Modelling

Ashwinder Khurana, Adam Sadiq

Bristol, United Kingdom

1. Question 1

What assumption does a Gaussian likelihood encode i.e what motivates the choice of this likelihood function?

We are aware that our data has been corrupted with some noise. We use an error function to model this, and assume that this function ϵ is normally distributed; based on this assumption, we know that the likelihood function is also normally distributed. This then allows us to model the likelihood function as a Gaussian distribution.

What does it mean when we have chosen a spherical co variance matrix for the likelihood, contrast with a non spherical case?

When you have a spherical co-variance matrix, it means that the variables you are measuring are completely independent from each other, whilst the non spherical case suggests that the variables you measure are dependent on each other. A spherical matrix usually takes the isotropic form of $\mathbf{B}^T \ \boldsymbol{I}$ where $\mathbf{B}$ is a scalar and $\boldsymbol{I}$ is the identity matrix.

2. Question 2

If we do **not** assume that the data points are independent how would the likelihood look then? Remember that $Y = [y_1, ...y_n]$

$p(y_1, ..., y_n|f, \mathbf{X}) = p(y_1|f, \mathbf{X})p(y_2|y_1, f, \mathbf{X})...p(y_n|y_{n-1}...y_1, f, \mathbf{X})$

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i \tag{1}$$

$where: \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2 I})$

3. Question 3

What is the specific form of the likelihood above, complete the right-hand side of the expression.

$p(\mathbf{Y} \mid \mathbf{X}\text{,}\mathbf{W}) = N(\boldsymbol{W}\boldsymbol{X}, \boldsymbol{\sigma^2 I})$

Preprint submitted to COMS30007 *29th October 2018*

4. Question 4

Explain the concept of conjugate distributions, why do they help us compute the posterior distribution?

Conjugate distributions are distributions that allow us to mathematically represent our belief for a specific parameter over a distribution, in a specific functional form. They are extremely useful to help us compute the posterior distribution because it allows us to know the functional form of the posterior before we even compute what the posterior. The posterior is proportional to the likelihood * the prior, and since the conjugate distribution allows us to fix the functional form of the posterior, we avoid the hard integration to calculate the evidence, that we would need for Bayes Theorem.

5. Question 5

Reason about the Gaussian distribution in this context, which distance function does it encode with a spherical co-variance matrix

A spherical co-variance is a diagonal matrix . In the Gaussian multivariate context, this means that the exponential equates to the euclidean distance between the data vector and the mean vector.

$\frac{1}{E}e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\boldsymbol{C}(\boldsymbol{x}-\mu)}$

where: $E = \sqrt{(2\pi)^M|C|}$

$C =$ An isotropic co-variance, in the form of :

$$\boldsymbol{B}^T I \tag{2}$$

6. Question 6

Write out the posterior over the parameters W. I recommend that you do these calculations by hand as it is very good practice and provides important intuitions. However, in order to pass the assignment you only need to outline the calculation and highlight the important steps. Justify the the posterior by providing an intuition of its form.

$$\begin{aligned} p(w|t) &= N(w|m_n, s_n) \\ where : M_n &= S_n(\tau^{-2}W_0 + \sigma^2 X^T Y) \\ , S_n &= (\tau^{-2} + \sigma^2 X^T X)^{-1} \end{aligned}$$

We begin, by having a general form prior with a Normal distribution over $\boldsymbol{w}$ with mean W_0 and co-variance τ^2.

M_n and S_n are the mean and co-variance respectively after seeing N data points. Intuitively reading the formula, if we have no data points, Sn will be reduced to sτ^{-2}, which equals our prior. However if we have an increasing amount of data points, β X^{T}X increases. Therefore when we take the inverse, our diagonals for the co-variance will become small. This is expected as the more data we have, the less our points co-vary, which is a consequence of incorporating the likelihood function.

7. Question 7

What is a non-parametric model and what is the difference between non-parametrics and parametrics? In specific discuss these two aspects of non-parametrics, Representation/parametrisation of data? Interpretability a of models?

Non parametric models are models that essentially throw away the idea of having parameters and we focus on encoding the relationship with how any new data point we have relates to the existing data that we have modelled, so in this case our "parameters" are simply the amount of data that we see. In comparison, parametric models force us to model the data by encoding an assumption in the form of a distribution and updating our parameters of these distributions to better the data. Linear regression is an example of updating the $\boldsymbol{W}$ parameters.

However, this problem may lead to an over-fitting of data. As we gather more data, our model becomes increasingly complex. So, if our model is already tightly fitted to existing data, when we receive new data that is dissimilar, our model will struggle to adapt to fit the new data. To deal with this, we have to add noise to the data, but finding the balance between adding too much noise and adding too little noise is a tricky task in itself because it could mean that we lose the general data pattern or follow the data too closely respectively.

8. Question 8

Explain what this prior represents and how it places structure on the space of functions?

With a Gaussian Process, our prior is formed over our assumption that all inputs are jointly Gaussian within the infinite input space. We can argue that for every point (x$_i$, f$_i$), we can use the kernel distance measure to calculate how another point (x$_j$,y$_j$) co-varies with our original point. Once we have this distance measure, we can use this to produce a Gaussian distribution where y$_j$ will be. Given the abstraction of x$_i$, we can apply this argument to the whole of the input space which therefore produces a tube-like prior, with the tube centred at an output of 0.

The structure this prior places on the space of functions is that it assumes that if we already have a data point, we know nothing about any new data points. This explains the tube-like shape, centred around a mean, for the prior because we have a fairly wide variance over the function space, where our uncertainty is not reduced.

9. **Question 9**

Does this prior encode all possible functions or only a subset?

The prior includes all possible functions. Because our prior produces infinite Gaussian as there are infinite input values, and since Gaussian in their nature are defined as being non-zero over an infinite space: we encode all of the output space - therefore all possible functions are encoded, just with varying probabilities.

10. **Question 10**

Formulate the joint distribution of the full model that you have defined above,

$$p(Y, X, f, \theta) \tag{3}$$

Draw the graphical model and clearly state the assumptions that has been made in bullet list.

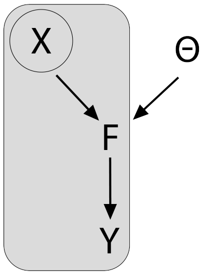


Figure 1: Graphical Model

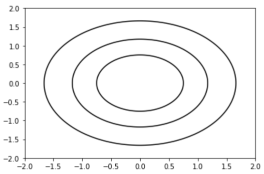
The assumption we make that every instantiation of a function follows a Gaussian distribution, which implies that the joint distribution of every instantiation follows a Gaussian distribution.

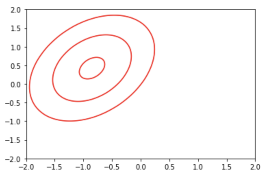
11. **Question 11**

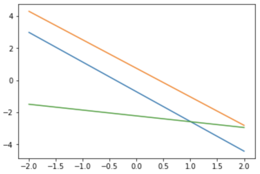
Explain the marginalization in Eq[2]

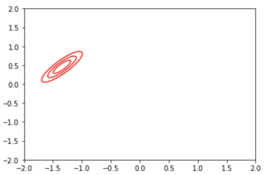
When we marginalize out over df, this means that we average out our beliefs over the space of all functions given the parameters X and θ. Since we have integrated, our likelihood finds the balance between the data that we have and incorporating our assumptions which connects our prior and the data. Our uncertainty is filtered through which is easy to visualize in the graphical model. The uncertainty is in our prior parameter θ and this parameter is used for our function f, so therefore the uncertainty is carried over. This occurs again because our function is a parameter for our data. Overall this means that θ is still a parameter we have to factor in, to produce a unique co-variance, to produce the marginalised likelihood.

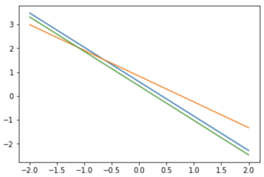
12. Question 12











5. Describe the plots, and the behavior when adding more data? Is this a desirable behavior? Initially, our prior has a generic mean at [0,0] and a huge variance. We chose an isotropic co-variance hence the circular shape in Figure 1, to represent the idea that the variables are assumed to be independent. Figure 2 shows the Co-variance after seeing 1 random data point. The posterior immediately changes, where our mean has shifted to be closer to the ”true” value of the our weight's function - showing that our model has learned to a degree as to what the parameters w_0 and w_1 could be, but still with a large variance as our prior is still a large factor. Figure 3 shows a few samples taken from the posterior, as it follows the general direction of our data, our model appears to be learning.

Figure 4 shows the posterior after seeing 199 data points. This is an extremely accurate co-variance as it is centered around the ”true value” of our W. This is desirable as it means that our model gets more sophisticated(complex?) and can predict more accurately with increased input data. Also, the samples taken from this posterior show something interesting. The samples diverge as the x value increases. This is because our marginal distribution's variance as to what the Y-intercept of our data is is much smaller than the variance of what our gradient value is, hence why the lines eventually diverge.

6. Relate to the expression of the posterior why you see the behaviour that you do when you add more data Since the expression follows the formula in Question 6, we can easily reason about the behaviour of the posterior. If we give our model no data, the model just believes that our prior is correct since $\beta X^T X$ and $\beta X^T Y$ are reduced to 0 which just leaves us with the prior mean and co-variance. However, as we have give more data to the model, $\beta X^T Y$ and $\beta X^T X$ increase. As a consequence, this means that our likelihood becomes a growing factor in comparison to the prior to produce our posterior, which then gives a more accurate updated belief.

13. Question 13

Explain the behavior of altering the length-scale of the covariance function.

Our squared co variance function calculates how ”close” two points are from each other. So if you increase the length scale l, this will increase the overall output value of the kernel function. This means that the points we are comparing in the kernel co-vary more, hence the smooth lines that it produces in our samples. This is because we have a very peaked Gaussian, so the sample functions change slowly, and that we are certain where x_j is, in relation to x_i. However, if we decrease the length scale, the kernel output is very small, implying that the points do not have strong co variance, so you would have functions that can change very quickly - hence the wavy lines in our samples

What assumption does the length scale encode?

The assumption is that between all points, there is always a non-zero co-variance. Even if you could use values with high orders of magnitude, there would still be some co-variance between data points.

14. Question 14

Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal co-variance matrix to the squared exponential?

15. Question 15

Elaborate on the relationship between assumptions, belief and preference.

An assumption is the first step we take in order to learn about something. When we take an assumption, we do not have anything to validate it. We state that one property of the data we obtain is true. Learning can only be done by through assumptions, as through our assumptions we generate priors, enabling us to learn via models incorporating these assumptions to generate priors. Belief is what we perceive to be true incorporating previous experiences. Within the context of machine learning, our experience are analogous to the parameters of our prior - we use these parameters to average out our prior. A preference is how we would like our model to represent and learn from our data. Much like the Type II maximum likelihood did. As we have integrated out the function, our preference is encoded, and it chooses a function that is a balance between being close to the observed data and matching our assumption.

16. Question 16

$$\textit{p(x) = N(0, I)} \tag{4}$$

What is the assumption/preference we have encoded with this prior?

Because it is a spherical Gaussian, we assume that the probability distribution has spherical (circular) symmetry - the covariance matrix is diagonal (so the off-diagonal correlations are 0), and the variances are equal. We assume independency.

17. Question 17

$$p(Y|W) = \int p(Y|X, W)p(X)dX. \tag{5}$$

Perform the marginalisation in Eq.2.1 and write down the expression. As previously, I do recommend that you do this by hand but to pass the assignment you only need to outline the calculations and show the approach that you would take.

$$p(Y|X,W) = \int p(Y|X)p(X|W)dX \tag{6}$$

We marginalise out the X value, since we are not interested it in per-say. We want to get the expected distribution of Y, given a W. We specify our belief in the function and marginalise it out, taking an average of the distributions. Computing this, we have an average of what Y is given W. We've taken all possible likelihoods and weighed them according to our prior distribution.

18. Question 18

How are ML, MAP, and Type II ML different?

These all try to find the optimal way to fit a distribution to the data by determining the optimal parameters of a model. Maximum Likelihood tries to do this via uniquely maximising the likelihood - it blindly trusts our data. Maximum a posteriori does so by taking into account the prior, and hence maximising the posterior. Type-II Maximum-Likelihood goes in-between. We integrate our belief in the function, giving a marginalised likelihood. Following this, we find the best fitting parameter to maximise it. This removes the blindly trusting of our data, yet also helps us navigate away from the difficulty marginalising all of our variables.

How are MAP and ML different when we observe more data?

When we observe more data, the MAP changes and it takes into account the prior, it will become more accurate and constrained to our data, whereas Maximum Likelihood does not change, it only focuses on maximising the likelihood wrt. parameters.

Why are the two expressions in Eq. 10 equal?

Because the Evidence always integrates to 1, maximising the LHS will always be equivalent as maximising the RHS.

19. Question 19

Write down the objective function log(p(Y—W)) = L(W).

20. Question 30