KTH Royal Institute of Technology

DD2434
Machine Learning Advanced Course

Year 2017 - 2018

Assigment 1

Individual report

Picó Oristrell Sandra 940531 – 2308 sandrapo@kth.se

Professors :
Herman Pawel
Kjellström Hedvig
Lagergren Jens

December 3, 2017

Contents

1 I The Prior $p(\mathbf{X})$, $p(\mathbf{W})$, $p(f)$

1.1 Theory

Question 1*: Why Gaussian form of the likelihood is a sensible choice? What does it mean that we have chosen a spherical covariance matrix for the likelihood?*

The Gaussian, or also called the *Normal distribution*, is one of the most important models used for continuous variables distribution.The reason why it is one of the most known is because it can be used in different contexts and perspectives. One situation in which the Gaussian fits is when we consider the sum of multiple random variables.[1]

As we learned studying Machine Learning, we use to face with a lot of situations where we have uncertainty, where we do not have any previous information about some data. That is one of the reasons why choosing a Gaussian distribution is a sensible choice.

The *Central Limit Theorem*(**CLT**), *"establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions"*.[2]

In other words, in practice, as the number of variables in the sum increase, the distribution will converge to a Gaussian.[3]

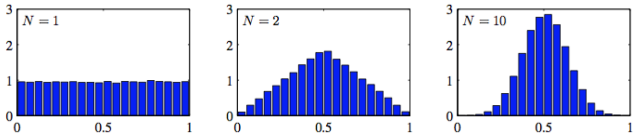


Figure 1: Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian. (Pattern Recognition and Machine Learning page 79)

At the same time, we also have to take into account that Gaussian distribution is mathematically friendly because only two parameters needed to be defined; the variance($\sigma^{\mathbf{2}}$) and the mean (μ).

Specifically, the Gaussian distribution can be written in the form: [1]

$$\mathcal{N}(x|\mu),\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \tag{1}$$

On the other hand, a spherical or *isotropic* co-variance matrix is a matrix that is proportionate to the identity matrix, *"i.e it is diagonal and all elements on the diagonal are equal"* [4].For this reason, in terms of the likelihood, it implies that all the terms included in **T** do not have any correlation between them. In other words, they are independent.[5]

$$\mathbf{C} = \lambda\mathbf{I} \tag{2}$$

***Question 2**:If we do **not** assume that the data points are independent how would the likelihood look then? Remember that* $\mathbf{T} = [\mathbf{t}_1,...,\mathbf{t}_N]$

As we have seen in the description of this question, if each output point is conditionally independent given the input and the mapping, the likelihood of the data will be:

$$p(\mathbf{T}|f, \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{t}_i|f, \mathbf{x}_i). \tag{3}$$

For this reason, if we do not assume that the data points are independent, applying the *Chain rule*(Product rule)[6], we are going to obtain:

$$p(\mathbf{T}|f, \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{t}_i|f, \mathbf{X}, t_{i-1}, ..., t_1). \tag{4}$$

The formula above state that every data point will take into account the previous data points that appeared before.

Nevertheless, it should also be kept in mind that in that case, computationally talking, it is more expensive, because the co-variance matrix is no longer symmetrical.

1.1.1 Linear Regression

In this section we are going to take into account two assumptions.

The first assumption is about the mapping and the model; the relationship between them variate as a *linear mapping*.

The second assumption is regarding the structure of the noise in the observations. The structure of the additive noise will be *the Gaussian*. We are assuming that they have been corrupted by noise.

For the two assumptions stated above, we have the following relationship:

$$\mathbf{t}_i = \mathbf{W}\mathbf{x}_i + \epsilon, where \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{5}$$

***Question 3**: What is the specific form of the likelihood above, complete the right-hand side of the expression in (6).*

In this question we are trying to calculate the likelihood of the data given the previous assumptions already stated before. For this reason, as we said, the target variable $\mathbf{t}_i$ is given by a linear model with additive Gaussian noise.

Then, considering that **T** represents the group of t_1,...,t_N and making the assumption, for simplicity, that these data points are drawn independently from the distribution, the likelihood will be defined by the following equation[7]:

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{W}x_n, \sigma^2 I) \tag{6}$$

***Question 4**: The prior in Eq.8 is a spherical Gaussian. This means that the "preference" is encoded in terms of a L2 distance in the space of the parameters. With this view, how would the preference change if the preference was rather encoded using a L1 norm? Compare and discuss the different type of solutions these two priors would encode.*

Before answering this question we should know that the equation 8 indicates us that regarding the different choices that we could made for the prior $p(\mathbf{W})$, the most sensible choice is to pick the following Gaussian prior,

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{W}_0, \tau^2\mathbf{I}) \tag{7}$$

However, considering the question about the difference between L1 and L2 norm, we should start talking about regularization. According to Bishop, *regularization is "one technique often used to control the over-fitting phenomenon, which involves adding a penalty term to the error function in order to discourage the coefficients from reaching large values"*.[8] There are many possible choices of regularization, but a common regularization is the LP vectors norms, with different values of P. In this question we are going to talk about L1, also known as Lasso. In the Pattern Recognition and Machine Learning book, Bishop show the main difference between L1 and L2 through the following image:

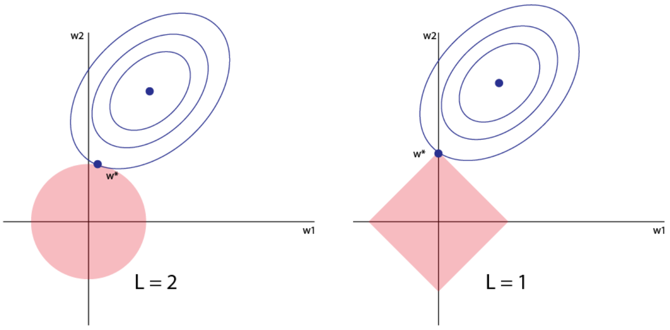


Figure 2: Plot of the contours of the unregularized error function along with the constraint region.

As we are able to see in the picture, with L2 norm the regularization term is a circle while with L1 norm the term is a rotated square. In the figure above we can see that the combination of the unregularized error function (blue) and the regularizer (red) will be minimized at some point that touches both surfaces. Nevertheless, L1 penalties have the effect that they can encourage sparse parameters vectors. The sparser the more parameters will be exactly zero, effect that not happens with L2.[9] With L2, some weights may tend to zero but really difficult to be exactly zero.

***Question 5:** Derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. However, in order to pass the assignment you only need to out- line the calculation and highlight the important steps. You can make derivations for individual samples ($\mathbf{x}_i$,$\mathbf{t}_i$) and then generalize to the data set or operate on matrices keeping the concept of vectorization in mind.*

- *Briefly comment/discuss the form (mean and covariance).*
- *What is the effect of the constant Z, are we interested in this?*

Based on: [10]

Since we know that the Gaussian Distribution is part of the exponential group of distributions we know that the property of *conjugate prior* is applied [11]. In other words, if the prior and the likelihood are Gaussian distributions, this will imply that the our posterior will have a Gaussian distribution as well. For this reason, we can express our posterior as:

$$P(W|X, T) \sim exp\frac{1}{2}(w-\mu_w)^T\sum^{-1}(w-\mu_w) \tag{8}$$

where we obtain three different terms; the quadratic term, the linear term and the constant term.

1. Quadratic term:

$$exp(\frac{1}{2}w^T(\sum_w)^{-1}w) \tag{9}$$

2. Linear term:

$$exp(w^T(\sum_w)^{-1}\mu_w) \tag{10}$$

3. Constant term:

$$exp(\frac{1}{2}\mu_w^T(\sum_w)^{-1}\mu_w) \tag{11}$$

Based on the answers given in the question 3, we can approximate our posterior as:

$$P(W|X, T) \sim exp\frac{-1}{2\sigma^2}(t-XW)^T(t-XW)exp\frac{-1}{2\tau^2}(W-W_0)^T(W-W_0) \tag{12}$$

because we know that the likelihood and the prior have the following form.

Likelihood:

$$P(T|X, W) = \mathcal{N}(Wx_i, \sigma^2 I). \tag{13}$$

Prior:

$$P(W) = \mathcal{N}(W_0, \tau^2 I). \tag{14}$$

Then,if we focus on the exponent, we can also divide it in the terms mentioned before; quadratic term, linear term and constant term.

1. Quadratic term:

$$-\frac{1}{2\sigma^2}(XW)^T XW - \frac{1}{2\tau^2}W^T W \tag{15}$$

2. Linear term:

$$+\frac{1}{\sigma^2}(XW)^T t + \frac{1}{\tau^2}W^T W_0 \tag{16}$$

3. Constant term:

$$-\frac{1}{2\sigma^2}t^T t - \frac{1}{2\tau^2}W_0^T W_0 \tag{17}$$

Using the equations stated before, we can formulate the variance over the parameters W as the following equation:

$$\sum_w = (\frac{1}{\sigma^2}X^T X + \frac{I}{\tau^2})^{-1} \tag{18}$$

And then, knowing the equation of the variance, we can also obtain the mean over the parameters W:

$$\mu_w = (\frac{1}{\sigma^2}X^T X + \frac{I}{\tau^2})^{-1}(\frac{1}{\sigma^2}X^T t + \frac{I}{\tau^2})^{-1} W_0 \tag{19}$$

On the other hand, the parameter Z is a normalization factor in order to achieve that the posterior sums up to 1 taking into account all the values of **w**.

1.1.2 Non-parametric Regression

In the previous task we made the strong assumption that the relationship between the two variables was linear. However, this assumption severely restricts the representative power of our model.

If we think about how a Bayesian would think in the actual situation, we can just say that we have a large uncertainty in the parameters of the mapping and also in the form of it.

For this reason, as the Bayesian allows us to deal with this scenario, we just need to formulate our uncertainty about the mapping in a prior over mapping and then use Bayes to reach the posterior. We are going to use Gaussian Processes.($\mathcal{GP}$s)

Question 6: *Explain what this prior does? Why is it a sensible choice? Use images to show your reasoning. Clue: use the marginal distribution to explain the prior.*

A Gaussian process ($\mathcal{GP}$) is a non-parametric model that represents a prior over a space of continuous functions. As the description of the assignment explain, using Gaussian Processes, the prior can be formulated as:

$$p(f|\mathbf{X}, \theta) = \mathcal{N}(\mu(x), k(\mathbf{X}, \mathbf{X})) \tag{20}$$

In the equation above, we could identify two different parameters: the mean function and the co-variance function [k(.,.)]. In that case, our mean function is zero. On the other hand, the θ represents the *hyperparameters* of the processes.

If we analyze the equation above we realize that the equation express the prior credence of the distribution that has created the data. In other words, it helps us about the mapping of f. On the other hand, we could say that is a sensible choice by the simple reason that determines which maps of f are more likely to represent the data.

Nevertheless, the sense of using $\mathcal{GP}$ is because it helps to marginalize out the functions. Assigning the mass to an arbitrary and finite set of points, we can discover the structure hidden in the data. The locations of these points will allow us, at the same time, to be able to make the average of all the possible functions and thus to know the structure of the co-variance function. This function will be used by $\mathcal{GP}$s to calculate the later.

Then if we use the marginal distribution to explain the prior, we can state that the marginal equation is the following[12]

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{f})p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{C}) \tag{21}$$

where the co-variance C has the elements:

$$C(x_n, x_m) = k(x_n, x_m) + \sigma^2 \delta_{nm} \tag{22}$$

The equation written above reflects clearly the concepts explained before; the two Gaussian sources of randomness are independent and that is the reason why their co-variances simply add (as we can see in the previous equation).

Question 7: *Formulate the joint likelihood of the full model that you have defined above*

$$p(\mathbf{T},\ \mathbf{X},\ f, \theta)$$

(Try to draw a very simple graphical model to clearly show the assumptions that you have made.)

The joint probability for the $\mathcal{GP}$ can be formulated as the following [13]:

$$p(\mathbf{T}, \mathbf{X}, f, \theta) = p(\mathbf{T}|f)p(f, \mathbf{X}, \theta)p(\mathbf{X})p(\theta) \tag{23}$$

To continue, it has to be also state that the following term, the prior over f, is given in the description of the assignment:

$$p(f|\mathbf{X}, \theta) = \mathcal{N}(\mathbf{0}, k(\mathbf{X}, \mathbf{X})) \tag{24}$$

Finally, considering the two next assumptions:

- Target features in the matrix **T** are independent.
- Between them θ and **X** are independent.

We can re-write the joint probability as:

$$p(\mathbf{T}, \mathbf{X}, f, \theta) = p(\mathbf{T}|f)\mathcal{N}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}))p(\mathbf{X})p(\theta) \tag{25}$$

Finally, to represent the previous final equation in a simple graph, we can use the next illustration[14]:

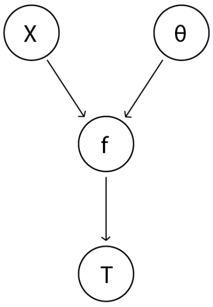


Figure 3: Graphical model for the joint probability using GP regression

Question 8: *Complete the marginalization formula in Eq.12 (general form) and discuss,*

- *Explain how this connects the prior and the data?*
- *How does the uncertainty "filter" through this?*
- *What does it imply that θ is left on the left-hand side of the expression after marginalization?*

It has to be clear that our goal is to find the probability that relates the targets T and the inputs X. For this reason, we need to marginalize the formula of the equation 12. In other words, we need to marginalize the mapping of *f*. That is because we really do not have interest in *f*.

To achieve the marginal distribution we need to integrate over f:

$$P(\mathbf{T}|\mathbf{X},\theta) = \int P(f|\mathbf{T})P(\mathbf{T}|\mathbf{X},\theta) \tag{26}$$

The previous equation connects the prior and the data through the average of the likelihood of each data point. We can also state that, using the property of conjugate priors, we can affirm that the result will also form a Gaussian Distribution.

On the other hand, the uncertainty in this distribution is given by two sources of Gaussian uncertainty. The first one, is related to the uncertainty in the relationship between the two variables **X** and **T** and the second one represents the uncertainty in the targets, **t**. We have to consider that we do not know about the noise that could have influence over it.

Nevertheless, these two forms of uncertainty could be defined by the next terms:

$$\text{p}(f|\mathbf{X},\theta) \text{ and } \text{p}(\mathbf{T}|f)$$

Finally, regarding to θ, the fact that the parameter is in the left-hand side of the expression after marginalization is because the hyper-parameters still have influence over f and also because it is a constant. The fact that the hyper-parameters still have influence over f also implies that they have influence over the final **T**.

1.2 Practical

1.2.1 Linear Regression

In this task we will implement the linear regression that we looked at in the previous theoretical task. To examine prior and posterior over the parameters **W** we will need to have some data.

In order to generate some data we will use the following parameters:

$$t_i{=}w_0x_i{+}w_1{+}\epsilon$$

$$\mathbf{x} = [-2,-1.98,...,1.98,2]$$

$$\epsilon{\sim}\mathcal{N}(0,0.2)$$

$$\mathbf{W} = [1.5,-0.8]$$

Question 9:

1. *Set the prior distribution over* **W** *and visualize it.*
2. *Pick a single data-point from the data and visualize the posterior distribution over* **W**.
3. *Sample from the posterior and show a couple of functions.*
4. *Repeat 2-3 by adding additional data points.*

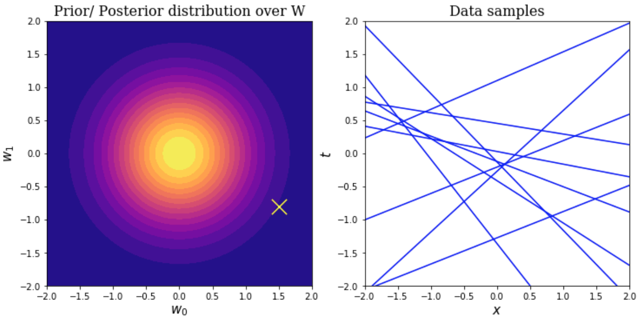
Describe the plots and the behavior when adding more data? Is this a desirable behavior? Provide an intuitive explanation..

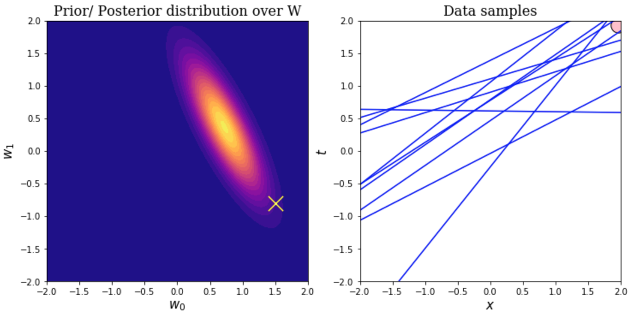
In the following images we demonstrate that as the number of the observations increase, the posterior becomes less uncertain [15].

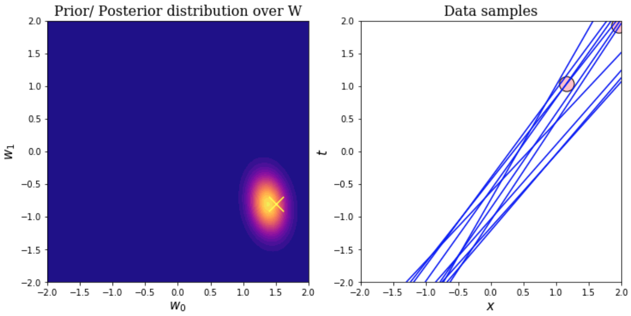
Each of the rows of the image corresponds to one of the different iterations of the code. And, regarding the columns, the first one represents the posterior over **W** and the second one show us the data samples space. As we said before, in each iteration, this data space is increasing. Specifically, we are evaluating four cases; four iterations.

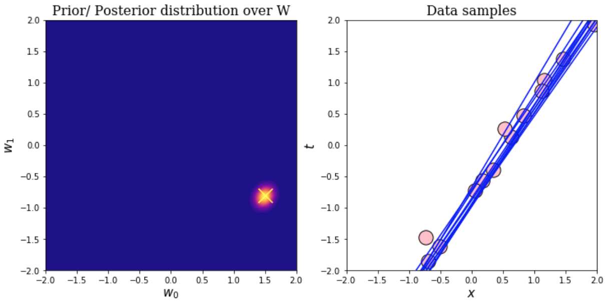
The first row of the next figure corresponds to the situation before any data points are observed. The row shows a plot of the prior distribution over **w** and the data space.

The second row is with one single data point; you can saw it painted in purple in the right-hand of the image. The third one is with two data point and, in the last row, the forth one, we can saw 20 different data points. If we analyze the data space, we can also conclude that as more data points we have in our analysis, more accurate become the linear regression model.









Finally, if we also look at the mean and the covariance of the posterior, we can analyze how they are changing in each iteration of the code:

First row:

$$\mu_{posterior} = \begin{bmatrix} 0. & 0. \end{bmatrix}$$

$$\sum\nolimits_{posterior} = \begin{bmatrix} 0.5 & 0. \\ 0. & 0.5 \end{bmatrix}$$

Second row:

$$\mu_{posterior} = \begin{bmatrix} 0.77070904 & 0.39325923 \end{bmatrix}$$

$$\sum\nolimits_{posterior} = \begin{bmatrix} 0.39839076 & -0.1991337 \\ -0.1991337 & 0.10973798 \end{bmatrix}$$

Third row:

$$\mu_{posterior} = \begin{bmatrix} 1.47533876 & -0.95989706 \end{bmatrix}$$

$$\sum\nolimits_{posterior} = \begin{bmatrix} 0.01978647 & -0.00198291 \\ -0.00198291 & 0.00707556 \end{bmatrix}$$

Forth row:

$$\mu_{posterior} = \begin{bmatrix} 1.51328772 & -0.83239684 \end{bmatrix}$$

$$\sum\nolimits_{posterior} = \begin{bmatrix} 0.00217967 & 0.0003826 \\ 0.0003826 & 0.00175874 \end{bmatrix}$$

In that point we can conclude that it is the desirable behavior based in two simple facts; the mean is close to **W** = [1.5,-0.8] and the co-variance has small values.

Finally, to sum up, we can highlight that as more observations we have, the more accurate representation we obtain and this is the desirable behaviour to be able to predict probabilities of different values of **t** for a given **x** based on the data points.

1.2.2 Non-parametric Regression

Question 10:

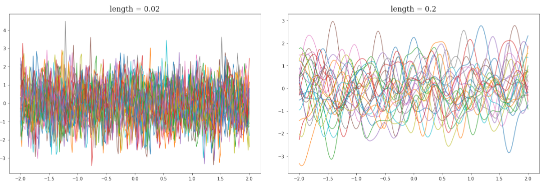
1. *Create a GP-prior with a squared exponential co-variance function.*
2. *Sample from this prior and visualise the samples.*
3. *Show samples using different length-scale for the squared exponential.*

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

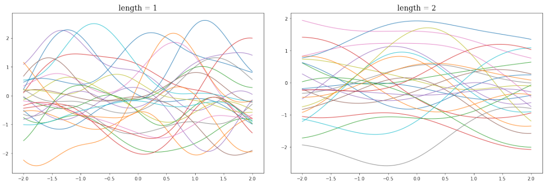
In this question, we used a squared exponential covariance function as kernel, as the description of this practical part indicates.

It has to be clear that, as the description indicates, we have two parameters that we could modify; the length-scale **l** and the variance $\boldsymbol{\sigma^2}$. The length-scale, as its name indicates, modifies the length-scale over the variations of the main function in the horizontal axis. On the other hand, the variance $\boldsymbol{\sigma^2}$ is in charge of controlling the variations in the vertical axis.

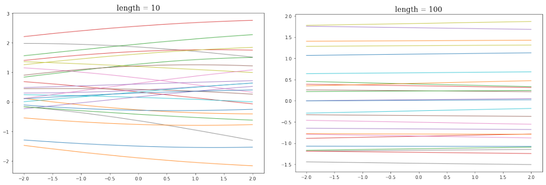
Using the Gaussian Processes (**GP**) and modifying the length of the covariance function stated before, I plotted the following figure:



(a) Length = 0.02 (b) Length = 0.2



(a) Length = 1 (b) Length = 2



(a) Length = 10 (b) Length = 100

As it can be observed, I basically plot six different images using the next different length-scales: [0.02,0.2,1,2,10,100]. We also need to state that our variance σ^2 was fixed to the value = 1.

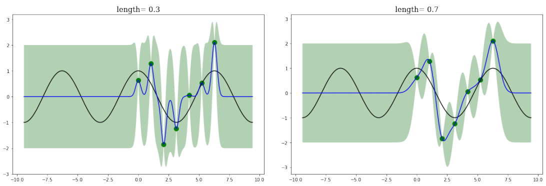
Iterating over these values, we can conclude that as we increase the value of the variable **l**, representing the length-scale, the samples from the Gaussian Processes become more soft. At the same time, we can state that as the length-scale increase, all the samples closer to the main function turn to be more correlated.

Question 11:

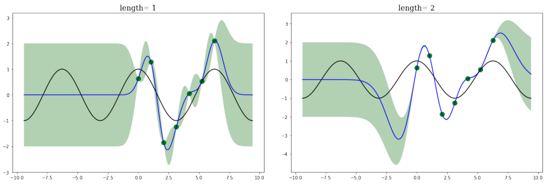
1. *How do we interpret the posterior before we observe any data?*
2. *Compute the predictive posterior distribution of the model.*
3. *Sample from this posterior with points both close to the data and far away from the observed data.*
4. *Plot the data, the predictive mean and the predictive variance of the posterior from the data.*

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 covariance matrix to the squared exponential?

In order to be able to answer question number 11, I decided to plot the following four images changing the value of the length-scale as I did for the question 10. Nevertheless, in this question, the values picked up for the length-scale are these ones: [0.3,0.7,1,2].



(a) Length = 0.3 (b) Length = 0.7



(a) Length = 1 (b) Length = 2

However, first of all, it has to be clear that in this exercise, we used the next statements:

$$y_i = \text{cos}(\text{x}_i) + \epsilon_i$$

$$\mathbf{x} = [0,...,2\pi]^T$$

$$\epsilon_i \sim \mathcal{N}(0, 0.5)$$

In other words, we are using a non-parametric regression with $\mathcal{GP}$ to a data-set of 7 data points y_i. As defined above, y_i is defined by the equation : $y_i = \text{cos}(\text{x}_i) + \epsilon_i$.

In the previous figure, the green points defined the seven observations requiered in the description. On the other hand, the black curve represents the function *cos*.Then, the mean of the $\mathcal{GP}$ of the predictive distribution is represented by a blue line and the painted green zone corresponds to the standard deviation, defined by the interval [-2,2].

Therefore, if we rely on the previously defined information and observing the green shaded zone, we can conclude that the farther we are from the 7 observations, the more uncertainty grows.

Regarding to the influence of a diagonal co-variance added to the kernel function, we could said that it will influence in the distribution. The conditional distribution will not pass through the data points as we saw in the other figures. This basically will happen because adding this co-variance matrix is producing the same effect than adding Gaussian noise to all the **t** values.

2 II The Posterior $p(\mathbf{X}|\mathbf{Y})$

In this second task we are going to look at representation learning. The notation used will be more consistent with the one used in most of the literature.

Y will denote output data instead of **T** used in the previous task.

In this section we only observe the outputs **Y** and we want to learn input X that can represent Y. Representation learning allow us to recover the needed parameters directly from the data. In other words, this relates to building a model of the data Y and then looking at the posterior distribution over the input to the model X. In this task, *learning* will be also introduced.

2.1 Theory

In this part we will focus on the same linear models as in the first part. However, input locations **X** are not known a priori. These locations will be referenced as the *latent representation* of the observed data **Y**.

The linear model will be the following one:

$$p(\mathbf{Y},\mathbf{X},\mathbf{W}) = p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})\mathring{u}p(\mathbf{W}). \tag{27}$$

At the same time, prior over the latent variables will be specified as a spherical Gaussian:

$$p(\mathbf{X}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{28}$$

***Question 12**: What type of “preference” does this prior encode?*

As we stated before in the description of the task, our prior is defined by a spherical Gaussian. For this reason, we are defining a preference on the shape of **X**. In other words, it is defining the structure f the input data.

Because the distribution is the spherical Gaussian, that also implies that there is no interaction between the dimensions and, as the second parameter of the distribution is defined by the identity matrix (**I**), all the correlations between dimensions are 0. That is, all the inputs [$x_1 ... x_N$] are independent.

At the same, as we described in the equation, **Y** will have zero mean.

***Question 13**:*

*Perform the marginalisation in Eq. 23 and write down the expression. As pre- viously, it is recommended that you do this by hand even though you only need to outline the calculations and show the approach that you would take to pass the assignment. **Hint:** The marginal can be computed by integrating out* **X** with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of **Y**(X).

Before starting with the question, we need to define that the equation 23 is the following one:

$$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|,\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X}. \tag{29}$$

The mapping from the latent variable to the observed variable is described by the next equation:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mu + \epsilon where \epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I}) \tag{30}$$

For this reason, based on the previous formula, the likelihood can be defined as:

$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \mathcal{N}(\mathbf{Y}|\mathbf{W}\mathbf{X},\sigma^2\mathbf{I}) \tag{31}$$

To continue, and to achieve the marginal likelihood distribution, we basically integrate the latent variable **X** using also the prior. We have also to state that this marginal distribution is Gaussian distribution by the simple fact that it corresponds to a linear-Gaussian model.

$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X} \tag{32}$$

Following the same argumentation as before, knowing that the marginal is Gaussian, we can analyze the covariance and the mean by this way:

$$\mathbb{E}[\mathbf{Y}|\mathbf{W}] = \mathbb{E}[\mathbf{WX} + \mu + \epsilon] = \mathbf{W}\,\mathbb{E}[\mathbf{X}] + \mathbb{E}[\mu] + \mathbb{E}[\epsilon] = \mu \tag{33}$$

$$cov[\mathbf{Y}|\mathbf{W}] = \mathbb{E}[(\mathbf{WX} + \epsilon)(\mathbf{WX} + \epsilon)^T] = \mathbf{WW}^T + \sigma^2\mathbf{I} \tag{34}$$

It has to be also clear that the mean and the covariance were achieved taking into account that ϵ and **X** are uncorrelated because they are independent.

Finally, we can state that the marginal distribution is:

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} p(y_i|\mathbf{W}) \tag{35}$$

Because we know that for each data point we have the next equation:

$$p(y_i|\mathbf{W}) = \mathcal{N}(y_i|\mathbf{WW}^T + \sigma^2\mathbf{I}) \tag{36}$$

2.1.1 Learning

By the moment, we have only looked at the posterior after creating a model but now we are going to learn the parameters of the model. We are going to start learning through a probabilistic model with the maximum-likelihood (ML) approach.

Using the maximum-likelihood approach, we will want to find the parameters of the likelihood that maximize it:

$$\hat{\mathbf{W}} = argmax_{\mathbf{W}} p(\mathbf{Y}|\mathbf{X}, \mathbf{W}). \tag{37}$$

To continue, we need to know about maximum-a-posteriori (*MAP*) estimation. This imply that we need to find the paramaters that maximize the posterior distribution:

$$\hat{\mathbf{W}} = argmax_{\mathbf{W}} \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})d\mathbf{W}} = argmax_{\mathbf{W}} p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W}) \tag{38}$$

And also an in-between stage which is called Type-II Maximum-Likelihood which implies maximisation of the marginal likelihood where you integrate out one parameter and then maximize over another:

$$\hat{\mathbf{W}} = argmax_{\mathbf{W}} \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X}. \tag{39}$$

Question 14*: Compare these three estimation procedures above in log-space.*

- *How are they different?*
- *How are MAP and ML different when we observe more data?*
- *Why are the two last expressions of Eq.25 equal?*

To answer properly the questions above we are going to follow the order stated before. First, we are going to talk about the three estimation procedures. Next, we are going to compare MAP and ML when we observe more data and, finally, we are going to analyze the equation 25 described in the assignment description.

To start comparing the two estimation approaches (ML and MAP) we are going to define the negative log function for each respectively:

- **ML** :

Maximum Likelihood estimation finds the values of **W** that maximize the opportunities of having observed the determined outputs **Y**.This is the same than minimizing the RSS[16]:

$$\mathcal{L}(\mathbf{W}) = \prod_{i=1}^{N} p(y_i - x_i, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}p(y_i|\mathbf{W}x_i + \mu, \beta^{-1}|\mathbf{I}) \tag{40}$$

We can transform that in the log-space as follows:

$$l(\mathbf{W}) = \frac{\beta}{2} \sum_{i=1}^{N} ||y_i - \mathbf{W}x_i - \mu||^2 + constant \tag{41}$$

- **MAP** :

Maximum-A-Posteriori estimation define the parameter distribution **W**, which the posterior function achieve its maximum as follows:

$$\mathcal{L}(\mathbf{W}) = p(w) \prod_{i=1}^{N} p(y_i - x_i, \mathbf{W}) = \mathcal{N}p(\mathbf{W} - 0, \alpha^{-1}\mathbf{I}) \prod_{i=1}^{N} \mathcal{N}p(y_i - \mathbf{W}x_i + \mu, \beta^{-1}\mathbf{I}) \tag{42}$$

In log-space:

$$l(\mathbf{W}) = \frac{\beta}{2} \sum_{i=1}^{N} ||y_i - \mathbf{W}x_i - \mu||^2 + \frac{\alpha}{2} \sum_{k} ||w_k||^2 + constant \tag{43}$$

If we evaluate MAP and ML when we observe more data we are going to find that MAP converges to the ML estimation equation. This is basically because, as more data we have, less important will be the term that differ from the two equations. Specifically, the following term:

$\frac{\alpha}{2} \sum_k ||w_k||^2$

This is because, the more data we observe, the prior distribution will have less importance and, on the other side, the posterior distribution will have more.

Finally, regarding the equation 25, the two terms are equal by the simple fact of the Baye's rule. The Baye's theorem tell us that:

$$p(\mathbf{A}|\mathbf{B}) = \frac{p(\mathbf{B}|\mathbf{A})p(\mathbf{A})}{p(\mathbf{B})} \tag{44}$$

If we applied this formula to equation 25, we can state that:

$$\frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})d\mathbf{W}} = Bayes = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{p(\mathbf{Y}|\mathbf{X})} \tag{45}$$

Nevertheless, as we are evaluating the *argmax* over **w** and the term *p*(**Y**|**X**) do not has any influence over w, that is why the equation 25 is defined as in the description.

In practice, when we have to perform optimization on probabilistic models we often have to deal with exponential. However, sometimes, exponential numbers are a bit tricky to play with. For this reason, rather than working directly with exponents we perform the learning in the log-space. That is why in practice, we often formulate the optimization problem as a minimization of the negative log of a probability:

$$\hat{\theta} = argmax_{\theta} p(\mathbf{Y}|\theta = argmin_{\theta} - log(p(\mathbf{Y}|\theta)) \quad (46)$$

***Question 15**:*

1. *Write down the objective function log(p(Y|W)) = $\mathcal{L}$(W) for the marginal distribution in Eq. 23.*
2. *Write down the gradients of the objective with respect to the parameters*

1. *Calculate the log likelihood for the marginal distribution:*

For each of the data samples we can state that:

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(y_i|\mu, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \quad (47)$$

We can finish the previous equation as:

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} \frac{1}{2\pi^{\frac{D}{2}} |C|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{y}_i - \mu)^T C^- 1(y_i - \mu)), \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}. \quad (48)$$

For this reason, based in the previous results, we are able to calculate the log(p(Y|W)) as:

$$logp(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_i|\mu, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}). \quad (49)$$

$$logp(\mathbf{Y}|\mathbf{W}) = \sum_{i=1}^{N} ln \frac{1}{2\pi^{\frac{D}{2}} |C|^{\frac{1}{2}}} exp(-\frac{1}{2}(y_i - \mu)^T C^{-1}(y_i - \mu)) = -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln\,|C| - \frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^T C^{-1}(y_i - \mu). \quad (50)$$

Concluding that the log likelihood would be:

$$logp(\mathbf{Y}|\mathbf{W}) = \frac{1}{2}(NDln(2\pi) + N ln\,|C| + Tr(C^{-1}\mathbf{Y}\mathbf{Y}^T))) \quad (51)$$

2. *Write the gradients of the objective with respect to the parameters:*

Using the result obtained in the previous equation, we are going to calculate the gradient by terms. To achieve that, we will divide the equation in three terms:

The first term; $\frac{1}{2}NDln(2\pi)$, the second term; $\frac{1}{2}N ln\,|C|$,and the third term: $\frac{1}{2}Tr(C^{-1}\mathbf{Y}\mathbf{Y}^T)$.

As the first term is basically a constant, we can affirm that:

1. First term:

$$\frac{\partial(\frac{1}{2}NDln(2\pi))}{\partial \mathbf{W}_{ij}} = 0 \tag{52}$$

To calculate the gradient in the second term we need to apply the following regulation:

$$\partial(ln(det(\mathbf{X}))) = Tr(\mathbf{X}^{-1}\partial\mathbf{X}) \tag{53}$$

And then, we are going to obtain :

2. Second term:

$$\frac{\partial(\frac{1}{2}Nln\,|C|)}{\partial \mathbf{W}_{ij}} = \frac{N}{2}Tr(\mathbf{C}^{-1}\frac{\mathbf{C}}{\partial \mathbf{W}_{ij}}) \tag{54}$$

And finally, regarding the third term we can state that:

3. Third term:

$$\frac{\partial(\frac{1}{2}Tr(\mathbf{C}^{-1}\mathbf{Y}\mathbf{Y}^T))}{\partial \mathbf{W}_{ij}} = \frac{1}{2}Tr(\mathbf{Y}\mathbf{Y}^T(-\mathbf{C}_{-1}\frac{\mathbf{C}}{\partial \mathbf{W}_{ij}}\mathbf{C}_{-1})) \tag{55}$$

And then, we should also calculate the derivative of C over W:

$$\frac{\partial \mathbf{C}}{\partial \mathbf{W}_{ij}} = \frac{\partial \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}}{\partial \mathbf{W}_{ij}} = \mathbf{J}^{ij}\mathbf{W}^T\mathbf{J}^{ijT} \tag{56}$$

And, finally, we can state that the gradient of the objective function with respect to the parameters $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ is:

$$\frac{N}{2}Tr[(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}x\mathbf{J}^{ij}\mathbf{W}^T\mathbf{W}\mathbf{J}^{ijT}]+\frac{1}{2}Tr(\mathbf{Y}\mathbf{Y}^T[-(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}x\mathbf{J}^{ij}\mathbf{W}^T\mathbf{W}\mathbf{J}^{ijT}x(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}]) \tag{57}$$

2.2 Practical

2.2.1 Linear Representation Learning

The data that we are going to generate in this practical part will be the following:

$$\mathbf{Y} = f_{lin}(f_{nonlin}(\mathbf{x}))$$

$$\mathbf{x} = [0,...,4\pi]^T$$

$$|\mathbf{x}| = 100$$

$$f_{non-lin}(x_i) = [x_i\sin(x_i),x_i\cos(x_i)]$$

$$f_{lin}(x') = \mathbf{x'}\mathbf{A}^T$$

$$\mathbf{A} = \mathcal{R}^{10x2}$$

$$\mathbf{A}_{ij} \sim \mathcal{N}(0,1).$$

Question 16: *Plot the representation that you have learned (hint: plot* **X** *as a two-dimensional representation). Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?*

In this question we will plot the representation that we have learn from our model. In order to plot different situations, I have divided the question into three cases. The first and the second case are basically without applying noise and the third case is the comparison obtained with noise. It is really interesting to realize of the importance of noise in our data.

Nevertheless, it is important to know how hour real data look like. The **X** that we want to recover is **X** $=$ [**X**-sin(**X**),**X**-cos(**X**)] where **X**=[0,...,4π]^T. In the following figure you can see some samples that represents the latent variable **X**.

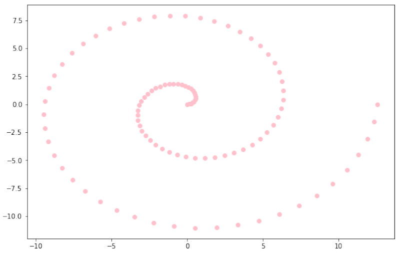


Figure 9: Real data.

The relationship between the variable **X** and the estimation of it, called **X'**, is the following one:

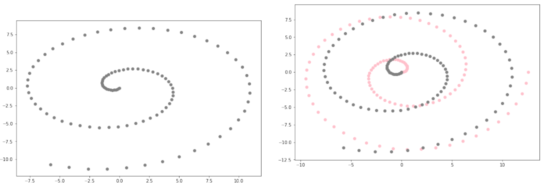
$$\mathbf{Y} = \mathbf{X}'\hat{\mathbf{W}}^T \tag{58}$$

$$\mathbf{Y}\hat{\mathbf{W}} = \mathbf{X}'\hat{\mathbf{W}}^T\hat{\mathbf{W}} \tag{59}$$

$$\mathbf{Y} = \mathbf{X}'\hat{\mathbf{W}}^T \tag{60}$$

Taking into account this relationship here I attached two different cases that represents the learned data and also, the learned data (plotted in grey) compared with the real data (plotted in pink).

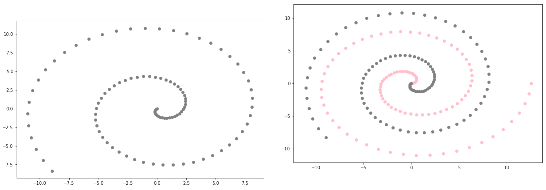
•**Case 1:**



(a) Learned data

(b) Learned data and Real data

•**Case 2:**



(a) Learned data

(b) Learned data and Real data

As you can realize in the two cases plotted above, *Case 1* is rotated compared with the real data figure. That is, the estimated data **X'** is basically a rotated version of **X**. Nevertheless, the marginal likelihood is invariant to a rotation of **W**. That is why we do not have a unique solution for **W** in linear mappings.

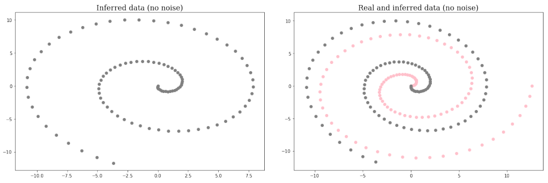
$$\hat{\mathbf{W}} = \mathbf{W}\mathbf{R} \tag{61}$$

$$p(\mathbf{Y}|\mathbf{W}) = \mathcal{N}(\mu + \mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}^T) \tag{62}$$

$$p(\mathbf{Y}|\mathbf{W}) = \mathcal{N}(\mu + \mathbf{W}\mathbf{W}^T) \tag{63}$$

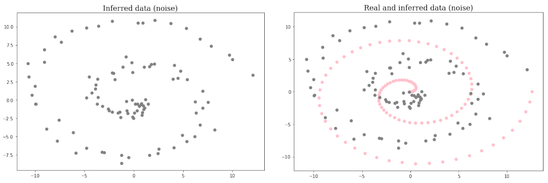
Finally, in the case 3, I decided to plot a comparison between having noise and to not, to be able to understand properly the importance of it in the estimation of our data. We are able to saw, perfectly, the difference between the "ideal" estimation and the "more realistic" one.

•Case 3:



(a) Learned data without noise

(b) Learned data and real data without noise



(a) Learned data with noise

(b) Learned data and real data with noise

3 III The Evidence $p(D)$

One of the main arguments behind the Bayesian reasoning is that it automatically implements Occam's razor and then it automatically will choose the "correct" model complexity to perform a specific task.

In this taks we will use the evidence of the data under the model as a means of measuring the complexity of the model.

3.1 Theory

In order to understand properly this section, you should read the paper of Murray,2005.

3.1.1 Data

Consider a simple data domain $\mathcal{D}$=t$^{i9}_{i=1}$ where t^{i} is betwwen -1,1. The data is structured according to a grid whos locations can be parametrised by $\mathcal{X}$=**x**$^{i9}_{i=1}$ where **x**^{i} = (-1,0,+1,-1,0,+1).

3.1.2 Models

Given the data defined above we wish to create a model that will explain the statistical variations that are possible in $\mathcal{D}$. The simplest model is something that simply takes all its probability mass and places it uniformly over the whole data space,

$$p(\mathcal{D}|M_0, \theta_0) = \frac{1}{512} \tag{64}$$

All the following questions are based in knowledge obtained from following paper:[17]

***Question 17**: Why is this the simplest model, and what does it actually imply? Discuss its implications, why is this a bad model and why is it a good model?*

The model M_0 is consider the simple model by the fact that all the possibles sets have the same probability to occur: $\frac{1}{512}$ In other words, this model does not gives any preference to any of the all possible sets, for this reason I can state that it is not a good model.

The fact of given equally probability in all the data sets implies that this first model is not providing any kind of information about $\mathcal{D}$.Any characteristics of $\mathcal{D}$ is captured.

Assignment theory:

The M_0 model does not take any parameters which implies that it has no flexibility and uses no information about $\mathcal{D}$ except for its cardinality. Nevertheless, we can use what we know about the data in order to create a more representative model. If we assume that all t^n are independent, we can factorise the model into 9 models:

$$p(\mathcal{D}|M_1, \theta_1) = \prod_{n=1}^{9} p(t^n|M_1, \theta_1) \tag{65}$$

Each model can be expressed using an exponential function which relates the value t^i to its location $\mathbf{x}^i$:

$$p(\mathcal{D}_1, \theta_1) = \prod_{n=1}^{9} \frac{1}{1 + e^{-t^n \theta_1^1 x_1^n}} \tag{66}$$

***Question 18**: Explain how 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?*

As we saw in the description of our assignment, we basically made two steps. The first one was focused in applying the independence condition between the data sets and the other one, was focused in changing the model of the prior probability from the first model M_0 to the new one M_1. In order to understand properly the difference of the two models, we are going to use the following image:

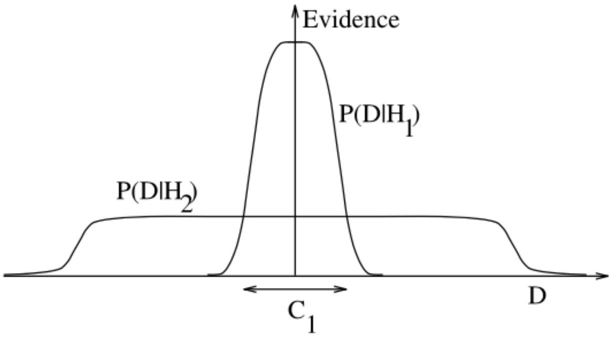


Figure 14: Figure from : "A note on the evidence and Bayesian Occam's razor", Murray. The image shows the different evidence given two different models; H1 and H2. The D-axis indexes all possible data sets.

The previous figure represents and compare the prior distribution of the two models; M_1 represented as H_1 in the figure and M_1 represented as H_2. If we analyze the picture, we could said that M_1 is the simplest model, because this model has a density of probability over $\mathcal{D}$ much narrower than the M_0 model.

We can also state that applying the M_1 also reduces the amount of sets that can be created compared with the M_0 model.

Assignment theory:

Then, we can continue adding more parameters and creating further models like M_2 and M_3:

$$p(\mathcal{D}|M_2, \theta_2) = \prod_{n=1}^{9} \frac{1}{1 + e^{-t^n \theta_2^1 x_1^n + \theta_2^2 x_2^n}} \tag{67}$$

$$p(\mathcal{D}|M_3, \theta_3) = \prod_{n=1}^{9} \frac{1}{1 + e^{-t^n \theta_3^1 x_1^n + \theta_3^2 x_2^n + \theta_3^3}} \tag{68}$$

***Question 19**: How have the choices we made above restricted the distribution of the model? What datasets 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.*

Regarding the M_2 model, we could state that as it increase its complexity adding the $\theta_2^2 x_2^n$ terms, it will be ble to generate more datasets respect the M_1 model. We can not compare it to the M_0 model by the simple fact that this model was able to generate all the possible datasets.

At the same time, M_3 model is even more complex than the others. The term θ_3^3 also implies that the probability density is more distributed, for this reason is a good model that fits when the datasets are no symmetric. The bias, θ_3^3 , let to have models that are not centered on the origin.

3.1.3 Evidence

The evidence of the model M_i is the distribution $p(\mathcal{D}|M_i)$. In the previous section we have defined a small simple data domain $\mathcal{D}$ and in this section we will evaluate where the different models defined above places their probability mass.

To be able to achieve the evidence of a model, we need to remove the dependency of the variable θ marginalising out the parameters from the model:

$$p(\mathcal{D}|M_i) = \int_{\forall\theta} p(\mathcal{D}|M_i,\theta)p(\theta)d\theta. \tag{69}$$

Question 20: *Explain the process of marginalisation and briefly discuss its implications.*

As we defined before this question, if we want to obtain the evidence, we have to be more focused in the model itself. In other words, we do not have so much interest in its parameters; $theta$.

That is why we are marginalizing out. If we are able to marginalize over $theta$, we will remove the dependence with the parameter. Nevertheless, to achieve the marginalization we should have already obtained $p(\mathcal{D}|M_i,\theta)$ and its prior. However, it has to be also clear that the prior might depend in the model. In that case, we should then, have the term $p(\theta|M)$ instead of $p(\theta)$.

Question 21: *What does this choice of prior imply? How does the choice of the parameters of the prior and effect the model?*

The choice in the description regarding the prior, is a simple Gaussian. Specifically, a prior with a zero mean ($\mu = 0$) and a high variance ($\sigma^2 = 10^3$).

$$p(\theta|M_i) = \int \mathcal{N}(0, 10^3\mathbf{I}) \tag{70}$$

The high variance implies that the distribution of $p(\theta|M_i)$ will be more wider. We can also state that the components of θ will be independent.

On the other hand, if we evaluate this prior in the different models (M_3, M_2, M_1, M_0) we can affirm that, as the variance has a high value, models like M_3, that has more flexible decision boundaries, are going to be more probable. In terms of variance, when $\sigma^2 = 0$, models like M_0 will be the more likely. And, regarding the models M_1 and M_2, we will need a smaller variance value.

3.2 Practical

Question 22: *Plot the evidence over the whole dataset for each model (and sum the evidence for the whole of D, explain the numbers you get). The* x-axis *index the different instances in D and each models evidence is on the* y-axis. How do you interpret this? Relate this to the parametrisation of each model.

In the following figure, we plot the distribution from the evidence over the whole data set for each model. Specifically, we are comparing with the four different models from we talked before: M_0, M_1, M_2, M_3. If we compare the results obtained with the ones obtained in "A note on the evidence and Bayesian Occam's Razor(2005) [17] we realize that they are similar,but not exactly the same. The reason behind that is basically because we have different parameters settings and also because I reduce the number of samples in order to be able to execute the code in my computer.

It is also clear that, as we said before, the simplest model M_0 spreads over the whole data set.

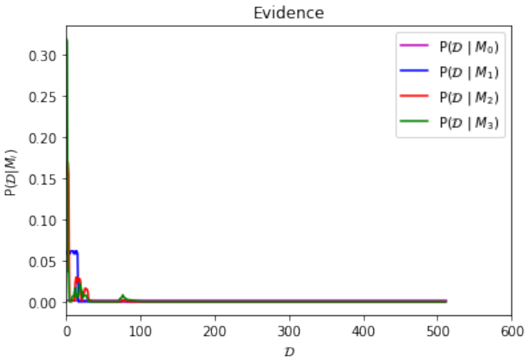


Figure 15: Evidence in the whole $\mathcal{D}$

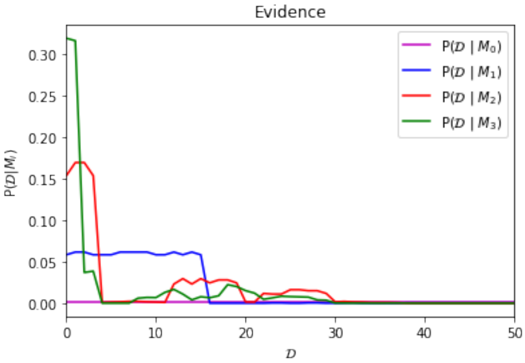


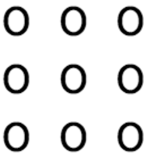
Figure 16: Zoom for 50 data sets.

***Question 23**:* *Find using* np.argmax *and* np.argmin which part of the D that is given most and least probability mass by each model. Plot the data-sets which are given the highest and lowest evidence for each model. Discuss these results, does it make sense?

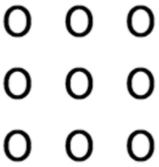
In this task we are going to show the most likely and the less likely generated dataset in the different models that we will evaluate ; M_0, M_1, M_2 and M_3. We are going to use 'O' and 'X' for 1 and 1, respectively.

•Model 0:

The result obtained in this model is the one that we have been commented during all the section ; all the datasets have the same probability given the model 0. However, the results show us that the affirmation was clear.



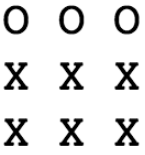
(a) Most probable data set given Model 0



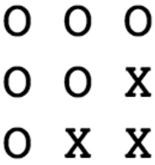
(b) Less probable data set given Model 0

•Model 1:

Taking into account the equation that define the model 1, since there is no bias term, this model is not able to place the decision boundary in another place than the origin. For this reason, the more probable dataset generated by M_1 have to be one that verifies the limitation stated above:



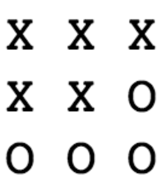
(a) Most probable data set given Model 1



(b) Less probable data set given Model 1

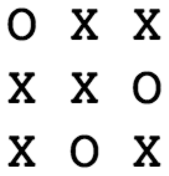
•Model 2:

As in the equation that define this model we have both axis ($x_1 and x_2$) we will be able to achieve diagonal decision boundaries, but centered in the origin (because we still do not have the bias term). Based on that:



X X X
X X O
O O O

(a) Most probable data set given Model 2



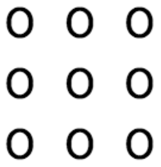
O X X
X X O
X O X

(b) Less probable data set given Model 2

•Model 3:

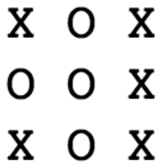
In the third model we have the bias term. This implies that the decision boundary would be located in any position between the two axis defined by $x_1 and x_2$. On the other hand, if we want to achieve a boundary located in the origin, the bias term should be zero. For this reason the less likely data set should be one that is located in the origin (a special case in given the equation that defines this model).

Based on that:



O O O
O O O
O O O

(a) Most probable data set given Model 3



X O X
O O X
X O X

(b) Less probable data set given Model 3

***Question 24**: What is the effect of the prior p(θ)*

- *What happens if we change its parameters?*
- *What happens if we use a non-diagonal covariance matrix for the prior?*
- *Alter the prior to have a non-zero mean, such that $\mu = [5,5]^T$?*
- *Redo evidence plot for these and explain the changes compared to using zero-mean.*

There are different parameters that modify the effect of the prior $p(\theta)$. Nevertheless, we should analyze the effect of changing the mean and the prior. Changing the variance (σ^2) we will have more or few variation in the parameters and, also, modifying the mean we can obtain smoother decision boundaries.

Finally, to continue, I will plot two different cases to compare and analyze what happens if we use a non-diagonal co-variance matrix for the prior and also, what happens if we alter the prior with a non-zero mean ($\mu = [5,5]^T$).

•Modifying parameters:

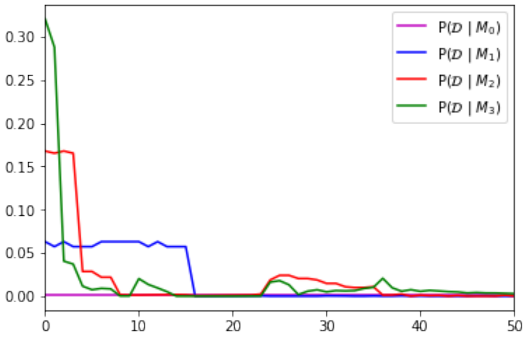


Figure 21: Case 1: non-diagonal covariance

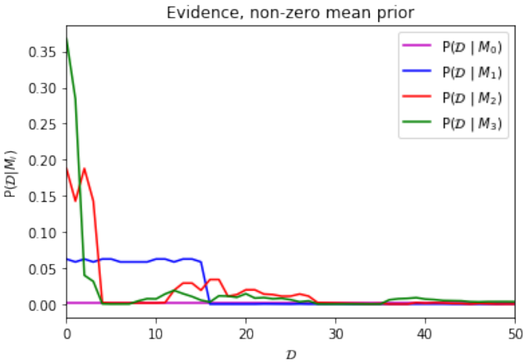


Figure 22: Case 2: non-zero mean.

In the first case, with the non-diagonal co-variance we are basically defining some correlation between the θ parameters. In the other case, using the mean as $\mu = [5,5]^T$ imply having more defined decision boundaries. That is why we are obtaining more "clear peaks" in the evidence plotted above.

Appendices

References

[1] C. M.Bishop, "Pattern recognition and machine learning," vol. 2: Probability distributions, p. 78, 2009.

[2] "Central limit theorem, wikipedia," https://en.wikipedia.org/wiki/Central_limit_theorem, note = Accessed: 2017-12-4.

[3] C. M.Bishop, "Pattern recognition and machine learning," vol. 2: Probability distributions, p. 79, 2009.

[4] "Cross validated: Isotropic covariance matrix," https://stats.stackexchange.com/questions/204595/what-is-an-isotropic-spherical-covariance-matrix, note = Accessed: 2017-11-10.

[5] C. M.Bishop, "Pattern recognition and machine learning," vol. 2: Probability distributions, p. 84, 2009.

[6] "Wikipedia: Chain rule," https://en.wikipedia.org/wiki/Chain_rule_(probability), note = Accessed: 2017-11-8.

[7] C. M.Bishop, "Pattern recognition and machine learning," vol. 3: Linear models for regression, pp. 140–141, 2009.

[8] ——, "Pattern recognition and machine learning," vol. 1: Introduction, p. 10, 2009.

[9] "Video tutorial of linear regression: Regularization," https://www.youtube.com/watch?v=sO4ZirJh9ds&t=413s, accessed: 2017-11-08.

[10] P. Herman, "Posterior exercise. machine learning advanced course."

[11] "Conjugate prior, wikipedia," https:https://en.wikipedia.org/wiki/Conjugate_prior, accessed: 2017-11-09.

[12] C. M.Bishop, "Pattern recognition and machine learning," vol. 6: Kernel Methods, p. 306, 2009.

[13] "Joint probability, wikipedia," https://en.wikipedia.org/wiki/Joint_probability_distribution, accessed: 2017-11-09.

[14] C. M.Bishop, "Pattern recognition and machine learning," vol. 8: Graphical models, p. 374, 2009.

[15] ——, "Pattern recognition and machine learning," vol. 3: Linear models for regression, p. 154, 2009.

[16] "Maximum likelihood estimation, wikipedia," https://https:/en.wikipedia.org/wiki/Maximum_likelihood_estimation, accessed: 2017-11-09.

[17] Z. G. Iain Murray, "A note on the evidence and bayesian occam's razor," 2005.