Coursework 1 - Modelling

Jonas Osborn jo14944 | **Tristan Saunders** ts16802 | **Corin Varney** cv14985

1 The prior

1.1 Theory

Question 1 *1.* As the instances of y are noisy observations of the underlying process and we do not know anything about this uncertainty we can assume it is the sum of independent and identically distributed errors. The Central Limit Theorem states the distribution of the sum of a large enough number of independent, identically distributed variables will be approximately normally distributed. From this we can say that our model of y has the following form:

$$y = f(x) + \epsilon$$
$$\text{where: } \epsilon \sim \mathcal{N}(0, \boldsymbol{I})$$

And so from this we have the likelihood of each y_i as a Gaussian distribution with mean $f(x_i)$.

2. Choosing a spherical covariance matrix for the likelihood means that we are assuming the different dimensions of y_i to be independent and identically distributed. As they are independent they do not covary with each other and so the covariance matrix is diagonal. As they are identically distributed they all have the same variance. Therefore the covariance matrix is spherical.

Question 2

$$\begin{aligned} p(\boldsymbol{Y}|f, \boldsymbol{X}) &= p(y_1, \ldots, y_{N-1}, y_N|f, \boldsymbol{X}) \\ &= p(y_N|y_{N-1}, \ldots, y_1, f, \boldsymbol{X}) p(y_{N-1}|y_{N-2}, \ldots, y_1, f, \boldsymbol{X}) \ldots p(y_1|f, \boldsymbol{X}) \end{aligned}$$

1.1.1 Linear regression

Question 3

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}) &= \prod_i^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{W}) \\ &= \prod_i^N \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i, \sigma^2 \boldsymbol{I}) \\ &= \mathcal{N}(\boldsymbol{W}\boldsymbol{X}, \sigma^2 \boldsymbol{I}) \end{aligned}$$

Question 4 A conjugate prior is one that is conjugate to the posterior, meaning they are in the same family of distributions. A conjugate prior is useful as it gives a closed-form solution for the posterior. If we didn't choose a conjugate prior then numerical integration may be necessary to calculate the posterior which may mean the solution is potentially intractable. The conjugate prior for a Gaussian posterior is a Gaussian.

Question 5 Just as encoding the prefrence in a L_2 norm is equivalent to having a Gaussian prior, encoding the preference using a L_1 norm is equivalent to having a Laplace prior. This is because the Laplace distribution estimates median rather than the mean estimated by the Guassian and median minimises the L_1 norm and mean the L_2

The shape of the laplace distribution's probability density function, with it's higher peak around zero compared to the probability density function of a Gaussian means that more co-efficients are likely to be equal to zero and this leads to a sparser model than those produced by a Gaussian prior.

Question 6

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) &= \frac{1}{Z} p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{W}) \\ &= \frac{1}{Z} \mathcal{N}(\boldsymbol{W}\boldsymbol{X}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{W}_0, \tau^2 \boldsymbol{I}) \end{aligned}$$

$\frac{1}{Z}$ is the normalising constant to ensure that the posterior is a probability density function by making the area under the graph equal to 1 this constant is called the evidence. We will ignore it for now.

By the probability density function of the multivariate normal distribution, we have:

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) \propto & \frac{1}{\sqrt{(2\pi)^N \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) \right) \\ & \cdot \frac{1}{\sqrt{(2\pi)^N \tau^2}} \exp\left(-\frac{1}{2\tau^2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0) \right) \end{aligned}$$

We ignore the normalising constants as we re-normalise with Z and then combine and multiply out the exponents.

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) &\propto \exp\left(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) - \frac{1}{2\tau^2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0) \right) \\ &\propto \exp\Bigg(-\frac{1}{2\sigma^2} (\boldsymbol{Y}^T \boldsymbol{Y} - 2\boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{W} + \boldsymbol{W}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{W}) \\ &\qquad - \frac{1}{2\tau^2} (\boldsymbol{W}_0^T \boldsymbol{W}_0 - 2\boldsymbol{W}^T \boldsymbol{W}_0 + \boldsymbol{W}^T \boldsymbol{W}) \Bigg) \end{aligned}$$

We know the posterior will be Gaussian as both the likelihood and prior are so can assume it will take the form: $\exp\left((\boldsymbol{W} - \mu)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{W} - \mu)\right)$. If we multiply the exponent out we get a quadratic so we try and make the posterior we have look like this quadratic.

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) \propto \exp\Bigg(& -\frac{1}{2\sigma^2} \boldsymbol{W}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{W} - \frac{1}{2\tau^2} \boldsymbol{W}^T \boldsymbol{W} && \text{quadratic term} \\ & + \frac{1}{\sigma^2} \boldsymbol{W}^T \boldsymbol{X}^T \boldsymbol{Y} + \frac{1}{\tau^2} \boldsymbol{W}^T \boldsymbol{W}_0 && \text{mixed term} \\ & - \frac{1}{2\sigma^2} \boldsymbol{Y}^T \boldsymbol{Y} - \frac{1}{2\tau^2} \boldsymbol{W}_0^T \boldsymbol{W}_0 \Bigg) && \text{constant term} \end{aligned}$$

By re-arranging and completing the square we can find both $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}^{-1}\mu$ in the quadratic and mixed terms respectively.

$$
\begin{aligned}
p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) \propto \exp\Bigg(& -\frac{1}{2}\boldsymbol{W}^T \overbrace{\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)}^{\mathbf{\Sigma}^{-1}} \boldsymbol{W} && \text{quadratic term} \\
& + \boldsymbol{W}^T \underbrace{\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\right)}_{\mathbf{\Sigma}^{-1}\mu} && \text{mixed term} \\
& - \frac{1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{2\tau^2}\boldsymbol{W}_0^T\boldsymbol{W}_0\Bigg) && \text{constant term}
\end{aligned}
$$

From these we can easily find $\mathbf{\Sigma}$ and μ.

$$
\begin{aligned}
\mathbf{\Sigma} &= \left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1} \\
\mathbf{\Sigma}^{-1}\mu &= \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0 \\
\mu &= \mathbf{\Sigma}\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\right)
\end{aligned}
$$

Thus we have our posterior as a Guassian with our values for $\mathbf{\Sigma}$ and μ:

$$
\begin{aligned}
p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{W}-\mu)^T\Sigma^{-1}(\boldsymbol{W}-\mu)\right) \\
&\propto \mathcal{N}\left[\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\right), \left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1}\right]
\end{aligned}
$$

1.1.2 Non-parametric regression

Question 7 Parametric models assume that the distribution the data comes from is based on a finite, fixed set of parameters and models future predictions based off these parameters, they capture everything there is to know about the data. Non-parametric models do not make such assumptions about the model structure and instead infer structure from the data, they have parameters but these are not fixed in advance and there can be an infinite set of parameters.

Non-parametric models are more flexible and can represent a wider variety of data and will represent the data better if the assumptions made in the parametric model are incorrect but are less precise and accurate than parametric methods if the right assumptions are made.

Parametric models are often easier interpreted as they are simpler to transcribe and are also often faster to compute due to lacking the complexity and flexibility of the non-parametric models.

Question 8 As we use a Gaussian process we define this prior over functions and want our prior to put some constraints on the space of functions. The fact that its a Gaussian process means that for an arbitary set of points $x_i, \ldots, x_j$ we assume that $p(f_i), \ldots, f(x_j))$ is jointly Gaussian with mean 0 and covariance function k, from this we have the equation for our prior:

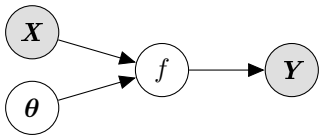
$$p(f|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(0, k(\boldsymbol{X}, \boldsymbol{X}))$$

The covariance function k allows us to set constraints from our assumptions about the mapping f. For example our assumption about the functions smoothness that if x_i and x_j are similar then we expect f_i and f_j to be similar too, with certain kernel functions for f we can only sample functions that have sufficient smoothness by ensuring that constraint.

Question 9 [TODO]

Question 10

$$p(\boldsymbol{Y}, \boldsymbol{X}, f, \boldsymbol{\theta}) = p(\boldsymbol{Y}|f)p(f|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{X})p(\boldsymbol{\theta})$$



[UNSURE]

- $\boldsymbol{X}$ and $\boldsymbol{\theta}$ are independent
- f is conditionally dependent on both $\boldsymbol{X}$ and $\boldsymbol{\theta}$
- $\boldsymbol{Y}$ is conditionally dependent on f and conditionally independent of $\boldsymbol{X}$ and $\boldsymbol{\theta}$

Question 11 [UNSURE] This marginalization shows the likelihood of the data we have observed over the function we are testing which is constrained constrained by the prior.

There are two sources of uncertainty here, that associated with f in the prior and that associated with ϵ in the likelihood, these are independent and as such are merged by simply adding to form the covariance of the marginal likelihood Gaussian.

Leaving the θ on the left-hand side of the expression implies that we still have specific hyperparameters rather than undefined ones. It remains throughout the integral.

1.2 Practical

1.2.1 Linear regression

Question 12 [TODO]

1.2.2 Non-parametric regression

Question 13 [TODO]

Question 14 [TODO]

2 The posterior

2.1 Theory

Question 15 Specifiying a prior assumption over $\boldsymbol{X}$ allows us to encode our preference about the nature of the properties that $\boldsymbol{X}$ should have. This assumption also constrains the variable $\boldsymbol{W}$ as $\boldsymbol{W}$ and $\boldsymbol{X}$ have a simple relationship.

Question 16 We have encoded the assumption that all the dimensions in each variable of $\boldsymbol{X}$ are independent and identically distributed as we have used a Gaussian with an identity matrix as its covariance matrix.

Question 17 [TODO]

2.1.1 Learning

Question 18 [TODO]

2.1.2 Practical optimisation

Question 19 [TODO]

2.1.3 Non-parametric

Question 20 [TODO]

2.2 Practical

2.2.1 Linear representation learning

Question 21 [TODO]