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

$$p(\boldsymbol{Y}|f, \boldsymbol{X}) = p(\boldsymbol{y}_N|\boldsymbol{y}_{N-1}, \ldots, \boldsymbol{y}_1, f, \boldsymbol{X})p(\boldsymbol{y}_{N-1}|\boldsymbol{y}_{N-2}, \ldots, \boldsymbol{y}_1, f, \boldsymbol{X}) \ldots p(\boldsymbol{y}_1|f, \boldsymbol{X})$$

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 [TODO]

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 probability distribution over the uncountably infinite space of functions. So every point in the input space is associated with a random variable that has been normally distributed. This prior represents the joint distribution of the random variables.

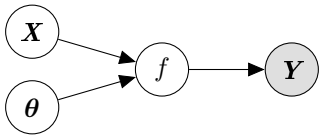
$$p(f|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(0, k(\boldsymbol{X}, \boldsymbol{X}))$$

The marginal [TODO]

Question 9 [TODO]

Question 10

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



- Are these the assumptions? [TODO]
- $\boldsymbol{X}$ and θ are independent
- f is conditionally dependent on both $\boldsymbol{X}$ and θ
- $\boldsymbol{Y}$ is conditionally dependent on f and conditionally independent of $\boldsymbol{X}$ and θ

Question 11 [TODO]

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

2.1.1 Learning

2.1.2 Practical optimisation

2.1.3 Non-parametric

2.2 Practical

2.2.1 Linear representation learning