Models

Mark Ergus Nicholl[1] and Faizaan Sakib[2]

University of Bristol
[1] *35358*
[2] *38655*

December 3, 2018

To truly understand data, we must understand the different aspects of building models of data and be able to choose the correct model for the data. Through the application of supervised and unsupervised learning, we can learn how to ensure that data can be fully represented in order to make the most out of it.

1 The Prior

1.1 Theory

Q1.1 A Gaussian likelihood would allow the assumption that it can be made up of the sum or average of multiple random variables that are IID (independent and identically random variables). As long as the random variables are mutually independent and each of them have the same probability distribution, one can use statistical inference procedures that are based on normal distributions, even if data is sampled from a population that is not normal. This is due to the extremely helpful 'Central Limit Theory'.

$$\sqrt{\frac{n}{\theta^2}}((\frac{1}{n}\sum_{i=1}^{n} x_i) - \mu) \xrightarrow{D} N(0,1) \tag{1}$$

As shown in (1), if n, the sample size, is very large, the distribution will be looking like a normal distribution with a mean of 0. CLT means that the sample mean will follow a Normal/Gaussian distribution for large sample sizes, regardless of the distribution from which it is sampled from. Due to the fact that it is possible to work with potentially unknown distributions and still come out with a Gaussian distribution, makes it a good choice for the likelihood function. The idea that noise is often assumed to be a Gaussian distribution arises from CLT, in the sense that the noise that is produced can be seen as an accumulation of many samples of noise that then smooth out into a normal distribution.

Q1.2 Choosing a spherical co-variance matrix, more specifically $\sigma^2 I$ means that we assume our data points are independent of each other. Alternatively, it can be said that an n-dimensional Gaussian with a mean μ and a diagonal co-variance matrix $\sum$ can be treated as a collection of n independent Gaussian random variables. If we do not assume independence, then a non-spherical co-variance matrix has to be used instead.

Q2 As long as each output point is independent given the input **X** and mapping **f**, the likelihood of the data can be represented as such:

$$p(\mathbf{Y}|f,\mathbf{X}) = \prod_{i}^{N} p(y_i|f,x_i) \tag{2}$$

However, if the data points are **not** assumed to be independent, all of the preceding mapped points must be taken into consideration. Without the variables being mutually independent, they would no longer be IID and fall under the Central Limit Theorem, so the likelihood would not be Gaussian. The likelihood would therefore look like this:

$$p(\mathbf{Y}|f,\mathbf{X}) = p(y_1,...,y_N|f,\mathbf{X}) \tag{3}$$

$$p(\mathbf{Y}|f,\mathbf{X}) = p(y_N|y_1,...,y_N,f,\mathbf{X})p(y_{N-1}|y_1,...,y_{N-2},f,\mathbf{X})p(y_1|f,\mathbf{X}) \tag{4}$$

1.1.1 Linear Regression

Q3 Given the formula, it can be rewritten as:

$$y - Wx = \epsilon \tag{5}$$

For simplicity, we now abbreviate the inverse variance, or precision, as β . As we assume that the data we get is corrupted with Gaussian noise, we can substitute in a Gaussian formula with the assumed mean and variance:

$$y - Wx = N(\epsilon|0, \beta^{-1}I) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-1/2(\epsilon - 0)\beta(\epsilon - 0)\right) \tag{6}$$

Substituting (4) into (5) :

$$N(y - Wx|0, \beta^{-1}I) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-1/2(y - Wx)\beta(y - Wx)\right) \tag{7}$$

The exponent portion of the equation resembles a Gaussian distribution of y. Hence:

$$N(y - Wx|0, \beta^{-1}I) = N(y|Wx, \beta^{-1}I) \tag{8}$$

$$p(y|W, x) = N(y|Wx, \beta^{-1}I) \tag{9}$$

Assuming independence, we can sum all of these values to formulate the likelihood of the data:

$$p(Y|W, X) = \prod N(Y|WX, \beta^{-1}I) = \prod N(Y|WX, \sigma^2 I) \tag{10}$$

Q4 A Gaussian likelihood would allow the assumption that there is a conjugate prior to the likelihood due to the fact that the prior and posterior are both Gaussian. This puts them in the same functional family. This is due to the fact that the posterior distribution is proportional to the product of the prior and the likelihood function.

$$p(Y|X) \propto p(X|Y)P(Y) \tag{11}$$

Since the form of the posterior is known, the evidence or the denominator in Bayes' Rule, which may be a complicated integral does not have to be calculated and instead, just the parameters can be identified.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\int p(X|Y)p(Y)} \tag{12}$$

Q5 The prior distribution allows us to visualise the probability of an event with certain parametric choices. If the prior

$$p(W) = N(W_0, \tau^2 I) \tag{13}$$

is a spherical Gaussian, this will be due to its spherical co-variance. The implications this has is the way that the value of the Gaussian falls off depending on the change of its parameters. Due to its spherical properties, if the Gaussian between two of its parameters is observed, for example X and Y, it will be evident that the same decrease in either X or Y will cause an identical decrease in the probability of such a parameter choice.

This would be different to a diagonal Gaussian model where the relationship between parameters is less regular than a spherical Gaussian. The contours of the diagonal Gaussian in 2D would appear squashed in either direction, showing that a change in X can cause a big shift in the probability while a change in Y causes a much smaller change in probability, and vice-versa.

With a spherical Gaussian distribution, it makes sense to use the Euclidean distance between the data point and the mean, since the parameters weigh equally in terms of distance. However, it is important to note that for a Gaussian that is not spherical, this is not the case. A point that is closer to the mean, or the centre of the Gaussian will not necessarily be of a high probability density than a point that is comparatively further away. This is because the ellipses of a Gaussian will be skewed if it is not spherical, meaning a point has to be in a contour more inwards than the other point to have a higher probability density. This is where the Mahalanobis distance comes in. It adds semantics to the distance of data points within Gaussian distributions, taking into account different variations for each directions, along with the co-variance between variables [1]. The squared Mahalanobis distance is given by

$$(x - \mu)^T \Sigma^{-1} (x - \mu) \tag{14}$$

where if Σ is the identity matrix, the Mahalanobis distance is reduced to a Euclidean distance between the data point and the mean.

Q6 Due to the property of conjugacy (10) and the knowledge that the prior is a Gaussian distribution, we can infer that the posterior is a Gaussian distribution as well :

$$p(W|X, Y) = N(W|m_N, S_N) \tag{15}$$

To find out the values of m_N and S_N, the posterior is written as proportional to the product of the prior and the likelihood function.

$$p(W|X, Y) \propto p(Y|X, W)P(W) \tag{16}$$

Using the prior (9) and the likelihood (12) chosen from earlier we obtain

$$p(w)\prod_{i=1}^{n} p(y_i|x_i,w) \tag{17}$$

which is proportional to the exponential of the expression.

$$exp(-\frac{1}{2\tau^2}(w-w_o)^2)\prod_{i=1}^{n} exp(-\frac{1}{2\sigma^2}(y_i-x_iw)^2) \tag{18}$$

$$exp(-\frac{1}{2\tau^2}(w-w_o)^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-x_iw)^2) \tag{19}$$

The exponential above (18) is actually proportional to a distribution over *w*. This expression can be multiplied together to get terms that can be collected to get a constant along with *w* and w^2 which would imply it being a quadratic and also therefore a Gaussian.

$$(-\frac{1}{2\tau^2}-\frac{n\sum x_i}{2\sigma^2})w^2+(\frac{w_0}{\tau^2}+\frac{\sum x_iy_i}{\sigma^2})w+c \tag{20}$$

Now that the like terms of constant, mixed and quadratic have been collected, the square must be completed in order to find the parameters by having an expression of the Gaussian form:

$$-\frac{1}{2\sigma^2}(x-\mu)^2 \tag{21}$$

By completing the square on (19) we get:

$$-\frac{1}{2}(\frac{1}{\tau^2}+\frac{n\sum x_i}{\sigma^2})(w^2-2(\frac{\frac{w_0}{\tau^2}+\frac{\sum y_ix_i}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n\sum x_i}{\sigma^2}})w+m^2)+c \tag{22}$$

$$-\frac{1}{2}(\frac{1}{\tau^2}+\frac{n\sum x_i}{\sigma^2})(w-m_N)^2+c \tag{23}$$

$$-\frac{1}{2S_N}(w-m_N)^2+c \tag{24}$$

where m_N and S_N is the mean and the variance of the posterior after the model has integrate the prior beliefs with the data having seen N data points.

$$m_N=S_N(\frac{w_0}{\tau^2}+\frac{\sum y_ix_i}{\sigma^2}) \tag{25}$$

$$S_N=\frac{1}{\frac{1}{\tau^2}+\frac{n\sum x_i}{\sigma^2}} \tag{26}$$

As a result, we can say that the posterior distribution can be modelled as a Gaussian with mean m_N and variance S_N.

$$p(W|X,Y)\propto exp(-\frac{1}{2S_N}(w-m_n)^2) \tag{27}$$

$$W|X,Y\sim N(m_N,S_N) \tag{28}$$

As explained through (10) and due to the conjugate prior (9) chosen, it can be assumed that the posterior distribution is of the same functional family, and as such can be treated as a Gaussian distribution. This is very helpful in obtaining the posterior from the prior as the form is already known. This reduces the complexity of the problem as it allows us to avoid computing the evidence and simply identify the parameters as done above.

1.1.2 Non-Parametric Regression

Q7 There are a few key differences between parametric models and non-parametric models.

Parametric models have a fixed number of parameters in which it captures all of the information available from the data. They also require assumptions on the underlying function to be made. With a parametric model, the only things that are needed in order to predict a new value from the current model state is the parameters. For example, with linear regression, only the coefficient and intercept is needed as the two parameters which will allow you to predict a new value. Parametric models usually work well with less complex problems, and can model problems with small data sets better than non-parametric models can.

Meanwhile, non-parametric models do not have an upper bound on the number of parameters used. This allows it to encompass the more subtle parts of the data without requiring assumptions, making the model more flexible and better at representing more complex problems, as long as there is a large enough set of data. This can be achieved using a Gaussian mixture model, where several Gaussian distributions can form to bring about a better representation of the data. However, non-parametric models do have a higher risk of overfitting the training data.

Q8 This prior represents a prior assumption over the space of functions, where the co-variance function k(.,.) encodes the behaviour of the function. The co-variance function is also known as the kernel of the Gaussian Process. As the kernel encodes the behaviour of the function, we can now establish a suitable kernel to act

as the prior of the data. To visualise this structure, we imagine that there are infinitely many 'slices' in our distribution space, and all slices are jointly Gaussian. Each 'slice' now is a Gaussian, and the Gaussian depends on our kernel which models the co-variance between an input data point and the Gaussian slice. Now, if we look at the infinitely many slices of Gaussian in our distribution space together, we can form a 'tube' of area where we predict a new data point to be. For every new data point, we can then narrow or widen the tube depending on our chosen co-variance. This gives us a belief of what we assume our model to look like, which is exactly what we need for a prior.

Q9 This prior encodes all possible functions. As the Gaussian Process is an infinite collection of random variables where any subset is jointly Gaussian, that means that there are infinitely many parameters. With infinitely many parameters, a line equation can take infinitely many forms. Hence, any possible function can be encoded by a Gaussian Process.

Q10 Below is the joint distribution of the full model where the x that is mapped by f. This allows us to contain the uncertainty with the mapping f.

$$p(Y, X, f, \theta) = p(y|f)p(f|x, \theta)p(\theta) \tag{29}$$

The assumptions made in the graphical model below are:

- f is conditional on the latent variable x
- y is the observed variable and is in turn conditional on f
- f is also conditioned by the hyper-parameter θ
- Any noise ϵ is assumed to be a Gaussian as $N \sim (0, \sigma^2 I)$

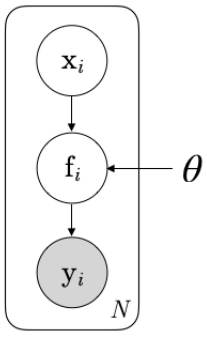

Figure 1: *A Graphical model of the non-parametric regression model*

Q11 An integral can be thought of as an expectation of the data. Usually, the expected value of a random variable y can be calculated as:

$$E[y] = \int yp(y)dy \tag{30}$$

This can be thought of as taking the sum of all the data with their associated probability. In the same way, to marginalise out the variable f, we calculate the expected distribution. In other words, an intuitive idea of this is to say "now we are summing up the probability conditional on f to each and every possible value of f, thus marginalising it." Hence,

$$\int p(Y|f)p(f|X, \theta)df \tag{31}$$

can be thought of as finding the expected distribution, given X and θ values thus connecting the prior and the data.

There are 2 sources of Gaussian uncertainty prior to the calculation of this integral, namely that associated with $f(x)$ and that associated with ϵ. By calculating the integral, we are taking into account all possible values of f, hence filtering it's uncertainty.

As the θ is left on the left hand side, this means that the output (Y) is still conditional on θ. The principle thing to do would be to specify our beliefs over them and try to integrate them out, as we have done with the f parameter. However, this can be hard to do, as it is in this case. Instead, we can optimise the θ parameter by finding the type-II maximum likelihood such that it maximises the marginalised likelihood.

Q12

12.1 Firstly, the prior distribution is visualised over W as a zero mean isotropic Gaussian, where a mean of 0 is used along with a diagonal co-variance $2.I$. It is visualised using contour lines to show the probability distribution function of the Gaussian. As seen below, it can be seen as a spherical Gaussian.

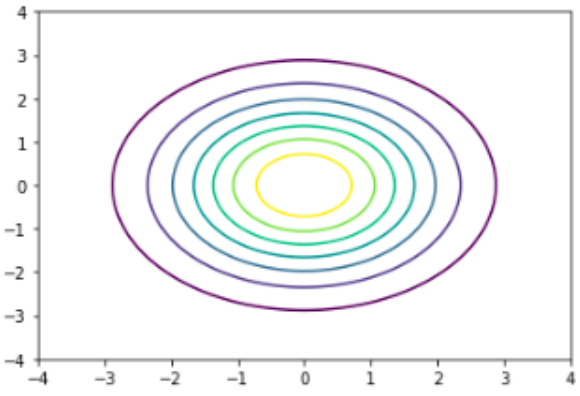

Figure 2: *The visualisation of the prior distribution over W*

12.2 As soon as one data point is observed, as seen as the red point, the posterior distribution can be seen to reduce its size and variance by quite a lot. However, this is only based on a single data point and is likely to be over-fitting the data at this stage.

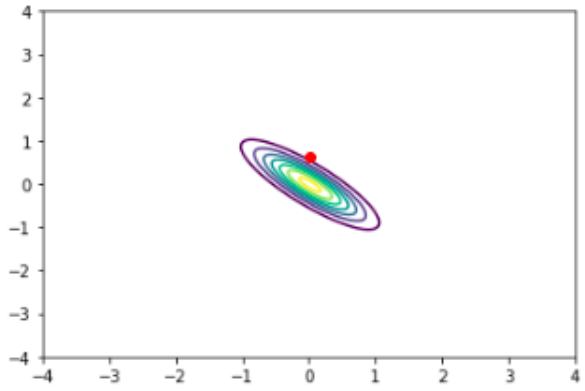

Figure 3: *The visualisation of the prior distribution over W after observing a single data point*

12.3 The posterior has now been sampled from, in particular the multivariate Gaussian of the posterior. The samples are then used as *w* values to plot different functions based on the posterior.

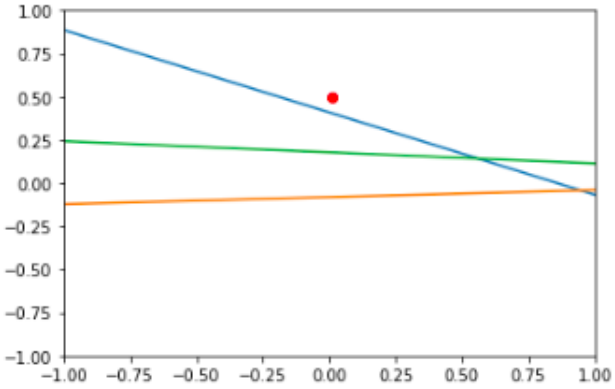

Figure 4: *The visualisation of three different functions sampled from the posterior*

It is important to note that even with just one data point, the sampled functions go through fairly close to the observed data point.

12.4 - 12.5 Once more data points are added to the model, the posterior distribution starts to hone in on the parameters that it is trying to predict. With the addition of each point, a line is becoming more and more apparent, and so do the possible parameters to produce a line of such nature.

When seeing even more data points, the posterior distribution becomes much more confident in the parameters. However, the posterior does reach a cut-off point where it has already observed enough data to make a good prediction and that every other point that is observed will help minimally.

This is the desirable behaviour expected from this linear regression as it manages to retrieve the parameters back even by simply just looking at the data without any intuition of what they were originally.

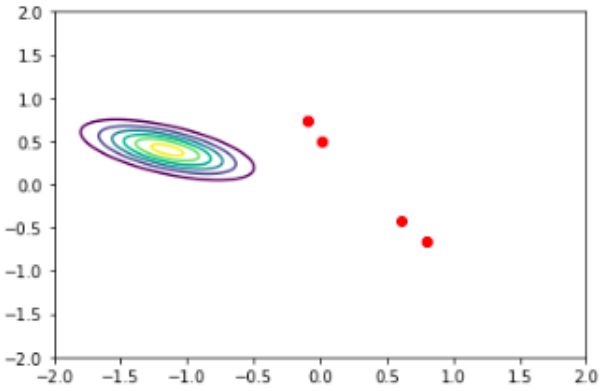

Figure 5: *Posterior distribution after 4 observed data points*

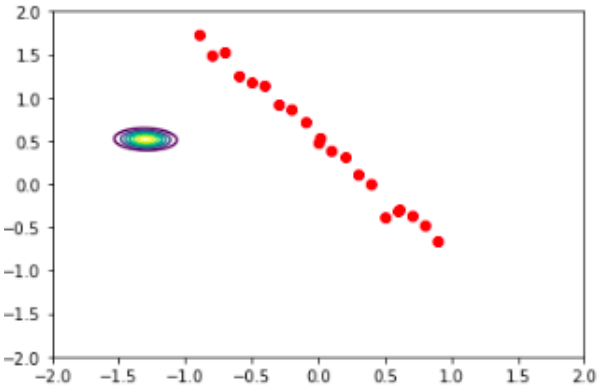

Figure 6: *Posterior distribution after 20 observed data points*

12.6 The expression of the posterior is not only dependent on the prior but also heavily dependent on the basis function $\phi(X)$ and the observed values t.

$$m_N = (S_0^{-1} + \beta\phi(X)^T\phi(X))^{-1}(S_0^{-1}m_0 + \beta\phi(X)^T t) \tag{32}$$

$$S_N = (S_0^{-1} + \beta\phi(X)^T\phi(X))^{-1} \tag{33}$$

The expressions above represent the mean and the co-variance after having seen N data points. With every data point added, both the basis function $\phi(X)$ and t encodes more information. This is an iterative process and with each data point, the posterior creates a new distribution based on the data points before it. This means the nature of the posterior getting closer to

the expected parameters after each data point makes sense, since if the data does actually have a meaningful distribution, then with each new data point, the posterior will understand the data better. It is interesting to see how the variance in the x-axis is much greater than the y-axis, as can be seen in the spread of the contours and the co-variance itself. This shows that the posterior has less confidence in w_0 = -1.3 than it does in w_1= 0.5. This could be due to the fact that w_1 is simply a constant that multiplies with 1 from the basis function, making it easier to predict.

Q13

13.1 At this stage, we create a GP-prior with a squared exponential co-variance function. To begin with this, we start by setting the parameters for the co-variance function. The 2 parameters we need to choose are the length-scale, l, and the variance, σ^2. For the initial co-variance function, we set the variance and length-scale to 1. The following is the result we get when we plot the full co-variance matrix, which represents the co-variance between respective points in x and x'.

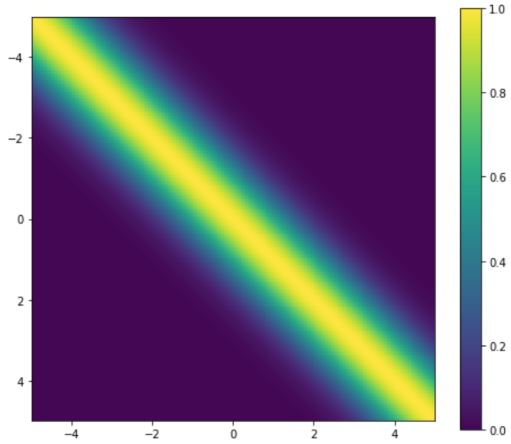

Figure 7: *squared exponential co-variance matrix*

As we can see from the figure, the value of the co-variance right down the diagonal is 1.0. This makes sense as these values are co-variances of the same data points against each other.

13.2 Next, we can sample from the prior with this co-variance matrix and visualise the samples. For this purpose, we take 10 samples from the prior which has a co-variance function with length-scale 1.

13.3 We will now alter the length-scale and repeat the sampling process. We set the length-scale at 0.5 and 5, then sample again.

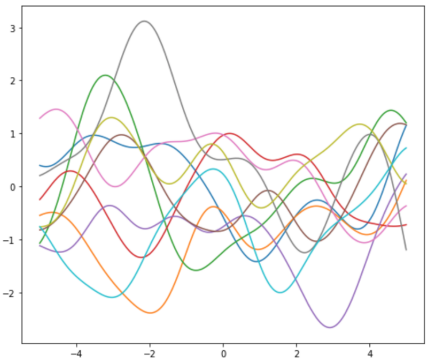

Figure 8: *10 samples from the prior, squared exponential co-variance with length-scale 1*

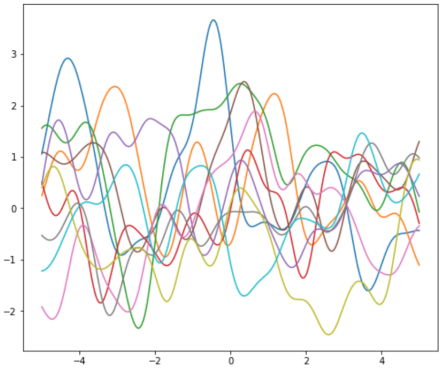

Figure 9: *10 samples from the prior, squared exponential co-variance with length-scale 0.5*

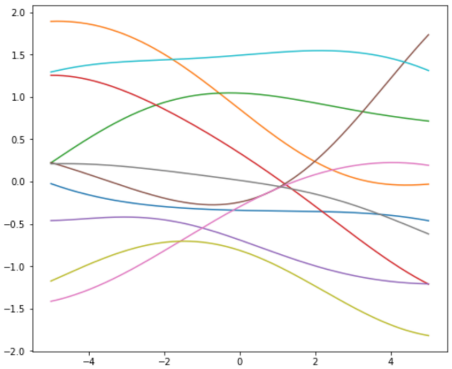

Figure 10: *10 samples from the prior, squared exponential co-variance with length-scale 5*

The incredibly interesting part about this is that we have successfully sampled a data with high dimensionality, in this case data over a 250 dimensional space, and displayed them in 2 dimensions.

13.4 As we alter the length-scale, we can observe that it changes the 'smoothness' of the samples in the prior. The larger the length-scale, the 'smoother' the samples. Or in other words, a small length-scale allows functions to change their values quickly.

13.5 The length-scale parameter encodes our assumption of the drop-off in co-variance between 2 data points.

Q14

14.1 Knowing that all instantiations are jointly Gaussian, we can say that:

$$\begin{bmatrix} f \\ f* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} k(X,X) & k(X,x*) \\ k(x*,X) & k(x*,x*) \end{bmatrix} \right) \tag{34}$$

extracting the conditional Gaussian, we now know that:

$$p(f*|x*,X,f,\theta) = N\Big(k(x*,X)^{\mathrm{T}}k(X,X)^{-1}f, \\ k(x*,x*) - k(x*,X)^{\mathrm{T}}k(X,X)^{-1}k(X,x*)\Big) \tag{35}$$

14.2 For this section, we take 20 samples from our posterior and plot them along with the locations of the training data to visualise our posterior.

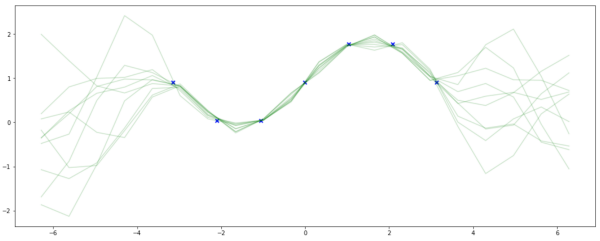

Figure 11

14.3 We can also plot our data with the predictive mean and predictive variance of the posterior. From

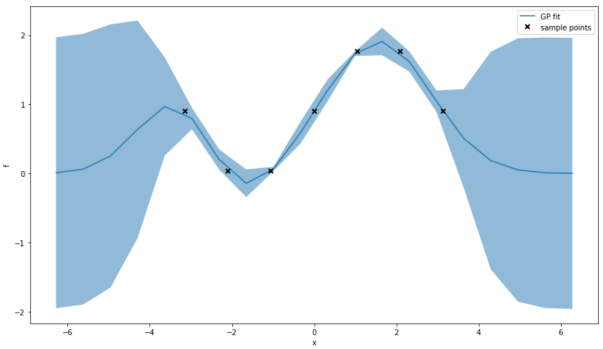

Figure 12

the figure, we can see that the samples from the posterior has narrowed down to those passing through the training data compared to the prior samples. We now have a tight margin for samples. This would not be desirable for data with noise, as little deviation will put them out of the variance allowance. To account for noisy data, we need to calculate the posterior as such :

$$\begin{bmatrix} f \\ f* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} k(X,X)+\sigma^2 I & k(X,x*) \\ k(x*,X) & k(x*,x*) \end{bmatrix} \right) \tag{36}$$

Now when we plot the data along with the predictive mean and predictive variance, we get more allowance for noisy data.

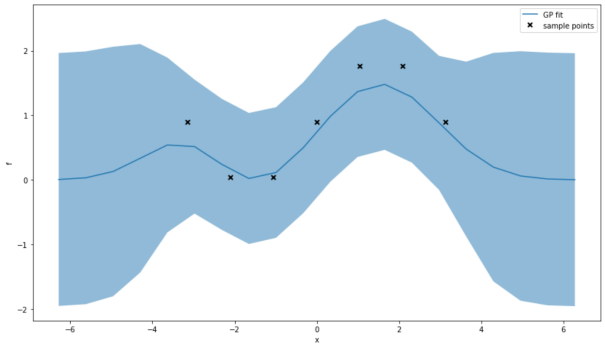

Figure 13

2 Posterior

Q15 While going through the machine learning course, we often come across 3 similar but semantically different words, namely assumptions, belief and preference. Assumptions are often used in cases where we assume an idea with little to no knowledge of it, such as in cases of assuming the prior to have a zero mean isotropic Gaussian. While there is a chance for this to be true, there is little basis for us to know if it is before dissecting the data. We usually stick with broad-based assumptions such as a high variance, in order to not form a prior strong enough to negatively affect results. With a large data set, we can update our assumptions as we gain more knowledge of the data.

Beliefs on the other hand usually stands on a certain basis of knowledge or intuition derived from data we have seen before in a certain field. For example, when modelling a coin toss, we can encode our belief that the probability of each outcome is usually 0.5 provided it is a fair coin as we make this guess from our experiences with coin tosses. In the same way, if we are trying to model an area that we are not an expert in, such as trends in obesity, we can then seek an expert's opinion, in this case a medically qualified doctor, to encode their belief of what they know into our prior.

A preference differs in a sense that we choose to encode them for simplicity sake, without sacrificing the variability of our work. By choosing a certain mean, variance or even distribution to work with to simplify the mathematics required to obtain our results, we can classify that as a preference.

In most cases, there is no clear line between these 3 words. There are overlaps in most cases such as when

choosing a prior distributed with zero mean isotropic Gaussian. There is an assumption aspect of it where we know it is possible for this prior to be true, but we cannot verify it as of that moment. we keep the assumption broad in order to update it after gaining knowledge. There is also a preference aspect of it where the chosen prior is desirable to work with due to the properties of conjugacy. This overlap often draws criticism to Bayesian probabilities as the prior is such a 'subjective' and flexible choice, but when used correctly with people knowledgeable in said field, can filter uncharacteristic data from tainting our learning of the model. Simply put, without a prior belief, we would believe the world is flat if 8 videos out of 10 has 'evidence' of it.

Q16 With this prior, we have assumed or rather in this case prefer that the probability of x is not bias in any dimension. In other words, we prefer that the x variables have an equal probability to take any value scattered around the mean, 0.

Q17 To perform marginalisation on **X**, we can assume the model to be a zero mean Gaussian with noise where it can be given in the form

$$p(Y|W) = \mathcal{N}(0, C) \tag{37}$$

where the mean is 0 and C is the co-variance. Due to the property of conjugacy as discussed earlier, the two Gaussians can be multiplied to get another Gaussian

$$p(Y|W) = \int \mathcal{N}(Y|WX, \sigma^2 I)\mathcal{N}(0, \sigma^2 I) \tag{38}$$

The exponential of this expression can then be taken and then multiplied

$$exp(-\frac{1}{2\sigma^2 I}(x-0)^2)\prod_{i=1}^{n} exp(-\frac{1}{2\sigma^2 I}(y_i - x_i w)^2) \tag{39}$$

From this we can separate the quadratic, mixed and constant and then complete the square with respect to *x*. This would give us the mean and co-variance to the distribution with **X** marginalised out. This would give us a mean of 0 and a co-variance of $\mathbf{WW}^T + \sigma^2 I$. Thus the marginal distribution would be

$$p(Y|W) = \mathcal{N}(0, \mathbf{WW}^T + \sigma^2 I) \tag{40}$$

Q18 Maximum Likelihood involves finding the value that maximises the likelihood function. This means choosing the value of **W** for which the probability of the observed data set is maximised. By maximising the likelihood, we are also minimising the error of the model. However, it is important to note that Maximum Likelihood does not incorporate prior knowledge. This can lead to over fitting in the model and it can fall hostage to random variations in the data that is introduced by noise or uncertainty. For example, an ML estimation based on 4 coin flips, where all flips land on heads would suggest that the coin would forever produce heads for all proceeding flips. Meanwhile, MAP estimation does incorporate prior knowledge. Instead it maximises the posterior distribution, taking into account the likelihood and also the prior. However, MAP estimation can end up being the same as ML if the prior that is used is a non-informative uniform prior. This is because the prior can be taken as a constant which does not increase the posterior distribution like a change in parameter would.

When we observe more data, since ML estimation does not incorporate any prior, it will simply start on the data set as a whole and maximise the likelihood on it. On the other hand, MAP estimation will take the prior, or the previous posterior, from the current data set as the foundation to build upon to maximise the new posterior distribution. However, it is important to note, that for large data sets, ML and MAP estimation will converge to the same value. This is because with more data, the information that would be encoded within the prior, now shifts to the larger amount of data that is available. Since the only difference between MAP and ML is the use of the prior, this would mean they would both end up giving the same result.

Since the posterior is proportional to the product of the likelihood and the prior, the denominator of the expression on the right does not affect **W**. Therefore, it can be treated as a constant, making both expressions equal.

Q19 Taking the negative log of $p(Y|W)$, we get :

$$\begin{aligned} -log(p(Y|W)) = \frac{ND}{2}log(2\pi) + \frac{N}{2}log|WW^T + \sigma^2 I| \\ + \frac{1}{2}\sum_{n=1}^{N}(y_n - \mu)^T(WW^T + \sigma^2 I)^{-1}(y_n - \mu) \end{aligned} \tag{41}$$

Now, setting the mean of the data to zero, we can remove the sum, and we get :

$$L(W) = \frac{ND}{2}log(2\pi) + \frac{N}{2}log|C| + \frac{N}{2}tr(Y(C)^{-1}Y^T) \tag{42}$$

where $Y = (y_n - \mu)$ and $C = WW^T + \sigma^2 I$ and tr is the trace of the matrix. After integrating, we get the gradient minimising function given by:

$$\frac{\delta L}{\delta W} = \frac{N}{2} tr\Big(C^{-1}\frac{\delta C}{\delta W_{ij}}\Big) + \frac{N}{2} tr\Big(Y^T Y(-C^{-1}\frac{\delta C}{\delta W_{ij}}C^{-1})\Big) \tag{43}$$

where $\frac{\delta C}{\delta W_{ij}} = WJ_{ji} + J_{ij}W^T$

Q20 f is a function that relates **X** and **Y** together. This means f is simply a mapping and can be marginalised out because the information that is required to model the data is held within **X** and **Y**. Before marginalisation, **Y** is conditional on f, and f is conditional on **X**. If **X** is removed, it breaks the whole model as f will no longer be able to push through data from **X**. However, if we marginalise the f, the middle man is essentially removed, and **Y** can be directly conditional to **X**, also encoding the parameter θ.

This will eventually allow us to predict the probability of **Y** from **X**, where the mapping f is not really important for our case.

Q21 To start off with, the objective function is minimised to obtain the maximum likelihood for **W** where

$$\hat{\mathbf{W}} = argmin_{\mathbf{W}}\mathcal{L}(\mathbf{W}) \tag{44}$$

The value and the gradient of the objective is optimised at each value of **X**. Once we are just left with the linear mapping of **X**, it must be removed using matrix algebra so we can visualise $\hat{\mathbf{X}}$.

$$\begin{aligned}
Y &= f_{lin}(x') \\
Y &= \hat{X}\hat{W^T} \\
Y\hat{W} &= \hat{X}W^T\hat{W} \\
Y\hat{W}\hat{W^{-1}} &= \hat{X}W^T\hat{W}\hat{W^{-1}} \\
Y\hat{W}\hat{W^{-1}} &= \hat{X}\hat{W^T}I \\
Y\hat{W}\hat{W^{-1}}\hat{W^{T-1}} &= \hat{X}\hat{W^T}\hat{W^{T-1}} \\
Y\hat{W}(\hat{W}\hat{W^T})^{-1} &= \hat{X}I \\
\hat{X} &= Y\hat{W}(\hat{W}\hat{W^T})^{-1}
\end{aligned}$$

The plot that has been created is a 2D visualisation of the latent space of **X**, where the 10 dimensional data set has been reduced to be represented in 2 dimensions where the variance in the data is the highest. This was achieved by minimising the objective function which would mean maximising the co-variance in (42), making $\frac{1}{C}$ as small as possible. This evident in the fact that the plot of **X** is well spread, giving us an optimal visualisation even from the high-dimensionality of the data set. Furthermore, the spiral nature of the parameter **X** is expected as it stems from the non-linear mapping of the *sin* and *cos* functions.

As the images show, the representation of after learning is very similar to the representation from the

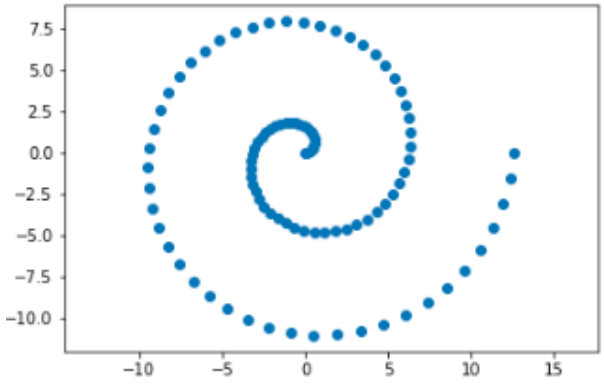

Figure 14: *Observed representation of X*

generated data. However, it is slightly rotated compared to the originally generated data. Nevertheless, this does not actually imply that the parameter learning has gone wrong. The rotation of **X** arises from the independent identically distributed additive Gaussian noise seen across all the dimensions. However, this can be looked over as the distribution is invariant under the effects of rotation. Taking a rotation matrix **R** which follows that $\mathbf{RR}^T = \mathbf{I}$ and applying it to **W** to have **WR** would have no effect on the marginal likelihood

$$\begin{aligned}
-log(p(Y|W)) = \frac{ND}{2}log(2\pi) + \frac{N}{2}log|(WR)(WR)^T + \sigma^2 I| \\
+\frac{1}{2}\sum_{n=1}^{N}(y_n - \mu)^T((WR)(WR)^T + \sigma^2 I)^{-1}(y_n - \mu)
\end{aligned} \tag{45}$$

where

$$(WR)(WR)^T = WRR^TW^T = WIW^T = WW^T \tag{46}$$

So even with the rotation, the generating parameter **X** will still be able to predict the same values of **Y**.

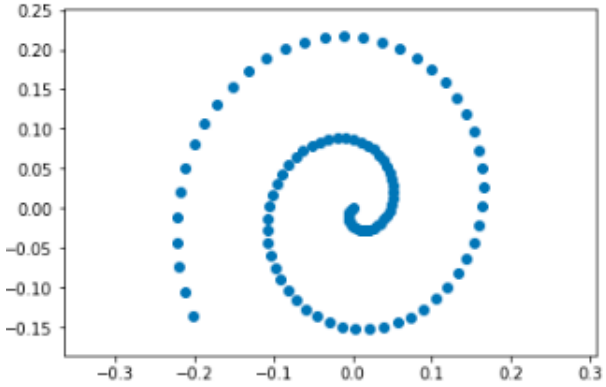

Figure 15: *Learned representation of X*

Q22 Now the data is plotted on a random 2D subspace, which was achieved by simply sampling from a normal distribution into a 2D matrix and plotting it against **Y**. Interestingly, the spiral form of the data is still represented, however the points are much closer together. This comes down to the fact that there is no optimisation being done on the objective function. Therefore, the co-variance isn't being maximised as it was like in the previous plot. This means, with the co-variance much smaller, the visualisation of **X** is in a dimension where the data is not very distinguishable from each other. Each of these plots are essentially seeing the same data in different ways, where certain dimensions or 'principal components' will give us a better representation of the data due to its larger spread, and finding the right dimension is dependent on being able to capture the maximum variance within the high-dimensional data set.

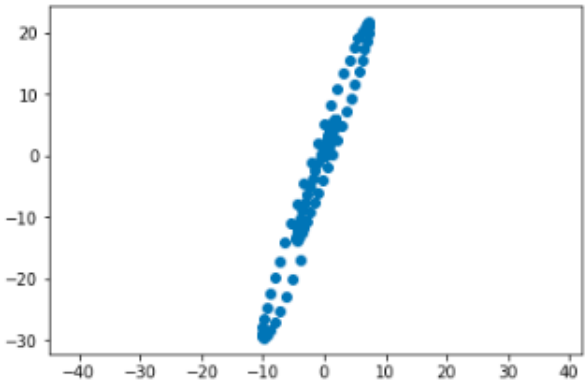

Figure 16: *Data plotted on different 2D subspace*

3 Evidence

Q23 This assumption implies that any outcome is equally likely. This is the simplest model as it takes into account the least number of features possible, zero. It only takes into account the cardinality of the output. It is also the most complex as the probability mass is spread aall outcomes, making it hard to infer from.

Q24 We have restricted the model by strictly modelling the outcome as independent of each other. This might or might not be the case. All 3 models are more flexible than the initial model,$M0$, as we now have parameters over discrete amount of variables. Giving the parameter vector more components increases the flexibility of each model. As such, $M3 > M2 > M1$ in terms of flexibility. Even if there is an overlap in the vector parameter, they can take the same value. However, the more parameters in our vector space increases restriction as well as now we have the probability mass spread over a larger area hence making them harder to infer from. $M3$ is suitable to model a more complex model whereas $M1$ would be suitable to model a relatively simpler model.

Q25 As mentioned in earlier sections, the prior encodes our belief in a certain distribution. Contrary to previous sections, we now have to encode our beliefs in the values a parameter can take. This prior implies that we are uncertain what value it can take, hence we give it zero mean and a large variance, letting it be equally likely to take a large range of values.
Choosing a prior with such a large variance would give the parameter more freedom as the sampling is done from a large range of value. However we lose, to an extent, the benefit that the prior gives; which is to give some level of protection against uncharacteristic results. Of course, in order to encode a stronger prior over the parameter space, it must be based on strong grounds, as a wrongly assumed strong prior would negatively affect our learning as well. The idea of a 'good' prior is definitely subjective, but it would be beneficial if we can encode an assumption we can prove into the distribution of the parameters.

Q30 The coursework has been a step-by-step process of understanding the fundamentals of machine learning. with two side-by-side pathways of learning, both integral to gaining an intuition of how machine learning works. On one hand, it is the theoretical knowledge of data and its characteristics of being represented as prior, posterior, uncertainty and using this to be able to model data best to our abilities. This can be by integrating our prior beliefs to augment the data with more meaning, or by adding additive noise to get a better representation of data that would otherwise be over-fitted. On the other hand, the theory led to us obtaining a 'toolkit' that allow us to act on the data using different models such as parametric/non-parametric models and different kinds of learning such as supervised/unsupervised. By seeing different techniques used in the practical sections on different representations of data, we were able to get a clearer understanding of which models apply to which circumstances. However, not only did we see the defining characteristics of each approach, but also how some techniques can actually overlap in terms of their usability. For example, with a large enough data set, an ML estimate is likely to be as good as a MAP estimation even though it accounts for the prior too.

4 Sources Used

[1] Rick Wicklin (2012) - What is Mahalanobis distance?
https://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance.html
[2] Jim Grange - (Pesky?) Priors
https://jimgrange.wordpress.com/2016/01/18/pesky-priors/