Models

Lin Feng[1] and Zilya Amerkhanova[2]

[1]*candidate number 97302*

[2]*candidate number 97497*

November 1, 2018

The purpose of this report is to examine the fundamental principles of machine learning and the process of building data models. It begins with the prior section and then focuses on the posterior and evidence. Finally, it concludes with reflection on our learning experience as a whole.

1 The Prior

Question 1

1.1 The assumption is that each data point is independent of the others and identically distributed (i.i.d.). That means that the likelihood is the product of the probabilities of each individual data point. Choosing a Gaussian likelihood is very useful – average/sum of i.i.d. samples from any distribution will follow a Gaussian distribution, and that means we do not necessarily have to know which distribution the data comes from.

1.2 It means we assume each parameter in two dimensions is independent, and they have the same probability density for a Gaussian distribution that the covariance matrix is proportional to the identity matrix.

Question 2

$p(Y \mid f,X) = p(y_1,..., y_n \mid f,X)$

If we do not assume independence of data points we will have to deal with the joint $p(y_1,..., y_n|f,X)$, instead of looking at individual probabilities $p(y_i| f,X)$.

Question 3

The likelihood is of Gaussian form. Given N training examples, the likelihood assuming i.i.d. is

$$p(Y|X,W) = \prod_{i}^{N} p(y_i \mid x_i, W)$$

Question 4

Conjugate distributions are prior and posterior distributions that are in the same probability distribution family. Choosing the conjugate prior (i.e. Gaussian) is a sensible choice as the posterior will also be Gaussian. A conjugate prior helps compute posterior distribution as it gives a closed-form expression for the posterior distribution, meaning the calculations can be done in a finite number of operations and typically involve no limit calculations. Knowing the form of posterior distribution helps to identify the parameters.

Question 5

For a spherical covariance matrix, since two independent parameters have the same probability distribution, so we can think it as a single Gaussian distribution, simply use Euclidean distance function to estimate the probability density function of a random variable, then we can find the maximum likelihood.

Question 6

We know that the posterior is proportional to the likelihood multiplied by prior due to conjugacy

$$p(W|X, Y) \propto p(Y|X, W)p(W)$$

Since we derived likelihood

$$p(Y|X,W) = \prod_{i}^{N} p(y_i | x_i, W)$$

and our prior is

$$p(W) = N(w_0, \tau^2 I)$$

we can now compute the posterior as per our mapping function.

$$p(W|X, Y) = \frac{1}{Z} p(Y|X, W)p(W)$$

$$p(W|X, Y) = \frac{1}{Z} \prod_{i}^{N} p(y_i | x_i, W)\, N(w_0, \tau^2 I)$$

As the likelihood and prior are Gaussians, the intuition is that posterior will also be of Gaussian form.

Question 7

Parametric models assume data can be defined by finite set of parameters. We encode relationship between variates using parameters w, which induced a corresponding prior distribution, to obtain the corresponding posterior distribution over parameters w. Given the parameters, predictions are independent of the observed data. Therefore, the complexity of parametric models are bounded even if the amount of data is unbounded, they only interpret the local data, they are not very flexible.

In contrast, non-parametric models assume data is defined by a function with infinite dimensional. We encode relationship between variates using variates themselves, define a prior probability distribution over functions directly, the corresponding posterior distribution may be various functions. Predictions are dependent of the observed data. Therefore, non-parametric models can grow as the amount of data grows and their complexity depends on data. They are more flexible.

The parametric model is easier to interpret than a non-parametric one, as the complexity of the model does not change. The non-parametric model on the other hand allows unlimited number of parameters - it can create new parameters, as opposed to the parametric model.

Question 8

The Gaussian process prior represents distribution over functions, and if we assume the mean function to be zero then the Gaussian process is fully defined by its covariance function k(X,X).

Question 9

This prior encodes all possible functions, because this prior can represent any function k(X,X) which is evaluated at any two values of x.

Question 10

The prior is

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

And therefore the joint distribution of training outputs f and test outputs f* is

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

Where function f is evaluated at test points X*. With n training points and n* test points K(X,X*) indicates a n×n* matrix of the covariances assessed at all pairs of training points and test points. The same applies for K(X,X*), K(X*,X) and K(X*,X*).

The Gaussian process graphical model is shown in figure 1 below, assuming the following

- No noise
- Each y_i is conditionally independent given f_i
- Observed data is in squares, circles are latent, the bar represents fully connected nodes

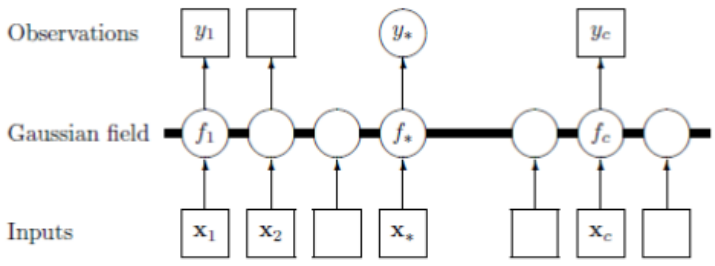

Figure 1: Graphical model for a GP regression. Figure from Rasmussen and Williams (2006)

Question 11

To perform marginalisation over the mapping *f* we calculate the integral of the likelihood multiplied by prior.

The Gaussian process marginalization allows for addition of more inputs *x* and targets *y* - it does not change the distribution of any other variables. This allows for uncertainty.

θ on the left means we calculate the marginal likelihood as a function of kernel hyperparameters θ.

Question 13

GP-prior with a squared exponential covariance function:

$$p(f) = N(0, k(X, X))$$
$$k(x_n, x_m) = \sigma^2 exp(-\frac{(x_n - x_m)^2}{2l^2})$$

After sampling from this prior and visualising the samples we get the following graph (see figure 2 below). The figures 3 and 4 then show samples using different length-scale for the squared exponential.

In those examples, we alter the length-scale to adjust how smooth the function is. Small length-scale value means that function values can change quickly (figure 3) and large length-scale values characterize functions that change very slowly (figure 4).

Length-scale assumes the how far we can reliably extrapolate from the training data. In general, we will not be able to extrapolate more than the length-scale away from the data.

Question 14

After computing the predictive posterior distribution of the model, sampling from this posterior with points both close to the data and far away from the observed data as well as plotting the data, the predictive mean and the predictive variance of the posterior from the data, the results can be seen in figure 5 below.

Using the non-parametric regression, the posterior has the similar shape with the observed data, it can go through the each observed data with small length-scale.

2 The Posterior

Question 15

Assumptions are the conditions added without proof. Those conditions are extremely sensitive and uncertain. For instance, if we assume some random variables can be categorized in linear regression model, so we can use a linear equations to represent the variables. Without this assumption, they maybe have any distribution we do not know.

Belief is the possibility added on reality, the possibility can be changed within the range. For instance, if we believe all answers have the same possibility, so that the final answer is the mean of all answers. Without this belief, the final answer may be any one of the answers, but still in the range of all answers.

Preference is the conditions we added with expecting the data, usually they are existing options, we choose them to formulate the data more easily. For example, after observing the data, we can choose using different models to represent the data. They are different ways how we obtain predictions, but we cannot tell any of them will get the correct predictions.

Question 16

We assume that $x = (x_0, \dots, x_i)$, where x_i is independent Gaussian distribution. Observed data shows a preference that in each dimension, x_i is a standard Gaussian distribution with variance 1 and mean 0.

Question 17

$$p(y, x) = p(y|x)p(x)$$
$$p(y, x|W, \mu, \sigma^2) = p(y|x, W, \mu, \sigma^2)p(x)$$
$$\text{as } p(x) = N(0, I), \text{ so } k = WW^T, \; C = k + \sigma^2 I = WW^T + \sigma^2 I$$
$$p(y|W) = \int p(y|x, W)p(x)dx = N(x|\mu, C) = N(x|\mu, WW^T + \sigma^2 I)$$

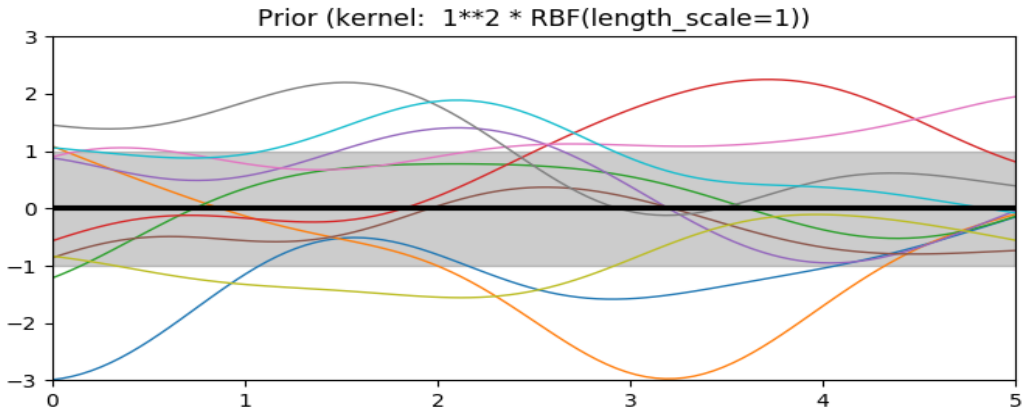

Figure 2: Sampling from the prior (Question 13)

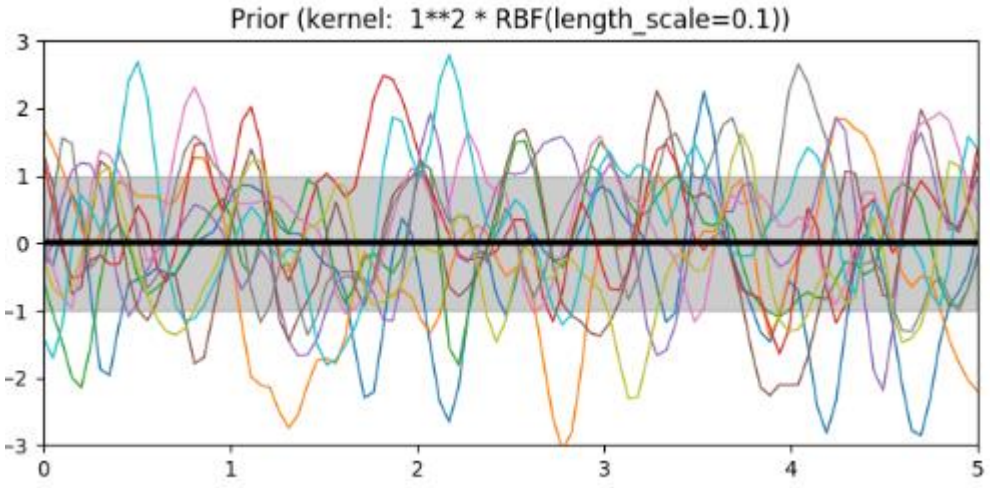

Figure 3: Samples using small length-scale values for the squared exponential (Question 13)

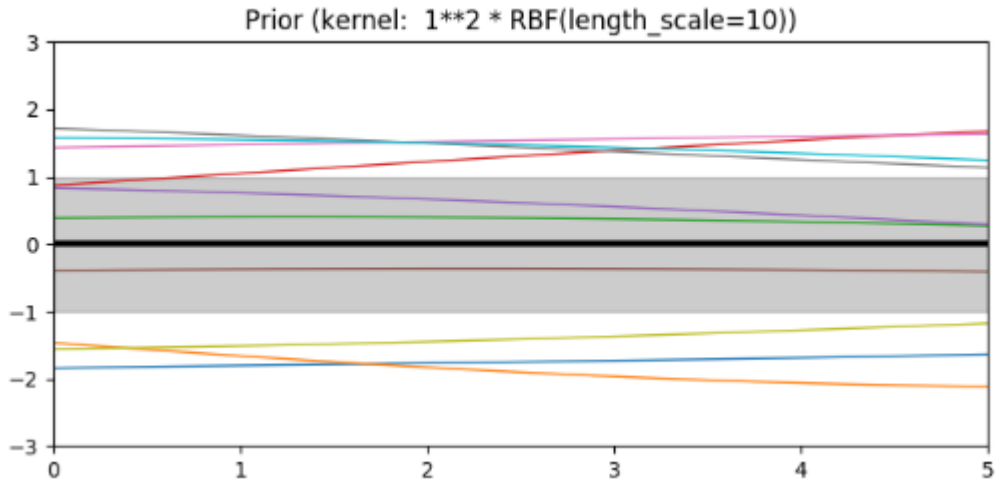

Figure 4: Samples using large length-scale values for the squared exponential (Question 13)

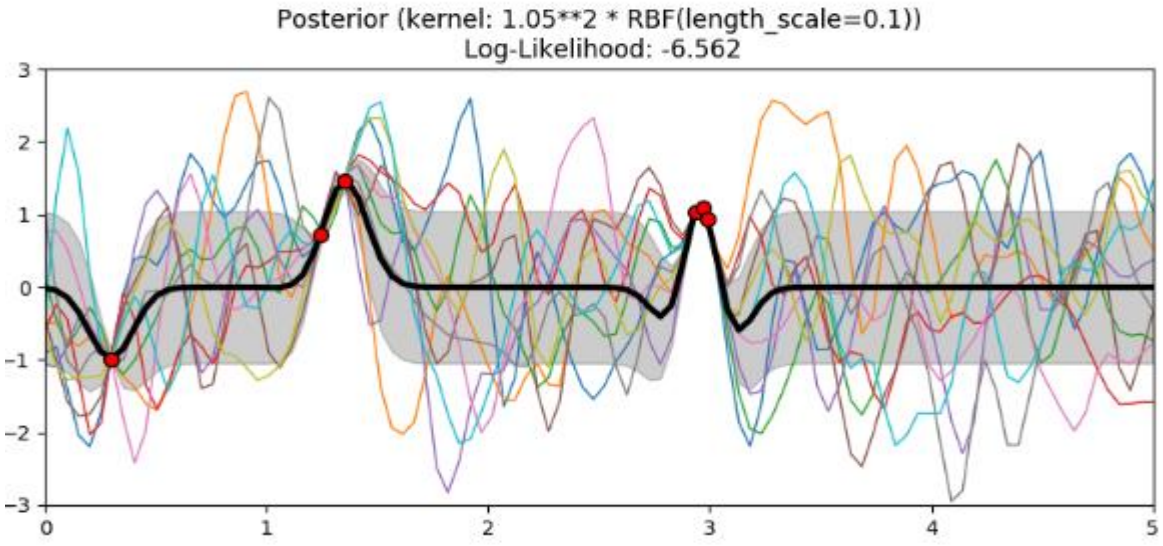

Figure 5: Plot of the data, the predictive mean and the predictive variance of the posterior from the data (Question 14)

Question 18

MAP estimation tries to find W that will balance between the prior and likelihood, it imposes a regulariser on W and is less biased than ML.

As we observe more data, the ML and MAP results will become more similar until the point when they are practically the same.

The expressions are equal as the marginal likelihood is positive and does not influence the optimisation in any way – it does not depend on θ.

Question 19

Assume

$$p(y = 1|x; w) = \sigma(w^T x)$$
$$p(y = 0|x; w) = 1 - \sigma(w^T x)$$
$$p(y|x; w) = (\sigma(w^T x))^y (1 - \sigma(w^T x))^{(1-y)}$$

The negative log likelihood is

$$L(w) = -\sum_i^N y_i log\sigma(w^T x_i) + (1 - y_i) log(1 - \sigma(w^T x_i))$$

The gradients when

$$y \in \{0, 1\}$$
$$w := w - \eta(y_i - \sigma(w^T x_i))x_i$$

Question 20

Marginalisation of f is much simpler to do than marginalising out X, as X represents observed data.

Question 21

Assume the model is in the form of

$$f(x) = \theta^T x$$

In figure 6 below, with 3 observed data points (blue point), draw a *Y=f(x)* (green line), find the minimum sum of the gradient of the data at *Y=f(x),* so we can get the optimization function which represents the average of all observed data (red line).

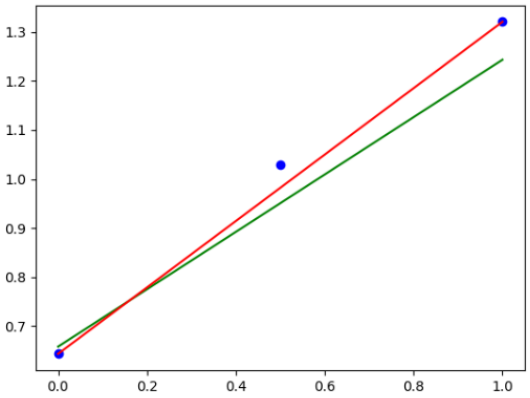

Figure 6: Representation of the data (Question 21)

3 The Evidence

Question 23

We can think of it as the M3 in Figure 1, it sets possibility distribution to $D_i = \frac{1}{512}$ which means it treats all data sets equally, so it is the simplest possible model because it simply includes all given data sets. However, in this case, the prior *p(θ|M),* the parameterization becomes very important, and the quantities of observed data will affect the function a lot, since the mean of possibility will not be $D_i = \frac{1}{512}$ any longer, it will be most complex model.

Question 24

$p(D|M_1, \theta_1)$ model1 only restricts x_1^i the first dimension of x_n^i , so it can imply the uncertainty of x_1^i, and it's more flexible if the data sets have the decision boundary about function of x_1^i but not x_2^i. But it cannot model decision boundaries if the data sets has rotation invariance.

$p(D|M_2, \theta_2)$ model2 restricts both dimension of x^i, it can imply the uncertainty of x^i, due to rotation invariance, it's more flexible if the data sets have $x_1^i = 0$ compared to model 1.

$p(D|M_3, \theta_3)$ model3 is a standard logistic regression with a bias weight which can account data sets with unequal distribution of +1 and −1.

If data set is not well modeled by any sharp linear boundary, none of models above is likely.

Question 25

It implies that in each dimension, θ_i is a Gaussian distribution with mean 0 and variance 10^3, the models with using θ correspond to a sharp linear boundary in x space, the models with a bias weight θ_i can account for data domain $y_i \in \{1, -1\}$, the models which add θ_i times x_i restrict the distribution of x_i, but cover the most data sets. However, if the variance of θ_i is too small, adding restrict to x_i will cause data loss, affect the data learning.

4 Conclusion

Having completed the Models assessment, we feel we now have more understanding of the underlying principles of machine learning and the process of building models. This understanding of the intuition behind the machine learning algorithms will help us make a more informed choice of model for a specific task, including model complexity and parameter choice, and get results that are more useful. We believe the purpose of this assessment was to offer a theoretical approach to machine learning that allows for a deep understanding of models and gives a good foundation to perform practical tasks in the future.

References

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.