Course Work 1 - Models

Samia Mohinta(97293) Tanya Sharma(26131) Tushar Bhatnagar(97607)
Srijan Singh Chhabda(97201) Jasmeet Khalsa(97313) Paola Brito(97340)

01 November 2018

1 The Prior

Question 1

1. When we assume that the random variables are mutually independent and are identically distributed (i.i.d.), we take a Gaussian Likelihood into account. We believe that these random variables follow the Central Limit Theorem and form a normal distribution having a specific mean and a variance. Gaussian is conjugate to itself. This helps in predicting the functional form of the posterior. If likelihood and prior are Gaussian, the posterior is also going to be a Gaussian.
2. A spherical co-variance matrix is of the form $M = \beta.I$, which is a diagonal matrix with all diagonal elements having the same value. This encodes X and Y are mutually exclusive. The probability mass is evenly distributed in a spherical co-variance matrix (circular symmetry). A non-spherical co-variance matrix is chosen when there is a dependency between the random variables. In non-spherical co-variance, probability mass is unevenly distributed (like an ellipsoid).

Question 2

If it is assumed that the output is not independent, then the likelihood should look like:

$$p(Y|f,X) = p(y_1, y_2, y_3, ...y_N|f,X) \tag{1}$$

$$p(y_N|y_{N-1}, y_{N-2}, ..., y_1, f, X)p(y_{N-1}|y_{N-2}, y_{N-3}..., y_1, f, X)....p(y_1|f,X) = \prod_{i=1}^{\infty} p(y_i|x_i \cup y_i) \tag{2}$$

Question 3

Assuming that the i.i.d. observations are corrupted by an additive Gaussian noise. The likelihood can be defined as :

$$P(Y|X,W) = \prod_{i}^{N} p(y_i|x_i, W) = \prod_{i}^{N} \mathcal{N}(Wx_i, \sigma^2 I) = \mathcal{N}(W^T X, \sigma^2 I) \tag{3}$$

Here the likelihood function is written as the Gaussian distribution function so that the form of the likelihood, prior and hence the posterior is the same.

Question 4

The concept of conjugate distribution means that the functional form of the posterior and the prior belong to the same family of distributions. The prior is then called a conjugate prior to the likelihood. We know that the product of two Gaussian functions gives us a Gaussian function (self-conjugate). Thus, the posterior has the same functional form as the prior which therefore leads to a simplified Bayesian Analysis. With conjugacy we can avoid computing the evidence in Bayes'theorem.

Question 5

Euclidean distance is a special case of the Mahalanobis distance with equal variances of the variables and zero co-variances.

$D_M(x) = \sqrt{\sum_i^N (x-\mu)^2/\sigma^2}$, where σ^2=variance and $D_M(x)$ = Standardized Euclidean distance

Since a spherical co-variance matrix has zero co-variance, it can be said that the Gaussian distribution referred in question 5 encodes the Euclidean distance function.

Question 6

We are given,

$$p(W) = \mathcal{N}p(W_0, \tau^2 I)$$

We have to find $p(W|X, Y)$.
Therefore,

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

$$= \frac{1}{Z}\mathcal{N}(WX, \sigma^2 I)\mathcal{N}p(W_0, \tau^2 I)$$

We can ignore the $\frac{1}{Z}$ as of now, because it is just a normalizing constant, called the Evidence.
So, we represent the multivariate pdf as under :

$$p(W|X, Y) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-1}{2\sigma^2}(Y-XW)^T(Y-XW))}$$

$$. \frac{1}{\sqrt{2\pi\tau^2}} e^{(\frac{-1}{2\tau^2}(W-W_0)^T(W-W_0))}$$

We equate the exponents.

$$p(W|X, Y) \propto e^{(\frac{-1}{2\sigma^2}(Y-XW)^T(Y-XW))} . e^{(\frac{-1}{2\tau^2}(W-W_0)^T(W-W_0))}$$

After expanding and re-arranging the above equation.

$$p(W|X, Y) \propto e^{[-(\frac{1}{2\sigma^2}(Y^TY) - \frac{1}{2\tau^2}(W_0^TW_0) - \frac{1}{2\sigma^2}W^TX^TXW - \frac{1}{2\tau^2}W^TW + \frac{1}{\sigma^2}Y^TXW + \frac{1}{\tau^2}W^TW_0]}$$

The quadratric part of the above equation is used to get the co-variance.

$$C = -\frac{1}{2}W^T(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)W$$

Hence, co-variance

$$\sum = (\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1}$$

Taking the linear term to solve for μ, we get

$$\mu = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1}X^TY$$

Therefore our posterior distribution now takes the form with co-variance and μ

$$p(W|Y, X) \propto \mathcal{N}(W|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1}X^TY, (\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1})$$

Question 7

A parametric model is used to describe a probability distribution having a specific form governed by a number of parameters. In order to make a prediction,knowing the values of parameters is sufficient. Non-parametric models are used to describe those probability distributions where it's hard to identify the form of the function and no explicit assumptions can be made. It relies on training data to make better predictions.

This is typically useful when there is a mixture of Gaussian or even a 3-4 order equation. Non-parametric models are usually said to be infinite dimensional (more degrees of freedom) and can hence express data better than parametric models. Non-parametric models are more flexible than parametric models and make better predictions as they are based on observing data. For a parametric model, the assumptions about the data are fixed and the model makes future predictions depending on these fixed assumptions. A parametric model is represented as: $Y = mx + c$ (linear regression). A non-parametric model is represented as: $Y = f(x) + \epsilon$ where $f(x)$ is some function over the data.

Question 8

With a GP, we define a prior over functions and want our prior to put some constraints over the space of the functions. The Gaussian process assumes that $p(f_1, f_2.., f_n)$ is jointly Gaussian for the set of data points $x_1, x_2..x_n$ with a zero mean and a co-variance k. k(X,X) is the co-variance matrix giving the inner product between the data points. The kernel function that determines K is typically chosen to express the property that for points x_i and x_j that are similar, the corresponding values $f(x_i)$ and $f(x_j)$ will be more strongly correlated than for dissimilar points [1]. We no longer need to know what the function looks like, just the inner product provided by the kernel will help us compute the posterior from the prior easily.

Question 9

The prior encodes all the properties and is not a subset. In a GP every function has some probability (recalling the pipeline example from lectures) at each point. If our argument for GP existing at a one point is valid, we can say it is valid at all points. Even the probability mass at infinity is something. Hence all possible functions are included as a GP encodes a probability distribution at each point.

Question 10

The joint probability distribution can be formulated as :

$$p(Y, X, f, \theta) = p(Y|X, f, \theta)p(f|X, \theta)p(X)p(\theta) \tag{4}$$

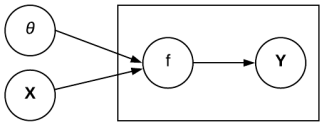


Figure 1: Graphical Model for the joint distribution

The graphical model assumes X and θ to be independent variables and hence the joint probability of them is the product of their marginals. f is conditionally dependent on the values of X and θ and hence is written as $P(f|X, \theta)$. Y is conditionally dependent on the values of f (not directly dependent on X and theta).

Question 11

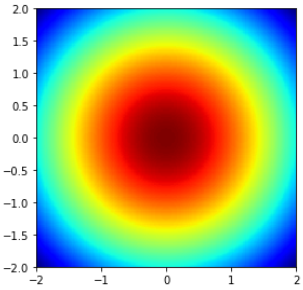
We marginalise to get rid of the extra term f. Intuitively we can say that to define our output we no longer need the function f and therefore f can be marginalized out.

As we see from the graphical model in Figure 1, we can say that initially uncertainty flows from X and *theta* to f and then to Y. However, when we marginalise over f, we remove Y's dependency on f, thereby, removing the uncertainty contributed by f to Y (filtering out uncertainty). After marginalisation of f, we can have our Y depending on the data and the parameters and these can now contribute directly towards Y's uncertainty.

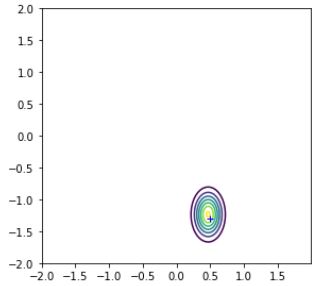
θ on left implies that *theta* we need to know its value to find Y and it has not been marginalised out. It is a hyper-parameter, which is a parameter of a prior distribution and is independent.

Question 12

The more data we observe, the lesser is the uncertainty. This behavior is desirable because it says that our model is learning the parameters from the data. Thus the values of the parameters gradually tend towards the actual (true) values. We use the previous posterior as the new prior for every new data-point sample. With increasing number of data-points, the posterior is updated and can therefore predict the output of a new data-point when multiplied with the likelihood. With infinite data-points samples, the posterior distribution centers on the true parameter values.



(a) Prior assumption (contour plot)



(b) Posterior using 200 sample points

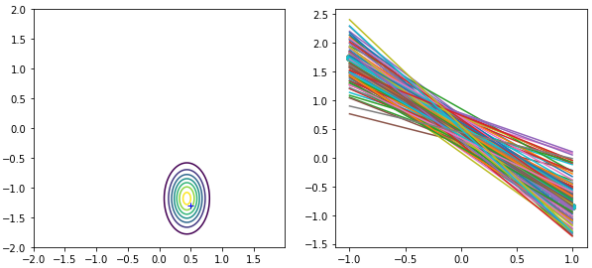


Figure 3: Sampled for 100 data points

Question 13

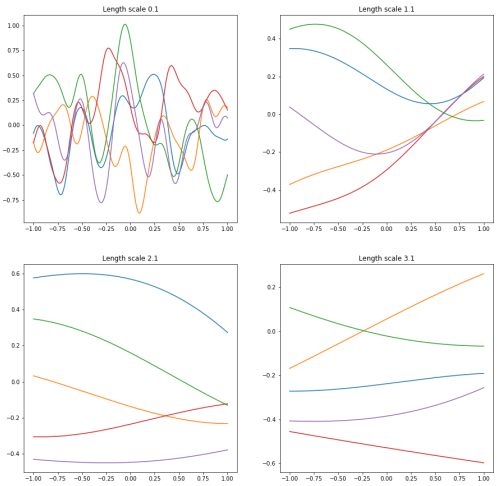
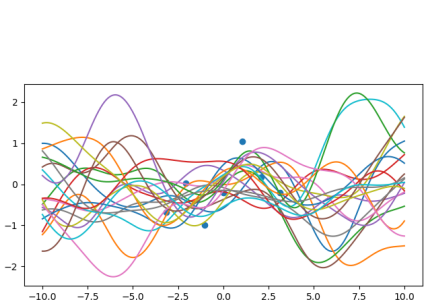


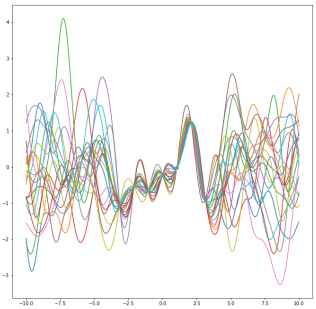
Figure 4: Plots with length-scale 0.1, 1.1, 2.1 and 3.1

The length-scale can be treated as a hyper parameter and encodes our belief over the smoothness of the underlying function. That is whether we believe that the function changes quickly or it changes slowly. When the length scale is small, we can see from the graph the functions are extremely wiggly (length-scale=0.1), however, when we increase the length scale to 1.1, 2.1 and 3.1, the functions become smooth and do not change rapidly.

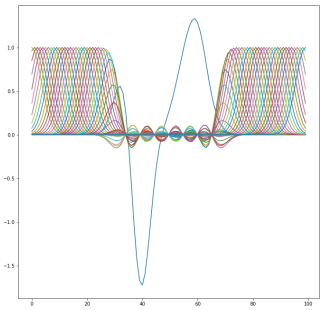
Question 14



(a) Predictive Posterior with Diagonal Covariance (length scale=0.1)



(b) Predictive Posterior Without Diagonal Covariance



(c) Predictive Mean and Variance of the Posterior

Figure 5: Predictive Posterior Distribution

When we club our prior with the data we get the posterior. The uncertainty in the posterior reduces as we observe more data. We can also see that the variance increases when we feed data that is out of bounds from our observed data. We have an assumption that if the points x1 and x2 are close, their corresponding y values will also be close. Our posterior encodes this assumption for the 7 data points we plot. Fig 5(b) is without the additive noise. Fig 5(a) is the plot after we have added the diagonal co-variance matrix to the squared exponential. The diagonal matrix adds uncertainty to the data. So now, the posterior does not exactly pass through the data points.

2 The Posterior

Question 15

Belief is based on observed data and is known truth. Preferences are usually fixed and externally given. Assumption is a starting point in modelling. We make some assumptions based on the beliefs we already have, and among those assumptions we prefer one over another to suit more to the context. For example, when we are given a set of data points we assume two or three functions to pass through that set based on a prior belief of seeing something like that. Now, among these set of functions, we prefer a linear equation more over a polynomial to pass through the data points. This is our preference over assumptions.

Question 16

We have made a preference of the prior being a zero mean Gaussian and assumed that the random variables are independent and identically distributed, therefore having an Identity matrix as co-variance.

Question 17

We assume to have an additive Gaussian noise for our linear non-parametric GP. Therefore, our likelihood is :

$$p(y_i, x_i, W) = \mathcal{N}(Wx_i, \sigma^2 I) \tag{5}$$

We need to marginalise X out of the below equation :

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

We multiply our Gaussian $p(Y|X,W)$ and $p(X)$ to get the Gaussian $p(Y|W)$.

$$p(Y|W) = \int \mathcal{N}(y_i|Wx_i, \sigma^2 I)\mathcal{N}(0,I) \tag{7}$$

$$p(Y|W) = [\frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2\sigma^2 (y-xw)^T (y-xw)}} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2(x-0)^T (x-0)}}]dX \tag{8}$$

We equate the exponents to get our quadratic term denoted by C.

$$C = -1/2\sigma^2 (x_i W)^T (x_i W) - 1/2 x_i^T x_i \tag{9}$$

$$C = -\frac{1}{2}[\frac{1}{\sigma^2} W^T x_i^T x_i W + x_i^T x_i] \tag{10}$$

$$C = -\frac{1}{2} x^T x[\frac{1}{\sigma^2} w^T w + 1] \tag{11}$$

Therefore our co-variance is

$$Cov = [\frac{1}{\sigma^2} W^T W + 1] \tag{12}$$

$$Cov = [W^T W + \sigma^2 I] \tag{13}$$

We next use the above co-variance to find μ.
Mean and co-variance after equating and re-arranging the exponents are :

$$\mu = 0 \qquad \sum = [W^T W + \sigma^2 I] \tag{14}$$

Therefore,

$$p(Y|W) = \mathcal{N}(0, [W^T W + \sigma^2 I]) \tag{15}$$

Question 18

ML when attached with a prior leads to MAP. ML is more of a frequentist approach, which becomes a special case of Bayesian MAP, when we assume a uniform prior for our MAP. Type II maximum likelihood is the maximization of the marginal likelihood over the variables (X) not required for the estimate.

When a MAP observes more data, the prior is overwhelmed with the data seen. That is, the estimates from MAP tend to match the estimates from ML.

The two expressions in (10) are equal because the denominator is a constant and can be ignored.

Question 19

The objective function can be written as in [2]:

$$L(W) = constant + log|C(W)| + \sum_i^N y_i^T (C(W))^{-1} y_i \tag{16}$$

In order to find the gradient of the objective function, it has to be differentiated by $\frac{\partial L}{\partial W} a$. The following matrices properties are used to solve for the gradient [2]:

$$\frac{\partial C}{\partial W_{ij}} = \frac{\partial W W^T}{\partial W_{ij}}$$

$$\partial(XY) = (\partial X)Y + X(\partial Y)$$

$$\frac{\partial X_{kl}}{\partial X_{ij}} = \delta_{ik}\delta_{lj}$$

$$\partial X^{-1} = -X(\partial X)X^{-1}$$

Therefore we can write the derivative of WW^T as:

$$\frac{\partial WW^T}{\partial W_{ij}} = W\frac{\partial W^T}{\partial W_{ij}} + \frac{\partial W}{\partial W_{ij}}W^T = WJ_{ij} + J_{ij}W^T \tag{17}$$

The $\partial log|X|$ term can be written as:

$$\partial log|X| = tr(X^{-1}\partial X) \tag{18}$$

The derivative of this equation can written as:

$$\frac{\partial}{\partial W_{ij}} log|C| = tr(C^{-1}\frac{\partial C}{\partial W_{ij}}) \tag{19}$$

This means that our second term becomes,

$$\frac{\partial}{\partial W_{ij}} tr(Y(C)^{-1}Y^T) = tr(\frac{\partial}{\partial W_{ij}} Y(C)^{-1}Y^T) \tag{20}$$

Using Chain rule we get:

$$tr(\frac{\partial}{\partial W_{ij}} Y(C)^{-1}Y^T) = tr(\frac{\partial}{\partial C}(YC^{-1}Y^T)\frac{\partial C^{-1}}{\partial W_{ij}}) = tr((YY^T)^T\frac{\partial C^{-1}}{\partial W_{ij}}) \tag{21}$$

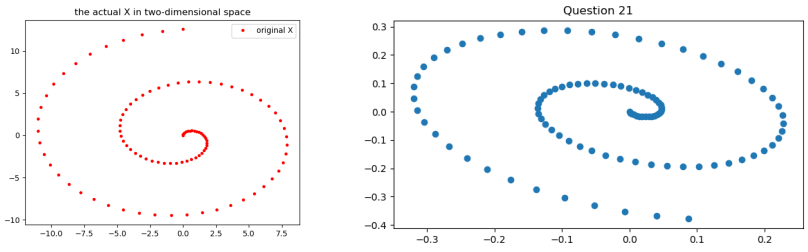
Now using derivative of a matrix inverse:

$$tr((YY^T)^T\frac{\partial C^{-1}}{\partial W_{ij}}) = tr(YY^T(-C\frac{\partial C}{\partial W_{ij}}C^{-1}) \tag{22}$$

Question 20

Note: The graphical model is presented in Figure 1. If we trace the arrows in the graphical model, f lies closer to Y, i.e., it is one arrow shorter than reaching X and hence marginalisation of f is much simpler(in terms of mathematical computation).

Question 21



(a) Representation of Original X (b) Learned X

Figure 6

Here we perform Type II maximum likelihood to optimise the W. Recall, knowing W means knowing X.

The output data is 10-D and therefore it's co-variances matrix will be 10x10. To generate a 10-D vector from a 2D space we need a matrix of the form 10x2. We know that the output after final derivation of matrix with a scalar is 10x10. Hence for this to work, we need [10x2][2x10]+[10x2][2x10]. W can be thought of as 2 basis vectors in 10-D space. Figure 6(b) shows us the projection of the 10-D data as recovered latent representation.

In the above figures, we have plotted the original X values against the latent variable X which we have learned from observed data Y (100x10) following a maximum likelihood estimate. The plot looks like this because we are plotting the multi-dimensional non-linear function in a 2D space. When we try to recover X from Y by calculating the weights, we cannot exactly get the values of X through the optimisation. This is because we marginalised out the latent space and optimised for the mapping.

Question 22

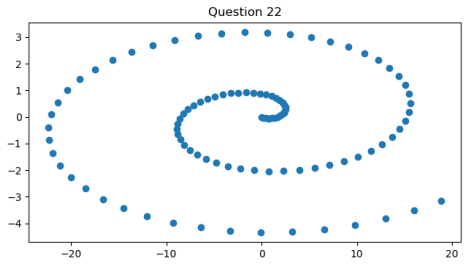


Figure 7: 2D subspace plot for question 22

This question required drawing a random two dimensional subspace and plotting the data. The main difference that we can see from this plot and the one we had from the previous question, is that the previous one followed a random normal distribution. From the plot we can see how the 20 random numbers that we have plotted appear, where although the centre values seem to be at (0,0), we can see how the values across the x-axis stretch from -15 to 20, and the y values from -20 to 20 across the y-axis. We can therefore see how the data plotted is not centred around a mean of zero like the plot in the previous question and the dimensions of the data have stretched significantly further.

3 The Evidence

Question 23

The assumption implies that all the data sets are equally likely because they have the same probability. It is a simple model because it has no free parameters and it assigns a uniform probability over the space. On the other hand, it can also be viewed as a complex model because it is not flexible, assigning same probability to a number of possibilities (utilizes no information from data sets except its cardinality).

Question 24

Model M3 is the most complex, most flexible and least uncertain model of all the given 4 models. It is a standard logistic regression. It is complex and less uncertain because it uses more parameters and it can realize other models by setting some of its parameters to zero. It is flexible because it can spread its probability mass over a wider range of data sets than the other models. Model M2 can be realized by setting the θ_3^3 parameter to zero. This model is the next most complex model after M3. M1 is the same as M2 when it ignores the second dimension of x. In context of the literature [3], model M3 gives higher probability to Data set (a) due to the bias term. Data set (b) is best captured by M1, because in this data set, the decision boundary is a function of x1 and not x2. M2 models the decision boundaries of data sets (c),(d) and (g). Data sets (e) and (f) are again favored by M3 due to the bias term. Finally, data set (h) is best modelled with the uniform model M0.[3]

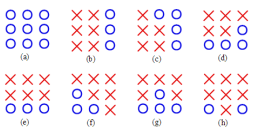


Figure 8: Literature reference model [3]

Overall, among all the 4 models, M0 is the simplest and M3 is most complex and certain. However, M0 is the least flexible, while M3 is most flexible. M1 is restrictive towards data sets that are functions of x1, while M2 can capture decision boundaries without the bias.

Explain the process of marginalisation. Discuss its implications.

Marginalization over θ here implies that we are eliminating the dependence of the probability distribution on θ. The evidence of this model will be a weighted sum and the weights shall be defined by a prior distribution of θ given the model.

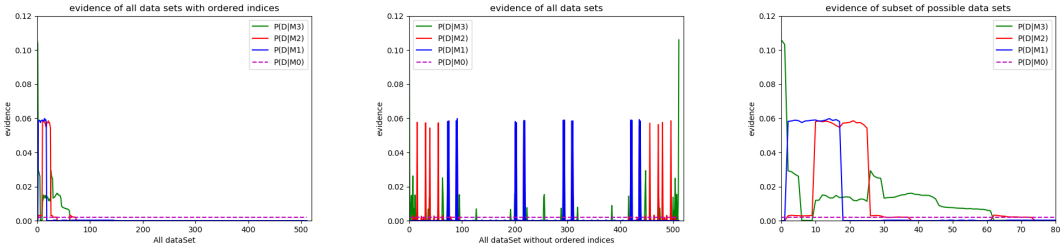
Question 25

We are completely uncertain of the parameters. By having a zero mean, we believe the probability distribution should center around zero. The co-variance is a diagonal co-variance matrix, i.e. the parameters have zero co-variance between them and are i.i.d. . The large variance of 10^3 will allow the parameters to vary in large range of values and hence the model leads to sharp linear boundaries in the data space.

Question 26

The sum of evidence for models M0, M1, M2 and M3 are 1, 0.99633341, 0.99962157 and 0.99094757 respectively. All of these sums are 1 or tending to 1 for the models because these are all probabilities distributions of weight and the sum of each probability distribution is ideally 1. The sum of evidence is 1 for M0 because it assigns uniform probability across all data sets. The sums for the other 3 models lack the accuracy of having an integral value of 1 because we have used an approximation technique (Monte Carlo).

Question 27



(a) Ordered (b) Unordered (c) Subset

Figure 9: Plot of evidence over the whole data set for each model

We can see from the ordered plot Fig 9(a) that model M0 has put a uniform probability mass across all the data sets. This is expected behavior. Model M3 has the highest evidence centered on specific data sets because it has 3 parameters including a bias term. M2 and M1 have the same evidence peaks for the data sets. However, M1 spreads across a few more data sets because of having one parameter less than M2 for defining the dimensions of X. The data sets defined by M2 and M1 can also be defined by M3 setting some of its parameters to 0.

Question 28

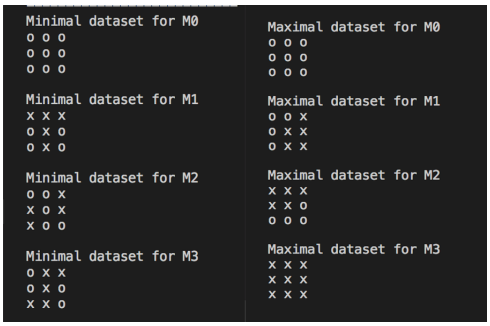
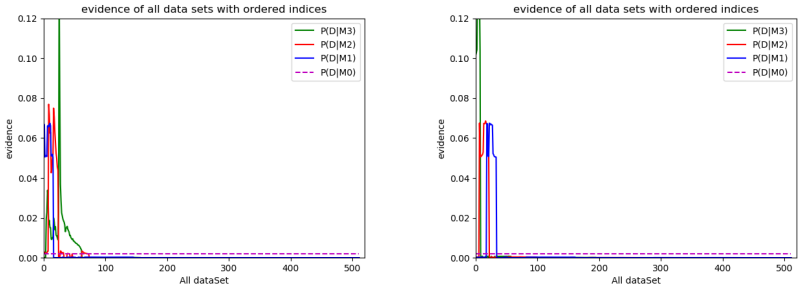


Figure 10: Datsets with Highest and Lowest Evidence of each model

Here we had to plot the data set which gives the minimal and maximal probability for each model in how it can classify between 1s and -1s. Let's begin with the maximal. We can see that the maximal for M0 consists of a data point with all Os, which is due to how the model treats all data equally where it returns 1/512. The maximal data-set for M1 is such because the model itself relies on x1, where we can see that along the x1 axis it is possible to differentiate between 1s and -1s. The maximal for M2 is explained by how the model uses x2 and not just x1. Here we can see that along the x2 axis you can differentiate between 1s and -1s. The maximal for M3 is explained where it includes an extra parameter, which it uses to shift from the data in essence to draw a line and classify. The justifications for the minimals are the converse of the above.

Question 29



(a) Prior with Non-Zero Mean (b) Prior with Non-Diagonal Covariance

Figure 11: Evidence (ordered) plots with changed prior parameters

We have changed the mean to [5; 5]. The peaks have shifted from the plots in question 27. Earlier for M1, we had evidence below 0.06 but with the mean changed it has more evidence for first 20 data-sets. The change in the mean has led to a change in the probability of the parameter for M1 to take more positive values now and hence this motivates to draw better decision boundaries for the first 20 data-sets. We have a similar case for M2 and M3 as well. However, for M0, which has no free parameters, the plot remains the same. When we use a non-diagonal co-variance matrix for the prior, we see that there is larger correlation between the parameters. This correlation also increased the peak of the evidence we have for each model (except M0).

Question 30

This course-work consisted of both theoretical and practical questions. The sections are sequentially laid out to help us learn in a systematic way. We are given the opportunity to go through and understand the theory part of each section first and then apply that on practical problems. We have gained hands-on experience at supervised, unsupervised, parametric and non-parametric machine learning problems. We have also learnt the nuances of model selection and how to integrate our beliefs with our observations as well as apply uncertainty and assumptions in our models. This course-work has equipped us with foundational knowledge of machine learning so that we can now apply this knowledge on real problems.

References

[1] Christopher Bishop. *Pattern Recognition and Machine Learning*(2006)

[2] K.B. Petersen and M.S. Pedersen *The Matrix Cookbook* (2012)

[3] Iain Murray and Zoubin Ghahramani. *A note on the evidence and Bayesian Occam's razor*. Technical Report GCNU-TR 2005-003 (2005).