Models

Justin Salmon[1] **and George Lancaster**[2]

[1] *wr18313*
[2] *qv18258*

October 30, 2018

1 The Prior

1.1 Theory

Question 1.1

A Gaussian likelihood encodes the inherent noise present in most real-world data. The central limit theorem states that in most cases, when independent random variables are added and normalised, the result is a gaussian distribution. In other words, most probabilistic processes in nature tend to be noisy, and that noise tends to follow a Gaussian distribution. Hence this is generally a good first assumption to make about unknown data.

Question 1.2

Choosing a spherical covariance matrix means that we are assuming that the distribution is equally likely to deviate from the mean in all directions. Additionally, we assume that all dimensions of y are independent, and therefore do not covary with one another. Again, this is a good place to start.

Choosing a non-spherical covariance would imply that we know something in advance about the relationship between the different dimensions of y.

Question 2

The covariance matrix would not be in terms of the identity matrix. We would have non-zero values in the offset diagonals which correspond to the correlations between different variables. If we apply the product rule,

1.1.1 Linear Regression

Question 3

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) = \prod_{i=0}^{N} \mathcal{N}(y_i|\mathbf{W}^T\phi(x_i), \beta^-1) \tag{1}$$

Question 4

A distribution is conjugate to another if they both take the same algebraic form, meaning that they are in the same probability distribution family. For example, Gaussians are conjugate to each other, and the conjugate to a Bernoulli distribution is a Beta distribution. Conjugates are used as a convenience to avoid calculating the denominator in Baye's rule (the evidence) which can often be an integral. If the prior and likelihood are conjugate, then their product will be proportional to the posterior.

Question 5

Euclidean distance from the mean X appears only in the gaussian exponential part Euclidean distance because of spherical covariance?

Question 6

Derive posterior mean and covariance, start at conjugacy. Watch video https://www.youtube.com/watch?v=nrd4AnDLR3U

1.1.2 Non-parametric Regression

Question 7

Non-parametric models are not focused on defining a set of parameters, but using what we know about the current data to classify new unseen data points. The data can be seen as analogous to the parameters. A good example of this is the K-nearest-neighbor model, which classifies new data points based on its surrounding classes. Unlike parametric models, non-parametric models do not assume that there is a finite set of parameters. Because of this, their complexity is not bounded by the number of parameters.

Non-parametric models may be more difficult to interpret as they do not have direct parameters to describe the model.

Question 8

This prior represents the space of all possible functions, however we want this to be constrained by making some functions more likely than others. The covariance $k(\mathbf{X}, \mathbf{X})$ allows us to assume that any two values x_i and x_j covary, therefore f_i and f_j can also be expected to covary. This means that we think that smooth functions are more likely, however the probability of saw-tooth functions are non-zero.

Question 9

All functions are possible, however some are more likely than others.

Question 10

$$p(\mathbf{Y}, \mathbf{X}, f, \theta) = p(Y|f)p(F|X, \theta)p(X)p(\theta) \qquad (2)$$

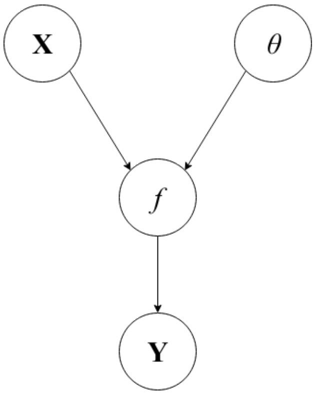


Figure 1: *Graphical model of the joint distribution.*

- $\mathbf{X}$ and θ do not depend on anything;
- $\mathbf{F}$ depends on θ and $\mathbf{X}$;
- $\mathbf{Y}$ depends on $\mathbf{F}$, but is is conditionally independent of $\mathbf{X}$ and θ.

Question 11

The marginalisation in Eq. 2 connects the prior and the data because we now have a way to directly generate $\mathbf{Y}$ values given $\mathbf{X}$ and θ value without knowing the actual form of f.

Because we are uncertain about f, when we marginalise it out, the uncertainty gets pushed onto $\mathbf{Y}$.

The fact that θ is left on the left hand side of the expression after marginalisation means that it is needed, with $\mathbf{x}$, to calculate $\mathbf{Y}$. This implies that $\mathbf{Y}$ is dependent on theta.

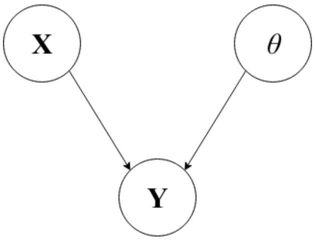


Figure 2: *Graphical model of the marginalised distribution.*

1.2 Practical

1.2.1 Linear Regression

Question 12.1

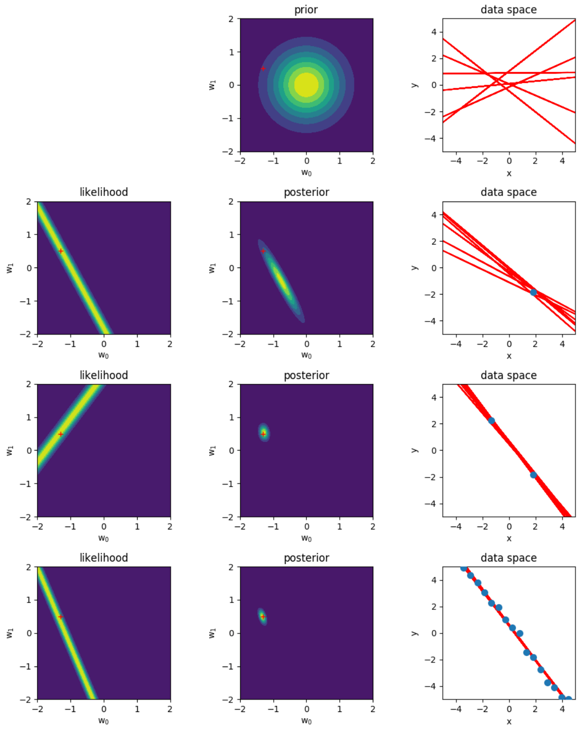


Figure 3: *Implementation of linear regression. The left hand column plots the likelihood. The middle column plots the prior/posterior, and the right hand column shows six random sample functions. Plots are drawn after one, two and twenty data points are revealed.*

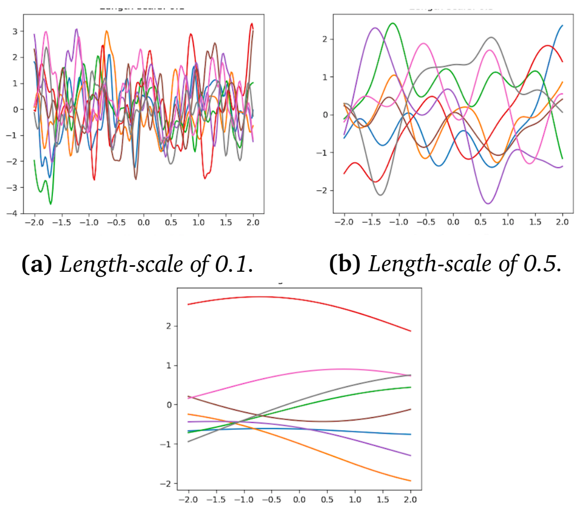
Question 12.5

After a single data point is added, we can see that the posterior distribution begins to squash in one direction to converge onto the parameters **W**. When we sample from this distribution, there are many lines passing through the point at different gradients, which is because we cannot define two parameters from a single data point. As we begin to reveal more data, the posterior distribution centres onto **W** and the sample functions fit more closely to the data points. This is a desirable behaviour as it shows that we have relearned the parameters of the model used to generate the data **X**.

Question 12.6

The posterior converges on the position of **W** because after a new data point is added, we update the parameters using the newly calculated mean and covariance.

Question 13



(a) *Length-scale of 0.1.*

(b) *Length-scale of 0.5.*

(c) *Length-scale of 5*

Figure 4: *Squared exponential with varying length scales.*

Question 13.4

Increasing the length-scale of the covariance function allows us to constrain the functions smoothness. A smaller length-scale creates functions with more rapid changes than those with a larger length-scale. This is because a higher length-scale means that, for two random variables x_i and x_j where the covariance is $k(x_i, x_j)$, their instantiations as f_i and f_j are similar. The samples in Fig 5 show the effect of varying length-scales.

Question 13.5

The lengthscale encodes how smooth the assume the function to be. A smaller lengthscale produces a function with higher peaks and lower troughs than a higher lengthscale.

1.2.2 Question 14

Need to describe the plots. What happens if we use a diagonal covariance matrix to the squared exponential.

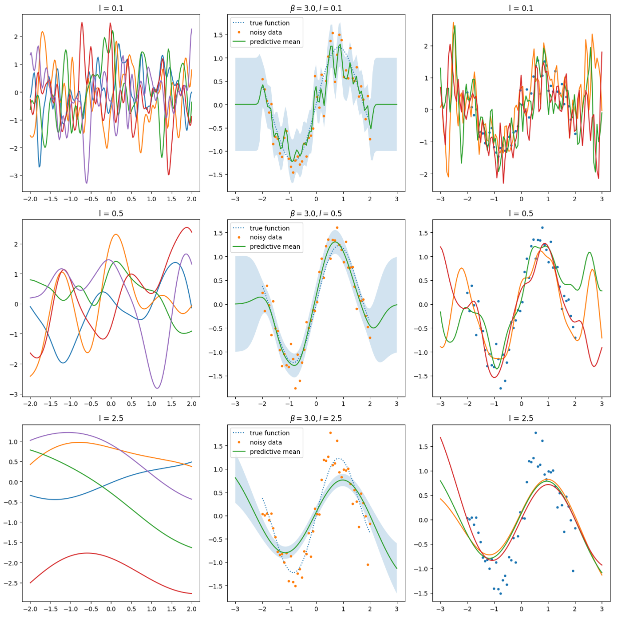


Figure 5: *Implementation of non-parametric regression with changing length-scales.*

2 Posterior

Question 15

We use our beliefs to find a starting point for the problem and to make assumptions about the data, or any other characteristics of the problem. Assumptions are what we assume to be true about something. If we had all the data we required for a problem, there would be no need to assume anything. This gives us a way to reason when we have very little information.

A preference is an assumption that we want to use because it is easier to do so. For example using a gaussian likelihood so that the posterior is also gaussian.

We use beliefs and assumptions to formulate a prior. Believe that the smooth lines are more likely. Prefer straight or flat lines.

Question 16

Since the covariance matrix is spherical, we have assumed that the latent input variables are independent and that they are centred around the origin.

Question 17

Look at the Gaussian Identities document to try to outline the steps to integrate out the variable we are not interested in

Question 18.1

A maximum-a-posteriori (MAP) estimation is equal to the mode of the posterior. MAP finds the parameters that maximise the posterior distribution.

The maximum likelihood (ML) can be seen as a type of MAP estimation, which assumes a uniform prior distribution of the parameters. In doing so, it finds the parameters that maximise the likelihood.

Type-II Maximum likelihood finds the parameters that maximise the marginal likelihood.

Question 18.2

With only one data point, ML will be less accurate than MAP since it does not take the prior into account. However as more data is seen, the two will begin to converge on one another.

Question 18.3

The denominator of the posterior distribution has no bearing on the result as it will always be positive, and depends on the parameters $\mathbf{W}$.

Question 19

The objective function can be written as:

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= -log(p(\mathbf{Y}|\mathbf{W},\boldsymbol{\mu},\boldsymbol{\sigma}^2))\\ &= -\frac{ND}{2}log(2\pi) - \frac{N}{2}log|\mathbf{C}| - tr((\mathbf{YC})^{-\mathbf{1}}\mathbf{Y^T}) \quad (3)\end{aligned}$$

where $\mathbf{C} = \mathbf{WW^T} + \boldsymbol{\sigma}^\mathbf{2}\mathbf{I}$. The gradients of the objective function with respect to the parameters can be written as:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{W}} &= tr(\mathbf{C^{-1}}(\mathbf{WJ_{ij}} + \mathbf{J_{ji}W^T}))\\ &+ tr(\mathbf{Y^TY}(-\mathbf{C^{-1}}(\mathbf{WJ_{ij}} + \mathbf{J_{ji}W^T})\mathbf{C^{-1}}) \quad (4)\end{aligned}$$

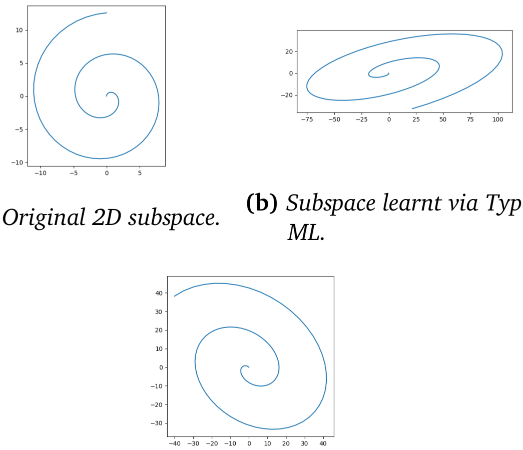
where $\mathbf{WJ_{ij}} + \mathbf{J_{ji}W^T} = \frac{\partial\mathbf{C}}{\partial\mathbf{W_{ij}}}$.

Question 20

The uncertainty in $\mathbf{X}$ and θ is pushed into f. Therefore, by marginalising out $\mathbf{X}$, we lose some of the information required to calculate an accurate $\mathbf{Y}$.

Question 21

We first mapped the input variables x non-linearly to a 2D space and plotted the result. This gives a uniform spiral shape as seen in Figure 4a. We then mapped this linearly to a 10D space. Given the objective function and gradients from above, we then minimised the log likelihood to "learn" the linear mapping parameters W. Using the new parameters, we performed the reverse of the linear mapping back to 2D space and obtained another (slightly different) spiral which can be seen in Figure 6b.



(a) *Original 2D subspace.*

(b) *Subspace learnt via Type-II ML.*

(c) *Random subspace.*

Figure 6: *Subspaces representing a 2D non-linear mapping from a 1D input vector.*

We expected the recovered shape to be more similar to the original shape. We suspect there may be an error in the minimisation process (probably somewhere in computing the gradients).

Since we have only learned the parameters for the linear part of the mapping, it does not immediately appear possible to recover the original latent variables themselves.

Question 22

Choosing any randomly initialised W matrix and performing the reverse mapping also results in a spiral shape, as can be seen in Figure 6c. However, the shape is slightly different, the dimensions are different.

3 The Evidence

Question 23

Question 24

Question 25

4 Final Thoughts

Question 30

I feel that the purpose of this assignment was to introduce the basics of machine learning using a statistical, low level approach. In doing this, we have been forced to learn the theory behind algorithms before we are able to implement them.

The main learning point of this assignment has been the methodology of machine learning, which involves defining a prior and likelihood with the goal of finding a posterior distribution.