^T (x - \mu)$, which is the square of the Euclidean distance scaled by $\frac{1}{\sigma^2}$. A spherical co-variance matrix $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$ encodes a Euclidean distance scaled by $\frac{1}{\sigma}$.

Question 6

As shown previously in question 3, we have a Gaussian likelihood function:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i | \mathbf{W}\mathbf{x}_i, \sigma^2 \mathbf{I})$$

and a Gaussian prior over parameters **W**:

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{W}_0, \tau^2 \mathbf{I})$$

We know that Gaussians self conjugate and so the posterior function must also be a Gaussian:

$$\begin{aligned}
p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W}) \\
&\propto -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{W})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X}\mathbf{W}) - \frac{1}{2\tau^2}\mathbf{W}^{\mathbf{T}}\mathbf{W} \\
&= -\frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{Y} + \frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{X}\mathbf{W}) - \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{W})^T(\mathbf{X}\mathbf{W}) - \frac{1}{2\tau^2}\mathbf{W}^{\mathbf{T}}\mathbf{W}
\end{aligned}$$

Using the term that is quadratic in **W**;

$$\begin{aligned}
A &= -\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{W})^T(\mathbf{X}\mathbf{W}) - \frac{1}{2\tau^2}\mathbf{W}^{\mathbf{T}}\mathbf{W} \\
A &= -\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{W}^T(\mathbf{X}\mathbf{W}) - \frac{1}{2\tau^2}\mathbf{W}^{\mathbf{T}}\mathbf{W} \\
A &= -\frac{1}{2}\mathbf{W}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\frac{1}{2\tau^2})\mathbf{W}
\end{aligned}$$

From this the covariance matrix of the posterior can be found:

$$\mathbf{S}^{-1} = (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{2\tau^2})$$

The mean can be found using the term linear in **W**:

$$B = \frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{X}\mathbf{W}) = \frac{1}{\sigma^2}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}$$

Comparing this with the general exponent of a Gaussian and solving for mean $\boldsymbol{\mu}$:

$$\begin{aligned}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{1}{\sigma^2}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}\\ \mathbf{W}^T\mathbf{S}^{-1}\mathbf{M} = \mathbf{W}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\tau^2})\boldsymbol{\mu} &= \frac{1}{\sigma^2}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}\\ \rightarrow (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\tau^2})\boldsymbol{\mu} &= \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{Y}\\ \rightarrow \boldsymbol{\mu} &= \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\tau^2})^{-1}\mathbf{X}^T\mathbf{Y}\end{aligned}$$

Therefore the final expression for the posterior over **W** is given by:

$$p(\mathbf{W}|\mathbf{X},\mathbf{Y}) = \mathcal{N}(\mathbf{W}|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\tau^2})^{-1}\mathbf{X}^T\mathbf{Y},(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\tau^2})^{-1}) \tag{1}$$

1.1.2 Non-parametric Regression

Question 7

If a model (M) of some data (Y) is defined as being a conditional distribution parametrised by some θ from a parameter space $\mathcal{T}$ such that:

$$M = \{P(Y|\theta)|\theta \in \mathcal{T}\} \tag{2}$$

If this parameter space $\mathcal{T}$ is a finite dimensional space, the model is parametric. If $\mathcal{T}$ is an infinite-dimensional space, the model is said to be non-parametric. In a parametric model, the number of parameters remains constant with respect to sample size however, in a non-parametric model, the number of parameters grows with the sample size, this can be thought of as the function gaining degrees of freedom as it sees more data.

Question 8

The prior represents the probability of a function f conditioned by **X** and some parameter θ of the kernel function. This is a Gaussian distribution, with mean 0 and a covariance matrix determined by a kernel function. This kernel function, which is parametrised by θ, gives us the covariance matrix where $\Sigma_{ij} = k(x_i, x_j)$, and it encodes generalisation properties of the GP model. Different choices of the kernel could encode an assumption of what our space of functions roughly looks like. For example, the 'smoothness' or 'wiggliness' that the functions in the space have.

Question 9

The prior must encode all possible functions as if this were not the case, it would be possible that some functions in the probability space would exhibit a probability of 0.

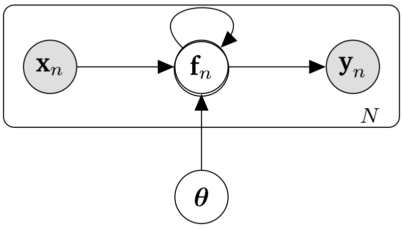
Question 10

We can write a compact expression for the joint distribution if we consider the dependence relationships between the variables.

$$p(\mathbf{Y},\mathbf{X},f,\boldsymbol{\theta}) = p(\mathbf{Y}|f,\mathbf{X},\boldsymbol{\theta})p(f|\mathbf{X},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})p(\mathbf{X}) \overset{1,2,3}{=} p(\mathbf{Y}|f)p(f|\mathbf{X},\boldsymbol{\theta})p(\boldsymbol{\theta})p(\mathbf{X})$$

1. Each $\mathbf{y}_n$ depends solely on the f_n value and the noise.
2. Each f_n value depends on the point x_n, the parameter $\boldsymbol{\theta}$ which affects its distribution, and also on other f_j values, as their distributions have a certain covariance given by the kernel function.
3. $\boldsymbol{\theta}$ and $\mathbf{X}$ have no dependence relations between each other.

This graphical model is the following:



Question 11

The marginalisation of f connects the prior and the data by essentially summing the probability of seeing data $\mathbf{Y}$ given every instance of f over the infinite functional space, weighted by the probability of each f respectively. It is this weighting that encodes the uncertainty in f into the likelihood. It can be seen that $\boldsymbol{\theta}$ remains on the left hand side of the expression after the marginalisation, the probability of any given value of $/y$ is still conditioned by θ despite f being marginalised. This is consequence of the weighting $p(f|\mathbf{X}, \boldsymbol{\theta})$ which accounts for every f conditioned by $\boldsymbol{\theta}$ in the marginalisation.

1.2 Practical

Question 12

We generate some data using the **W** parameters and adding a $\mathcal{N}(\epsilon|0, 0.03)$ noise. Then, as explained in !!!![cite lecturenotes4 here], we get the following likelihood:

$$p(y|\mathbf{W}, \mathbf{X}) = \mathcal{N}(y|\mathbf{W}^T\mathbf{X}, 0.03)$$

Thus, having a Gaussian likelihood, we will now choose a conjugate prior. We take a Gaussian prior, so that the posterior will also result in a Gaussian. We then assume, quite arbitrarily, that our prior is $p(\mathbf{W}) \sim \mathcal{N}((0, 0), I)$.

Figure 1 shows what the prior we chose looks like. As we keep adding points, the mean values of $p(\mathbf{W})$ are going to give us the parameters of the line we are looking for, and the posterior will keep changing shape, as shown in Figure 2.

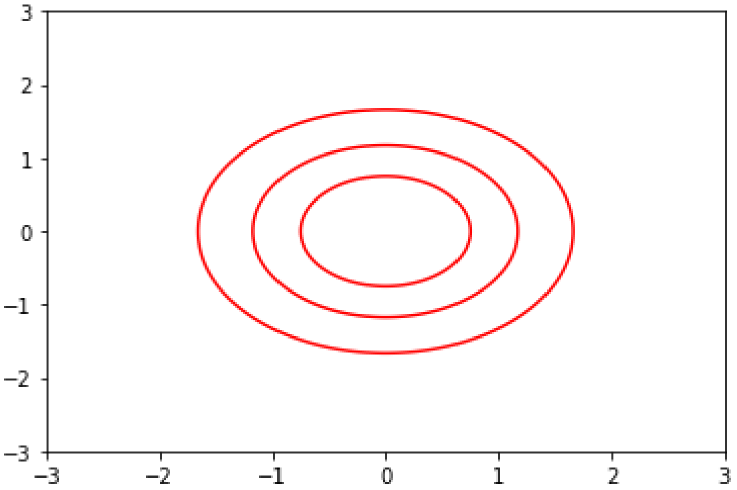


Figure 1: *Visualization of the prior*

We can observe that the more points we add, the more our lines look like the **W** we generated the points with.

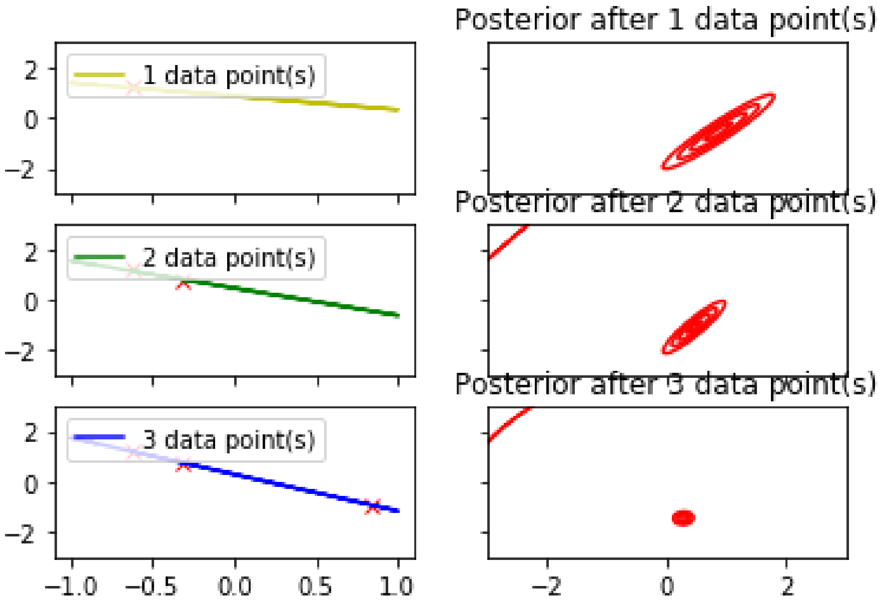


Figure 2: *What our lines and our posterior look like when adding more points*

Question 13

We create a prior over f given $\mathbf{X}$ and $\boldsymbol{\theta}$, and we have $p(f|\mathbf{X},\boldsymbol{\theta}) \sim \mathcal{N}(0, k(\mathbf{X},\mathbf{X}))$. If we assume that $\boldsymbol{\theta} = (\theta_0, \theta_1) = (\sigma^2, l^2) = (1, 5)$, our function space is going to look like Figure 3.

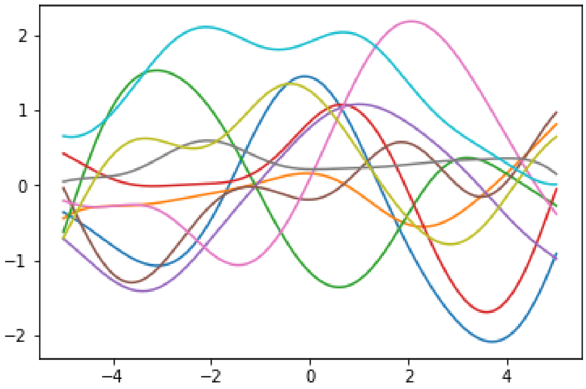


Figure 3: *Sample with lengthscale* $l^2 = 5$

If we introduce different values of l^2, our distribution yields completely different functions. Figure 4 corresponds to $l^2 = 0.03$.

Making the length-scale so small has altered the functions a lot. Let's go the other way around and try a very big length-scale, as in Figure 5, with $l^2 = 9999$:

We can observe that the bigger the length-scale is, the flatter the functions look, 'approaching' linear functions as l^2 approaches infinity. On the other hand, the smaller the l^2 is, the more 'wiggly' the functions look.

Therefore, the length-scale encodes how 'wiggly' or flat we assume the functions to be. If we think they resemble a linear model, we go for higher length-scales. If we assume the f_i to be quite wicked and take very different values over short intervals, we choose low length-scales.

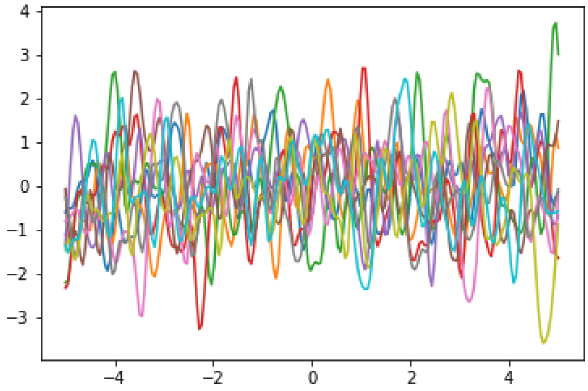


Figure 4: *Sample with length-scale* $l^2 = 0.03$

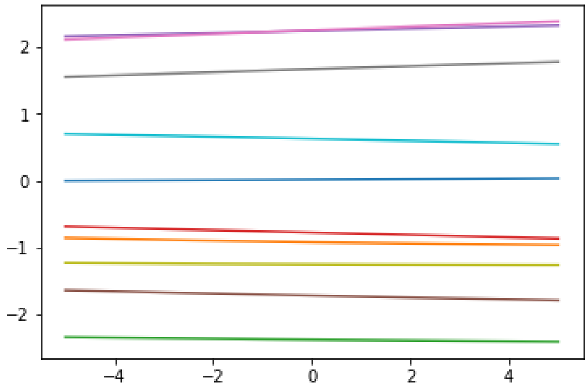


Figure 5: *Sample with length-scale* $l^2 = 9999$

Question 14

As can be seen in Figure 6, the variance of the predictive posterior increases as the distance from the data is increased. This is shown in Figure 6 by the blue shaded region which can be seen to 'swell' in the areas between the data points. Figure 6 also clearly shows regions of high uncertainty outside the range of the data which indicates that the posterior is no longer able to refine the prior, due to the lack of information, and so the system defaults to the prior, shown by the red line, indicating a mean of 0 and the blue region showing a standard deviation equal to 1.

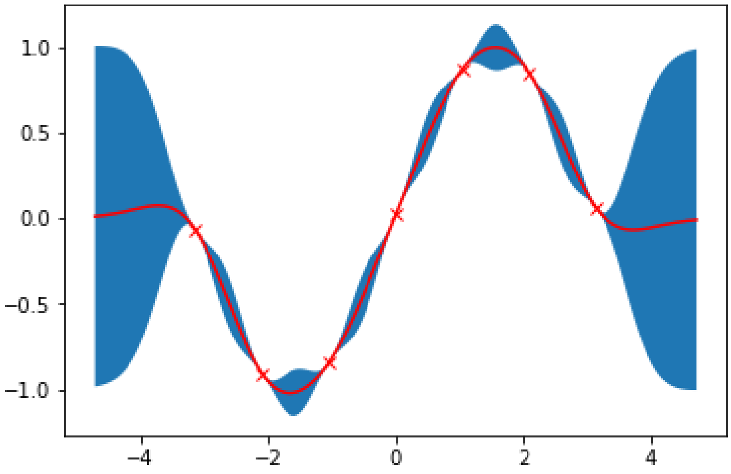


Figure 6: *Data, predictive mean and predictive variance of the posterior from the data*

2 The Posterior

2.1 Theory

Question 15

It is necessary to make assumptions at the beginning of the model so that you are able to formulate your beliefs - without assuming anything it is impossible to form beliefs. Your preference for the model is correlated to the value of the likelihood function. A high likelihood means we have a high preference for the model as the model fits the data closely. This behaviour is desirable however, in order to be able to make more accurate predictions, more data would be needed, particularly outside the range of $[-\pi, \pi]$.

Question 16

The first assumption that we can make is that x follows a Gaussian distribution, with mean 0. Also, the covariance matrix is the identity matrix and so we can assume there is zero covariance with other dimensions, therefore x is fully independent.

Question 17

Instead of computing the whole integral, which could be a tricky procedure, we can try and calculate the parameters that define $p(\mathbf{Y})$.

We have written **Y** as a linear mapping of Gaussians plus another Gaussian, thus we can say for sure it is another Gaussian. Therefore, knowing its mean and covariance will suffice to determine the marginalisation:

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{WX} + \epsilon] = \mathbb{E}[\mathbf{WX}] + \mathbb{E}[\epsilon] = \mathbb{E}[\mathbf{WX}] = 0$$

Bearing in mind that ϵ has mean zero, and that the elements of **X** also have mean zero, and we are taking a linear combination of them, the linearity of the expected value operator gives us that $E[\mathbf{WX}] = 0$ and so the mean of $p(Y)$ is also zero.

$$\mathbb{E}[\mathbf{YY}^T] = \mathbb{E}[(\mathbf{WX} + \epsilon)(\mathbf{WX} + \epsilon)^T] = \mathbb{E}[\mathbf{WXX}^T\mathbf{W}^T] + \mathbb{E}[\epsilon\epsilon^T]$$

We know the second term of the sum is $\sigma^2\mathbf{I}$. The linearity of the expected value operator allows us to get the **W** and $\mathbf{W}^T$ out, and we know $\mathbb{E}[\mathbf{XX}^T] = \mathbf{I}$, as the **x**'s covariance is **I**, thus we get the expression

$$\mathbb{E}[\mathbf{YY}^T] = \mathbf{WW}^T + \sigma^2\mathbf{I}$$

Therefore, $p(\mathbf{Y}) \sim \mathcal{N}(0, \mathbf{WW}^T + \sigma^2\mathbf{I})$

2.1.1 Learning

Question 18

Maximum Likelihood Estimation and Maximum A Posteriori are both methods for estimating some variable in the setting of probability distributions or graphical models. They are similar as they compute a single estimate instead of a full distribution. The ML estimate is the mode of the outcome of the likelihood function, however it tends to over-fit the data. As a result, the variance of the parameter estimates is high. We can reduce this variance by introducing a degree of bias to the estimate, this is called regularisation. In MAP, regularisation is achieved by assuming that the parameters are drawn from a random process. MAP is only possible if we are given the prior or able to assume it, whereas MLE takes no consideration of the prior.

MAP and MLE will give the same result if either there is an infinite amount of data or the prior belief is uniform - that is to say it is infinitely weak. These two conditions are linked, as the more data you have the weaker the prior becomes. You could say that MLE is a special case of MAP where the prior is uniform.

W is a mute variable in the denominator. Therefore, the **W** that we choose to try and maximise the fraction does not affect the denominator, as its value only depends on the value of the integration over the whole space where **W** lies. Thus, we could say the denominator is 'constant' with respect to **W**, and hereby the **W** that maximises the fraction corresponds with the **W** that maximises the numerator.

2.1.2 Practical Optimisation

Question 19

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{K}^{-1}\mathbf{y}_n) + \frac{N}{2}ln|\mathbf{K}| + \frac{N}{2}ln(2\pi) \tag{3}$$

2.1.3 Non-parametric

2.1.4 Question 20

The marginalisation of f is much easier to do than marginalising out X as the relationship between output Y and f is much closer than the relationship between X and Y. This is due to the fact that X must first go through f before going to Y as can be seen in the graphical model in question 11. This is different to the linear regression case, where marginalising out X would not be such a task, however in this case, there are many instances of f which makes marginalising out X very difficult.

3 Evidence

4 Final Thoughts

Question 30

We felt this coursework helped to reinforce our understanding of the lecture material and showed us what machine learning is in a practical sense, instead of only learning about the theory. It also was a good exercise in Python and our general understanding of probabilistic techniques.

However, we believe there were certain things that made the coursework more necessary than it needed to be. We don't have complaints about the actual teaching per se, but we all would have appreciated a clearer, more compact and more structured reference. Often we felt it was necessary to watch lectures on RePlay which we had already attended because the lecture slides on the repo were quite brief. It was frustrating having to dig for information from so many different sources, whether it be Bishop, lecture slides, the summary document, Bernoulli trial etc - it wasn't obvious where to find things. We found the lab sessions helpful but probably needed more of them as we attended nearly all of them but still struggled to complete before the deadline. We found the notation often confusing and inconsistent but understand that this was perhaps down to conventions in Machine Learning rather than any flaw in the unit. We are aware that this coursework covers most of the unit but it was quite intimidating to be launched into such a big coursework in the first few weeks. It would have been nice to build up our understanding in steps. Nonetheless the TAs were all helpful and knowledgeable and we don't doubt the enthusiasm with which the unit is taught.