Modelling

Ashwinder Khurana - 36208, Adam Sadiq - 34795

Machine Learning

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 $\mathbf{B}\ \boldsymbol{I}$ where $\mathbf{B}$ is a scalar and $\boldsymbol{I}$ is the identity matrix.

Question 2

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

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

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

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

Question 3

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

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

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.

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 :

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

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.

$$\begin{aligned} p(w|t) &= N(w|m_n, s_n) \\ where: M_n &= S_n(\tau^{-2}IW_0 + \sigma^2 X^T Y) \\ , S_n &= (\tau^{-2} + \sigma^2 X^T X)^{-1} \end{aligned}$$

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} and M_n to $\tau^{-2}W_0$, which equals our prior. However if we have an increasing amount of data points, β X^{T}X increases.

This is expected because we want our machine to learn, so therefore as we churn through the formula with more X values, we want our likelihood function to have more influence and to shift our mean and co variance away from our prior belief which happens when X^TX increases in size, so we learn more from the data we have.

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 fit 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.

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, where our f represents a Gaussian distribution, and our co-variance being a kernel function. If we have seen no data points whatsoever, our mean function tells us that have no information where another data point can lie, making all intersections of the function and points equally likely. Given this abstraction, if we can apply this argument to the whole of the input space infinitely, we produce a tube-like prior, where every slice has the same distribution, around the same mean. However, with our co variance represented by a kernel function, it specifies our belief on how the functions should behave. We can argue that for every point (x_i, y_i) in our input space, we can use the kernel to calculate how another point (x_j,y_j) co-varies with it. Once we have this kernel, we can use this to produce a Gaussian distribution f to predict where for a given x, our y_j will be.

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 has having non-zero probability mass over an infinite space: we encode all of the output space - therefore all possible functions are encoded, just with varying probabilities.

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.

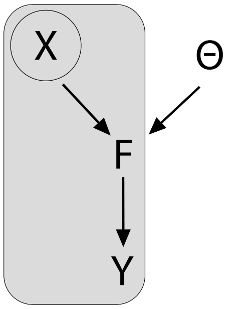


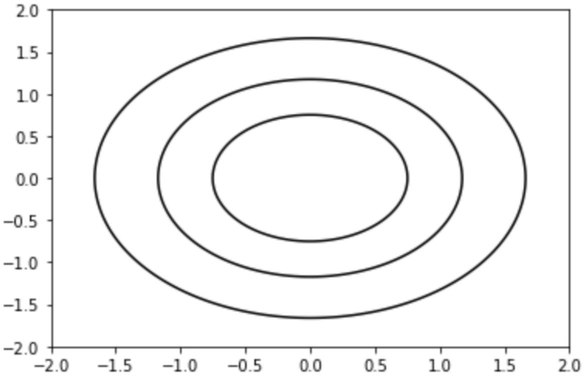
Figure 1: Graphical Model

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.

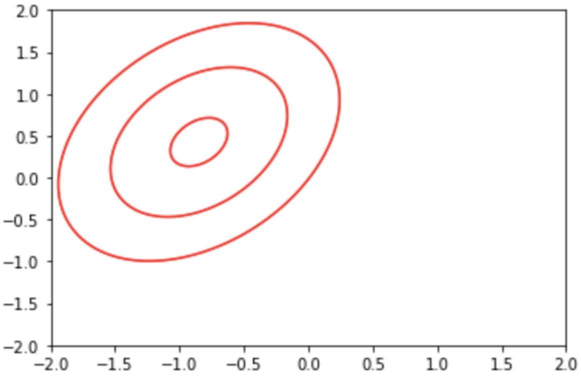
Question 11

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, to eventually encode our preferences from our beliefs. 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.

Question 12

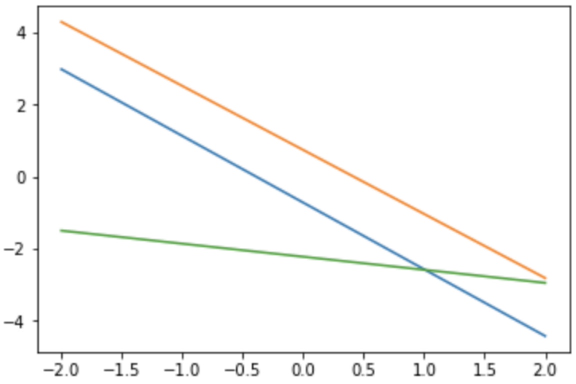


(a) Figure 2: Our prior

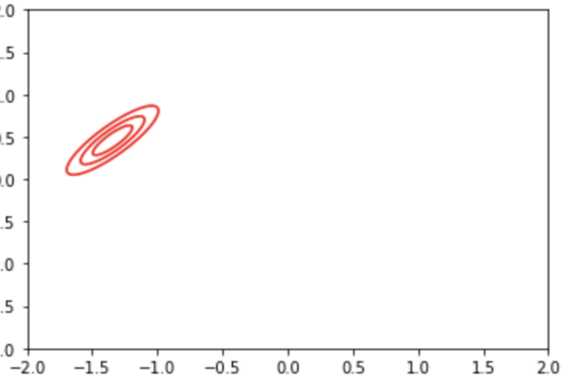


(b) Figure 3: Posterior after 1 data point

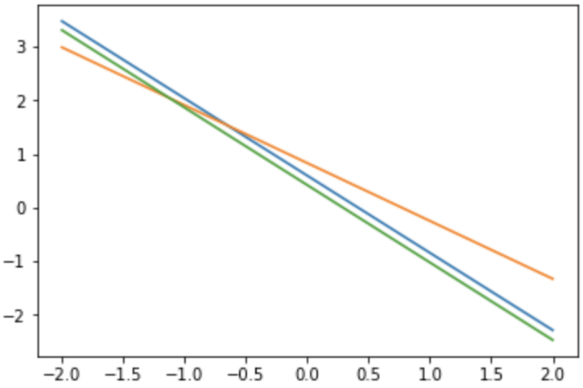
5. Describe the plots, and the behavior when adding more data? Is this a desirable behavior?
Initially, our prior has a generic mean at [0,0] and a huge variance. We chose an isotropic co-variance hence the circular shape in Figure 2, to represent the idea that the variables are assumed to be independent. Figure 3 shows the posterior 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 4 shows a few samples taken from the posterior, as it follows the general direction of our data, our model appears to be learning.



(a) Figure 4: Samples from posterior in figure 2



(b) Figure 5: Posterior after 199 data points



(a) Figure 6: Samples from Figure 4

Figure 6 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.

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 marginal 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?

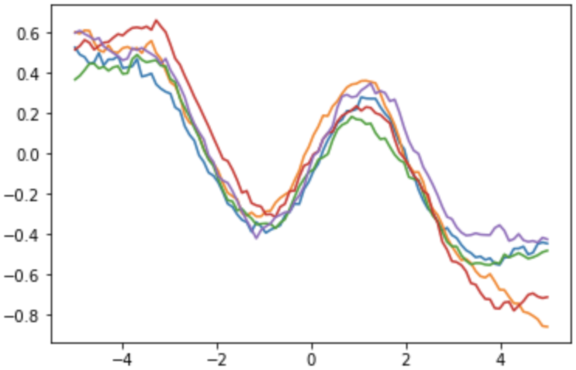
The assumption is that between all points, there is always a non-zero co-variance. Even if you could use values with high orders of magnitude, there would still be some co-variance between data points.

Question 14

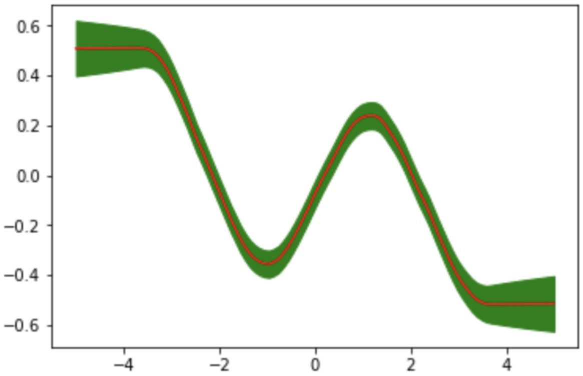
Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal co-variance matrix to the squared exponential?

If we sample from points that are far away from our initial data, our variance increases. This is due to our kernel function returning a low value co variance value, so we are more uncertain (smooth conditional Gaussian) about where the function value is at points that are far away, and we have peaked conditional Gaussian's for the sample points that are closer to our original data points because of the higher kernel function results. This behaviour is desirable because we are accurately able to predict the function value if it lies within our observed input domain, this does mean however, that we are not able to extrapolate well as points, that are not in our input space, have high uncertainties.

If we add a diagonal co-variance matrix to the squared exponential kernel matrix, our diagonal values would increase. This means that our data points that we have a higher variance at the observed data points.



(b) Figure 7: Samples from predictive posterior



(c) Figure 8: Plotting variance and mean

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.

Question 16

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

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 independence.

Question 17

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

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

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.

Because the gaussians are closed under linear transformations, we know that p(Y, X, W) / p(Y|W) is a gaussian distribution.

We must find the mean and covariance of this distribution via finding the mean and covariance: Mean: E[Y] = E[WX + ϵ]
As the expected value of ϵ and X is 0, E[Y] = 0
Covariance: E[yy^T] = E[(WX + ϵ])(WX+ϵ)^T]
= E[WXX^{T}w^T] + E[$\epsilon\epsilon^T$] Since the covariance of the noise is $sigma^2 I$, and the x has covariance of an identity matrix, ww^T = I, we can simplify our answer as:

$p(y|w)\ N(0, E[ww^T + \sigma\ I])$

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, even though the Maximum a posteriori takes into account the prior, it converges to be equal to the maximum likelihood.

Why are the two expressions in Eq. 10 equal?

Because in our Evidence, we are integrating out W. Therefore maximising over W with or without the evidence will always give the same result.

Question 19

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

$$\mathbf{L(W)} = constant + log|\mathbf{C(W)}| + \sum_{N}^{i} y_i^T(\mathbf{C(W)})^{-1}\mathbf{y}_i \tag{7}$$

Write down the gradients of the objective with respect to the parameters $\frac{\delta L}{\delta \mathbf{W}}$

$$tr(\mathbf{C}^{-1}(\frac{\delta \mathbf{C}}{\delta \mathbf{W}_{ij}})) + tr(\mathbf{Y}\mathbf{Y}^{T}(-\mathbf{C}^{-1}\frac{\delta \mathbf{C}}{\delta \mathbf{W}_{ij}}\mathbf{C}^{-1})) \tag{8}$$

Question 20

Marginalisation of f is much simpler to do than marginalising out X, the latter is actually in most cases analytically intractable. Provide a simple reason why this is, the argument should be general about marginalisation and not about this model in specific. A good idea is to draw the graphical model and then follow the arrows (hint hint).

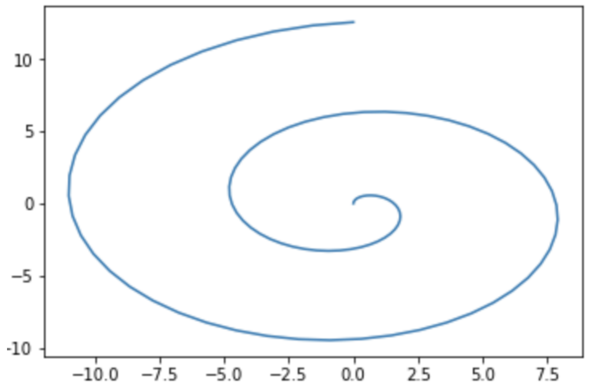
Given the complexity in the relationship with f and x, the marginalisation over x is very difficult because the calculation of the integral with respect to x is difficult. However, the relationship between f and y is much simpler to marginalise out, so we use marginalise out y for the maximum marginal likelihood.

Question 21

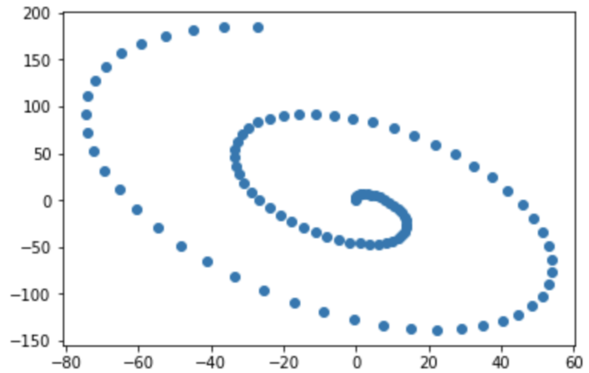
Plot the representation that you have learned. Explain why it looks the way it does. Was this the result that you expected? Hint: Plot X as a two-dimensional representation.

We are mapping **x'** to a 10 dimensional space using the matrix A. The objective function represents our error, so we use the gradient descent to optimise and minimise this objective function through a number of iterations. This iterative optimisation changes the values of the mapping of x' to an x*. So when we plot x* in Figure 10, it doesn't look like our initial x' in Figure 9.

We do get a spiral like curve, but it seems rotated and scaled in some way in comparison to our original x' Figure 9 plot. This means that our learned x* Figure 10 plot can be mapped back to the original x' with a 2x2 matrix consisting of a rotation R and scale S - (R · S). This makes sense intuitively as our minima should be unchanged by a rotation or a scaling.



(a) Figure 9: Our initial data

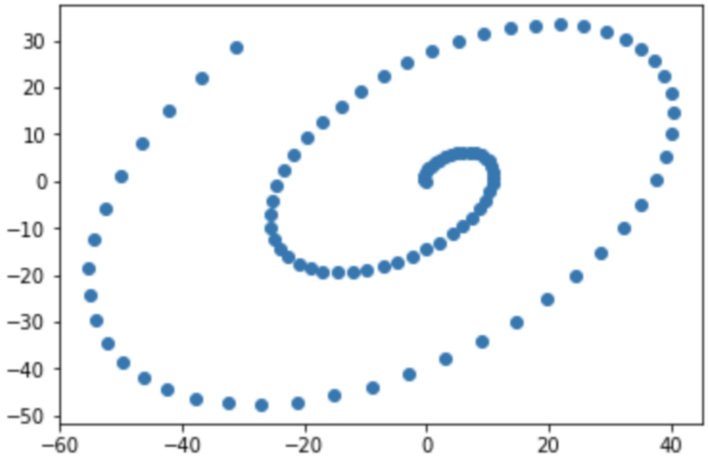


(b) Figure 10: Learned representation

Question 22

Draw a random two dimensional subspace (does not have to be an orthogonal basis) and plot the data. How is this result different compared to the subspace that you learnt? Provide a justification for the result.

Plotting a random 2 dimensional subspace gives us a curve that looks like the result of our x* Figure 11 plot. This shows us that, it is not a problem that the result of the optimization gives us a plot that looks dissimilar to our original data. It is just our original data under a rotation and a scaling, because any 2x2 matrix with full rank is a rotation and a scaling.



(a) Figure 11: Random 2D Subspace Plot

Question 23

$$p(D|M_{0,\,0}) = \frac{1}{512} \tag{9}$$

What does this assumption actually imply? Make an argument for why,

0.1. This is the simplest possible model

Because the model represents everything with the same probability mass, it could be argued that this is a simple model. For all parameters entered into the model, it comes out with the same probability and result. The conditionals between different parameters are eventually non-existent

0.2. this is the most complex model

However, taking into account Occam's Razor, this model represents everything in the data space with probability mass over the whole model. In Occam's Razor it acts much like M3, modeling all of the data domain. In addition, it is very likely that this would not be chosen as a model, as it argues there would always be a simpler model.

Question 24

Explain how the each separate model works? In what way is this model more or less flexible compared to M_0? How does this model spread its probability mass over D? How have the choices we made above restricted the distribution of the model? What data sets are each model suited to model? What does this actually imply in terms of uncertainty? In what way are the different models more flexible and in what way are they more restrictive? Discuss and compare the models to each other?

The model seems to be a conditional Gaussian, where the different models are made up of one input parameter - a point in the vector space along with a weight. The second model uses two points as parameters, and two weights, and the third, three parameters and three weights. The models are more flexible than M_0 as they take into account more data points, and more complexity. In relative terms, it seems to be more flexible in representing, where M_0 seems is static, especially how it spreads its probability mass. The first model seems very restrictive, as it it only incorporates one single data point in it's calculations. It seems that it would be very general and not very accurate at all, alike to a large length scale in the Gaussian Process. The second seems to be better, taking into account more data points, and even more for the third, however, because it is more complex than the other models, an argument can be made that it may be guilty of overfitting data. In this case, a case can be made that the first model is less restrictive than the 3rd model. However, in reality, complexity is not a global measure, it is relative always according to another standard of measure. It depends on the data that you want to represent that changes the choice of model you would use.

Question 30

Summarise the assignment in one paragraph, what have you learnt and what do you feel have been the prupose/message of performing this.

We believe that the purpose of the assignment was to essentially to begin to formalise the way we learn as humans in order to start implementing machines that learn. Intuitively, we learn about a topic via incorporating and weighting our prior beliefs about something and refining this belief by seeing new events.

Within the context of this coursework, we have used various different models like linear regression, non parametric regressions to find different ways to model our study of data. Whilst these may be basic models relative to other complex models we have not seen, the conceptual process that we have gone through will remain the same.