DD2434 Machine Learning, Advanced Course

Fernando García Sanz - fegs@kth.se

December 5, 2019

Contents

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

1.1 Theory

Question 1:

It is known that many natural phenomenon follow a normal distribution. From a mathematical point of view, according to *Central Limit Theorem*, the sum of a large number of independent random variables becomes increasingly Gaussian as the number of terms in the sum increases. Therefore, Gaussian distribution is a good way to represent a large amount of data since we assume that the amount of data is large enough and that random variables are represented on it.

Considering covariance matrices that are diagonal, so that $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$, there are a total of 2D independent parameters in the density model. If the covariance matrix is restricted to be proportional to the identity matrix, it is said that $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. This is known as *isotropic covariance*. Isotropic means that all dimensions vary on the same rate, so it makes the variation of values follow a spherical shape. Therefore, this matrix has the same value in all components of its diagonal, being this value σ^2, so it can be said that every dimension is independent of the others and all have the same variance, variance of the distribution σ^2.

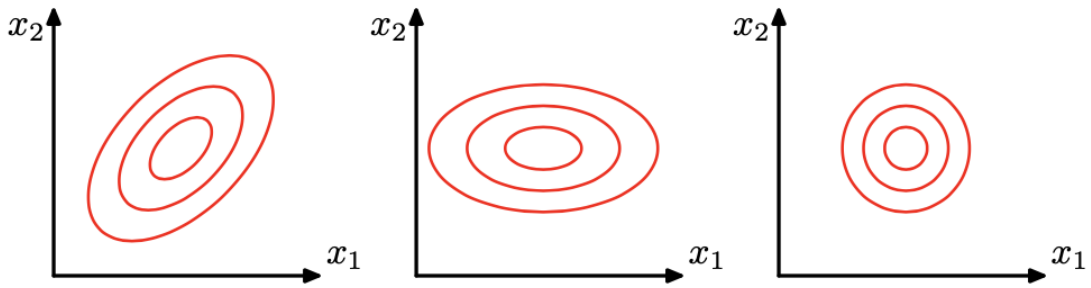


Figure 1: Covariances

Question 2:

If conditional independence is assumed between the output points, they can be represented as:

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

Nevertheless, this representation is only valid when there exists conditional independence in the outcome points. When there is some kind of dependence, it is necessary to apply *Bayes Theorem*:

$$p(\mathbf{T}|f,\mathbf{X}) = p(\mathbf{t}_1,...,\mathbf{t}_n|f,\mathbf{X})$$

So, from here, it can be known that:

$$p(\mathbf{t}_1,...,\mathbf{t}_n|f,\mathbf{X}) = \frac{p(\mathbf{t}_1|\mathbf{t}_2,...,\mathbf{t}_n,f,\mathbf{x}_1)\, p(\mathbf{t}_2,...,\mathbf{t}_n,f,\mathbf{x}_2,...,\mathbf{x}_n)}{p(f,\mathbf{X})}$$

So, finally, expanding all expressions:

$$p(\mathbf{T}|f,\mathbf{X}) = p(\mathbf{t}_1|\mathbf{t}_2,...,\mathbf{t}_n,f,\mathbf{x}_1)\, p(\mathbf{t}_2|\mathbf{t}_3,...,\mathbf{t}_n,f,\mathbf{x}_2)...p(\mathbf{t}_n|f,\mathbf{x}_n)$$

Question 3:

The real value of a target in linear regression is defined by:

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

So it can be said that, due to conditional independence:

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} p(\mathbf{t}_i|\mathbf{X},\mathbf{W})$$

And, substituting from the terms of linear regression equation, it is known that:

$$p(\mathbf{t}_i|\mathbf{X},\mathbf{W}) \sim \mathcal{N}(\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}) \implies p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I})$$

Question 4:

For this specific case, it is given that the prior is defined by $p(\mathbf{W}) = \mathcal{N}(\mathbf{w}_0, \tau^2\mathbf{I})$.

Defining the function as:

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\epsilon} \quad \rightarrow \quad \prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n|\mathbf{w}_n\mathbf{x}_n)$$

According to the prior probability, this is defined as:

$$\prod_{n=1}^{N} \mathcal{N}(\mathbf{y}_n|\mathbf{w}_n\mathbf{x}_n)\,\mathcal{N}(\mathbf{w}_0, \tau^2\mathbf{I}) \quad \rightarrow \quad \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{y}_n|\mathbf{w}_n\mathbf{x}_n) + \ln \mathcal{N}(\mathbf{w}_0, \tau^2\mathbf{I})$$

Focusing on the prior, for a multivariate distribution:

$$\sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{w}_0, \tau^2\mathbf{I}) \quad \rightarrow \quad \sum_{n=1}^{N} \ln \frac{1}{\sqrt{(2\pi)^k|\tau^2\mathbf{I}|}} + \ln exp(-\frac{1}{2}(\mathbf{w}_n - \mathbf{w}_0)^T(\tau^2\mathbf{I})^{-1}(\mathbf{w}_n - \mathbf{w}_0))$$

And from here, it can be derived that:

$$\sum_{n=1}^{N} const - \frac{1}{2\tau^2}(\mathbf{w}_n - \mathbf{w}_0)^T(\mathbf{w}_n - \mathbf{w}_0) \quad \rightarrow \quad \lambda \sum_{n=1}^{N}(\mathbf{w}_n - \mathbf{w}_0)^T(\mathbf{w}_n - \mathbf{w}_0)$$

This formula is equal to the L_2 regularization term. The idea of adding a regularization term is providing a mechanism to control over-fitting.

The case of L_2, as explained before, is equal to:

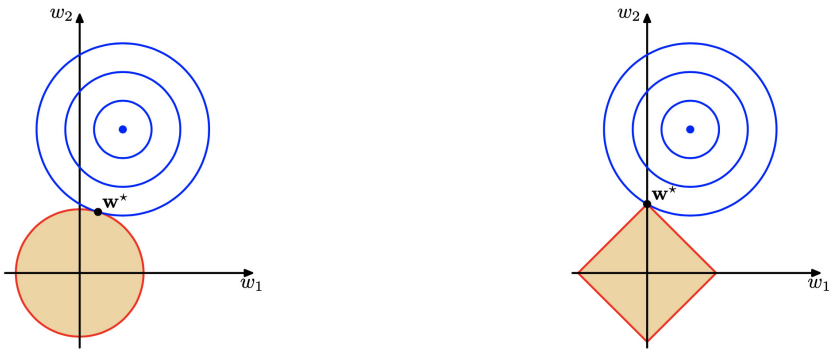
$$RSS + \lambda \sum_{n=1}^{N} \mathbf{w}_n^2$$

L_1 expression is equal to:

$$RSS + \lambda \sum_{n=1}^{N} |\mathbf{w}_n|$$

L_1 norm is also known as *Lasso* and L_2 as *Ridge Regression*.

A representation of how the spherical covariance matrix deals with this terms can be graphically found below:



(a) Ridge Regression (b) Lasso

Figure 2: Ridge Regression vs Lasso

With *Lasso*, some features that are irrelevant for the model will become zero, removing any influence they will have. Meanwhile, with *Ridge regression*, irrelevant features will see how their influence decreases until reaching values close to zero, but never zero; so their influence will be small, but never eliminated.

Question 5:

The target values $\mathbf{T}$ are represented by the estimated values $\mathbf{Y}$ and the error $\boldsymbol{\epsilon}$ in the following way:

$$\mathbf{T} = \mathbf{Y} + \boldsymbol{\epsilon} = \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}$$

The posterior is expressed as:

$$P(\mathbf{W}|\mathbf{X},\mathbf{T}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{W}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) \quad where \quad \mathbf{X} = \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(D)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \dots & x_n^{(D)} \end{bmatrix} \implies \mathbf{T} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}$$

From the definition of *normal distribution*, it can be derived that:

$$\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{W}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) = \underbrace{\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_{\boldsymbol{W}}|^{\frac{1}{2}}}}_{\mathbf{C}} e^{-\frac{1}{2}(\mathbf{W}-\boldsymbol{\mu}_{\boldsymbol{W}})^T\Sigma_W^{-1}(\mathbf{W}-\boldsymbol{\mu}_{\boldsymbol{W}})} =$$

$$= \mathbf{C} \underbrace{(e^{-\frac{1}{2}\mathbf{W}^T \boldsymbol{\Sigma_W}^{-1} \mathbf{W}})}_{quadratic\,(\mathbf{W}^2)} \underbrace{(e^{\mathbf{W}^T \boldsymbol{\Sigma_W}^{-1} \boldsymbol{\mu_W}})}_{linear\,(\mathbf{W})} \underbrace{(e^{-\frac{1}{2}\boldsymbol{\mu_W}^T \boldsymbol{\Sigma_W}^{-1} \boldsymbol{\mu_W}})}_{constant\,(no\,\mathbf{W})}$$

The posterior can also be expressed as:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{T}) = p(\mathbf{W})\, p(\mathbf{X}, \mathbf{T}|\mathbf{W}) \sim e^{likelihood\,exp} e^{prior\,exp} = e^{likelihood\,exp + prior\,exp}$$

So it is necessary to find the values of those exponents, likelihood and prior.

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \qquad \mathbf{T} = \mathbf{Y} + \boldsymbol{\epsilon} = \mathbf{XW} + \boldsymbol{\epsilon}^{(\mathbf{0}, \boldsymbol{\Sigma_\epsilon})} \qquad \boldsymbol{\Sigma_\epsilon} = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Probability of $\mathbf{W}$ can be represented as a normal distribution with mean = 0 and covariance $\boldsymbol{\Sigma}$. Meanwhile, the error in the expression is assumed to have mean zero and a diagonal covariance matrix.

Therefore, the posterior can be expressed as:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{T}) = p(\mathbf{T}|\mathbf{W}, \mathbf{X})\, p(\mathbf{W}) = \mathcal{N}(\mathbf{XW}, \sigma^2 \mathbf{I})\, \mathcal{N}(\mathbf{w}_0, \tau^2 \mathbf{I})$$

Where σ corresponds to the likelihood covariance and τ corresponds to the covariance of the prior.

Expanding this expression and focusing only on the exponents:

$$-\frac{1}{2\sigma^2}(\mathbf{T} - \mathbf{XW})^T(\mathbf{T} - \mathbf{XW}) - \frac{1}{2\tau^2}(\mathbf{W} - \mathbf{w}_0)^T(\mathbf{W} - \mathbf{w}_0) =$$

$$= \underbrace{-\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T} - \frac{1}{2\tau^2}\mathbf{w}_0^T\mathbf{w}_0}_{Constant} + \underbrace{\frac{1}{\sigma^2}\mathbf{T}^T\mathbf{XW} + \frac{1}{\tau^2}\mathbf{W}^T\mathbf{w}_0}_{Linear} - \underbrace{\frac{1}{2\sigma^2}(\mathbf{XW})^T(\mathbf{XW}) - \frac{1}{2\tau^2}\mathbf{W}^T\mathbf{W}}_{Quadratic}$$

So now, it is possible to match the expressions obtained here with those previously computed, obtaining the following:

1. For the quadratic terms:

$$\boldsymbol{\Sigma_w} = (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})^{-1}$$

2. For the linear terms:

$$\boldsymbol{\mu_w} = (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})^{-1}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{w}_0)$$

So, finally, it can be said that:

$$p(\mathbf{W}|\mathbf{T},\mathbf{X}) = \mathcal{N}((\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})^{-1}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{w}_0), (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I})^{-1})$$

Where σ corresponds to the likelihood covariance and τ corresponds to the covariance of the prior.

Assuming conditional independence, this expression can be generalized to:

$$p(\mathbf{W}|\mathbf{X},\mathbf{T}) = \prod_{i=1}^{N} \mathcal{N}((\frac{1}{\sigma^2}\mathbf{x}_i^T\mathbf{x}_i + \frac{1}{\tau^2}\mathbf{I})^{-1}(\frac{1}{\sigma^2}\mathbf{x}_i^T\mathbf{x}_i + \frac{1}{\tau^2}\mathbf{w}_0), (\frac{1}{\sigma^2}\mathbf{x}_i^T\mathbf{x}_i + \frac{1}{\tau^2}\mathbf{I})^{-1})$$

Regarding the posterior form, this relates to the least square estimator of W in the way that the obtained distribution has as mean the most probable value of w (maximum likelihood).

According to *Bayes Theorem*:

$$p(\mathbf{W}|\mathbf{X},\mathbf{T}) = \frac{p(\mathbf{X},\mathbf{T}|\mathbf{W})\, p(\mathbf{W})}{p(\mathbf{X},\mathbf{T})}$$

It can be seen that the denominator, which is substituted by Z in the given expression has no relation to w, so it does not affect to the calculated solution.

Nevertheless, it is important to keep the value of Z because it is considered as a normalization factor, which shapes the posterior in the format of a probability density functions. It is called the evidence and it is useful to select between different models.

Question 6:

According to the expression:

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

The prior constraints the values the posterior can take. The kernel function k typically chosen to express the property that for x points that are similar, the corresponding values of y will be more strongly correlated that for points that are not.

As it can be seen in Figure 3, those points which are close have smaller margins around. Nevertheless, those which have greater distances in between produce greater margins, since uncertainty will be higher.

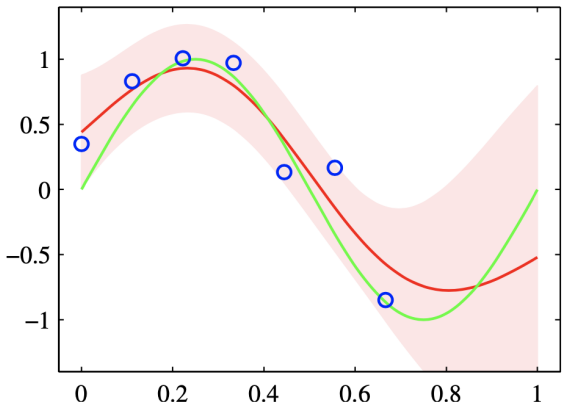


Figure 3: Confidence Intervals

Now, focusing on the marginal distribution, represented by the following formula:

$$p(\mathbf{T}|\mathbf{X}, \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{T}|f)}_{likelihood} \underbrace{p(f|\mathbf{X}, \boldsymbol{\theta})}_{prior} df$$

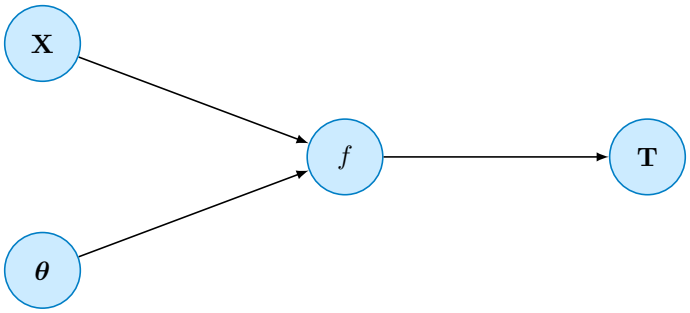
Since the prior is defined by the normal distribution $\mathcal{N}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}))$, it can be seen that if two points are close to each other, this will increase their correlation and the probability will be higher, due to the kernel function. Therefore, it will define with higher probability the values of a target, setting a smaller margin.

Question 7:

The joint distribution for the model:

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

A graphical representation of the model is given by:



Question 8:

Given the formula, it can be said that:

$$p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta}) = \int \underbrace{p(\mathbf{T}|f)}_{likelihood} \underbrace{p(f|\mathbf{X},\boldsymbol{\theta})}_{prior} df$$

The function f is estimated from the data $\mathbf{X}$ and the hyperparameters $\boldsymbol{\theta}$, and that it is this function f the one used to predict the target $\mathbf{T}$. Therefore, the function f is generated from the data and it is then used to predict the targets. The result of the integral is another Gaussian function, and it allows to directly relate the values of the data with the targets.

Uncertainty filters trough the marginalization because data points that are close to each other define lower uncertainty, but those which are further away define a bigger uncertainty area for target prediction. As points relate to each other, the uncertainty is affected by them.

The hyperparameter $\boldsymbol{\theta}$ still appears when integrating over f, so even after marginalisation, it is necessary to be taken into account. Also, the kernel function is used in the covariance, and it requires $\boldsymbol{\theta}$ too.

1.2 Practical

Question 9:

Assuming that there is not previous information in the system, the prior will be defined by a mean equal to zero and an isotropic covariance matrix, which will have the value σ^2 in its diagonal.

The result of the prior distribution over $\mathbf{W}$ is:

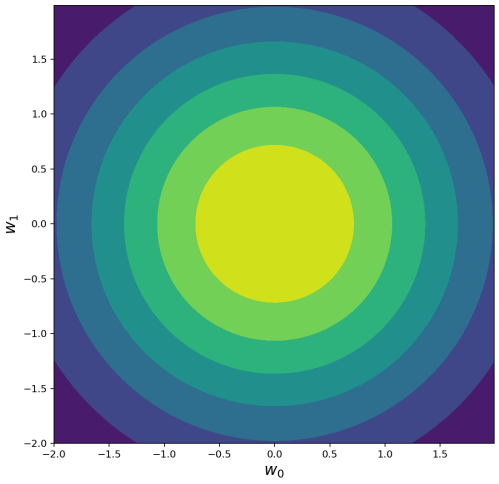
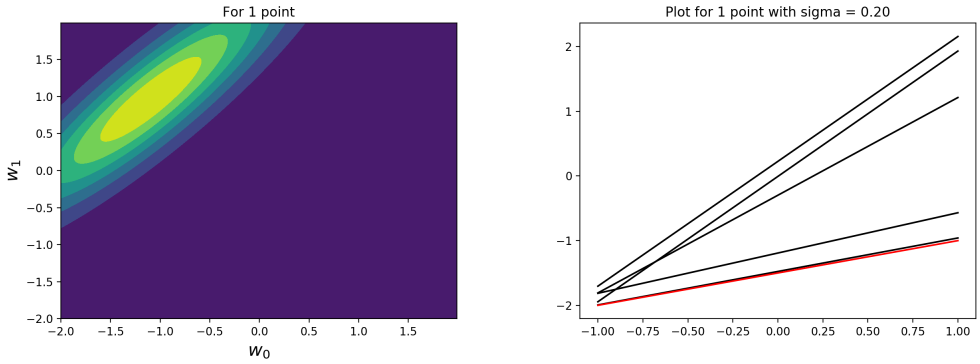


Figure 4: Prior distribution over $\mathbf{W}$

According to these initial assumptions, a point x_i belonging to the set $[-1, 1]$ is chosen and the target t_i corresponding to that point is generated. The following plots show the possible values of w_0 and w_1 and the lines they represent, being the red one the theoretical one that better represents the distribution.

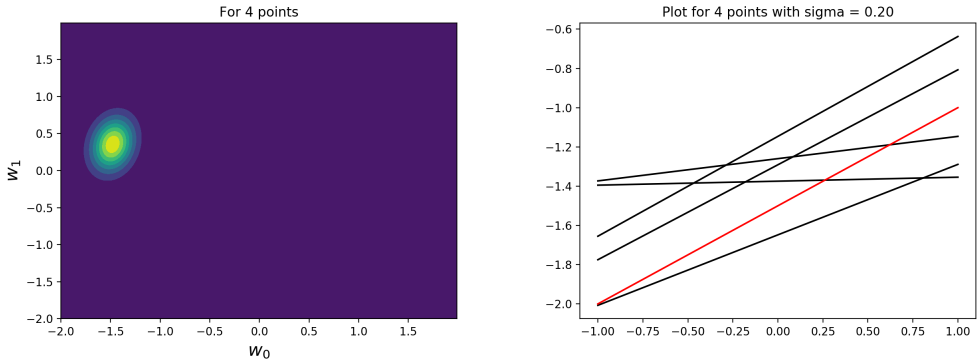


(a) Posterior for (x, y)

(b) Samples for (x, y)

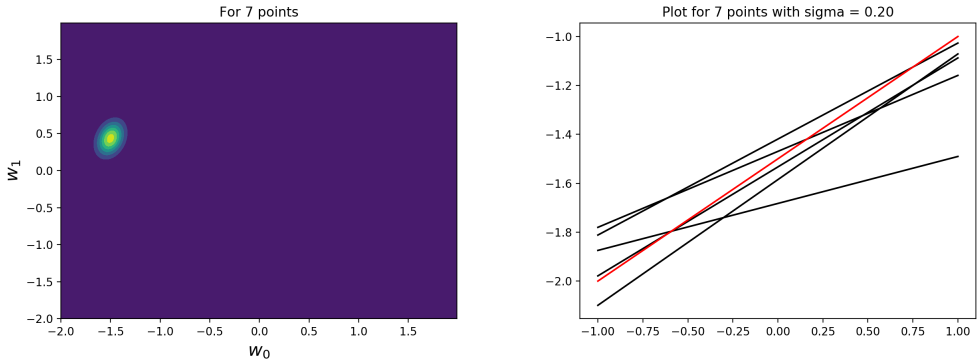
Figure 5: Posterior and samples for (x, y)

Using several points for calculation provides different results:



(a) Posterior for 4 (x, y) (b) Samples for 4 (x, y)

Figure 6: Posterior and samples for 4 (x, y)

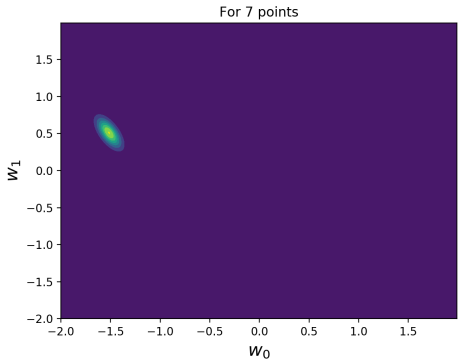


(a) Posterior for 7 (x, y) (b) Samples for 7 (x, y)

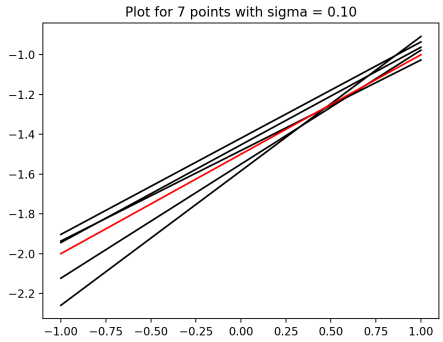
Figure 7: Posterior and samples for 7 (x, y)

As it can be seen, the bigger the amount of points, the bigger the amount of provided information. The first graphic shows an eventually smaller area in which the values of w_0 and w_1 can be and the second one shows lines which are closer each time to the theoretical distribution line.

If the value of σ which represents the variance varies, this is translated in the following events, visualized in the case of 7 points:

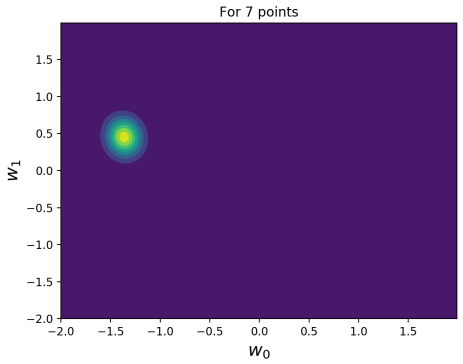


(a) Posterior for 7 (x, y) with $\sigma = 0.1$

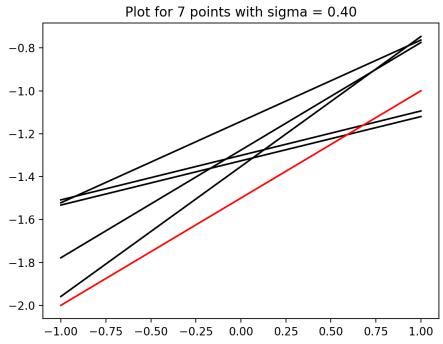


(b) Samples for 7 (x, y) with $\sigma = 0.1$

Figure 8: Posterior and samples for 7 (x, y) with $\sigma = 0.1$

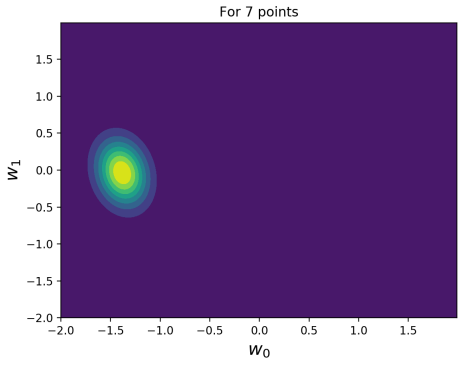


(a) Posterior for 7 (x, y) with $\sigma = 0.4$

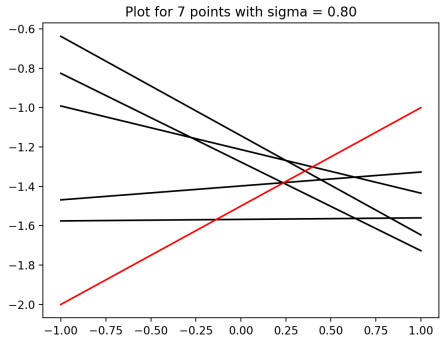


(b) Samples for 7 (x, y) with $\sigma = 0.4$

Figure 9: Posterior and samples for 7 (x, y) with $\sigma = 0.4$



(a) Posterior for 7 (x, y) with $\sigma = 0.8$



(b) Samples for 7 (x, y) with $\sigma = 0.8$

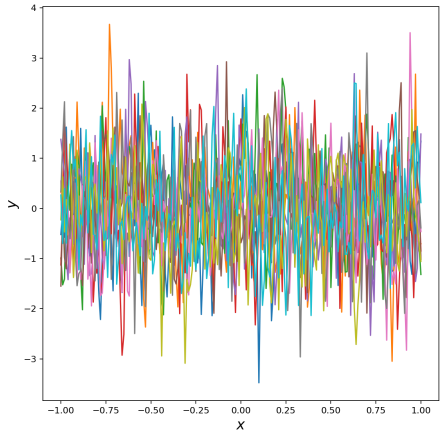
Figure 10: Posterior and samples for 7 (x, y) with $\sigma = 0.8$

As these plots show, the bigger the σ, the bigger the error, so the bigger the uncertainty when

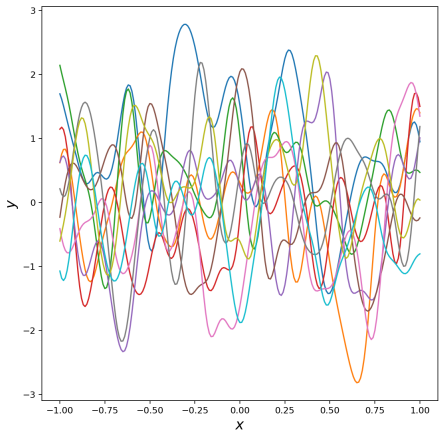
defining the model parameters. For the smaller value of σ, the values of w_0 and w_1 are well defined in a small area. The same happens to the plotted sample, which tend to converge in the original line. Nevertheless, in the last case, the drawn area is much bigger and the plotted lines do not converge at all. Therefore, the probability expressed by the posterior is more accurate when the σ value is small.

Question 10:

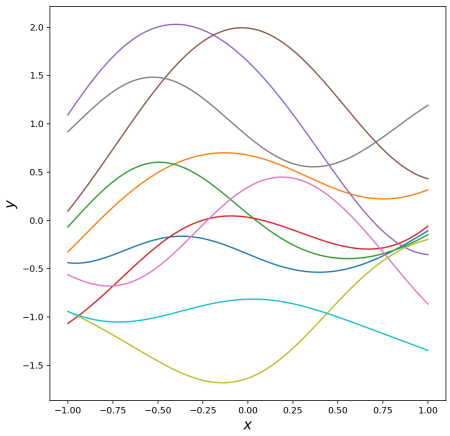
The $\mathcal{GP} - prior$ defined using the provided equation allows to draw the following samples according to the length scale used:



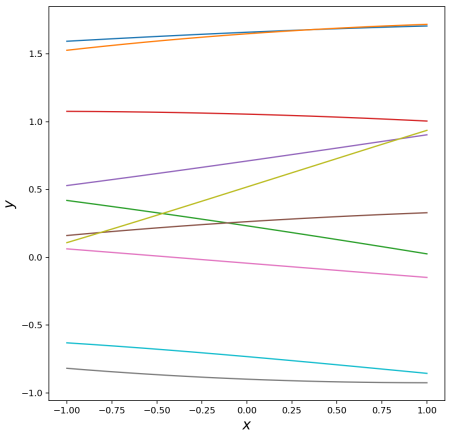
(a) Sampling with $l = 0.01$



(b) Sampling with $l = 0.1$



(c) Sampling with $l = 1$



(d) Sampling with $l = 10$

Figure 11: Samples with different length scales

As it can be seen, the bigger the value of l, the flatter the drawn samples are. The kernel

function used to express the covariance is defined by the following equation:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 e^{-\frac{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}{l^2}}$$

Which implies that:

$$lim_{l\to\infty} \implies exp[0] \implies k \to \sigma^2$$

$$lim_{l\to 0} \implies exp[-\infty] \implies k \to 0$$

Therefore, for a big value of l, the covariance will be closer to σ^2, and for a small one (being l a positive value), it will be closer to zero. For a small covariance value, there will be no dependency between the points, so the sample will be noisy. On the other hand, if the value is away from zero, there will be more linear relation between the points, and this is why the last plot shows more linear samples.

Question 11:

Before observing any data, the only knowledge available is the prior, so the posterior will be given by the prior itself.

Now, it is provided some initial information, and it is required to predict the target values that several not given data points will produce. To do so, the following formulas have been used:

$$p(\boldsymbol{f}^*, \boldsymbol{f}|\mathbf{x}^*, \mathbf{X}, \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{x}^*) \\ k(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right) \implies$$

$$p(\boldsymbol{f}^*|\mathbf{x}^*, \mathbf{X}, \boldsymbol{f}, \boldsymbol{\theta}) = \mathcal{N}(\underbrace{k(\mathbf{x}^*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}\boldsymbol{f}}_{mean}, \underbrace{k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}k(\mathbf{X}, \mathbf{x}^*)}_{variance})$$

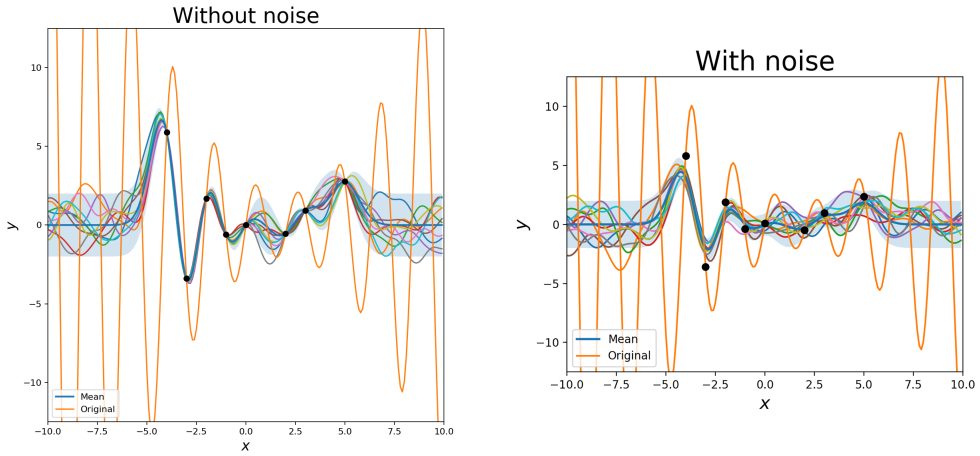
Nevertheless, noise must be also taken into account. When generating the values of the targets according to the provided data points, it is done by means of $t_n = f_n(\mathbf{x}_n) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$

The addition of noise also implies that the kernel function $K(\mathbf{X}, \mathbf{X})$ has to have some noise, causing the following effect:

$$\begin{bmatrix} \mathbf{t} \\ \boldsymbol{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I} & k(\mathbf{X}, \mathbf{x}^*) \\ k(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right)$$

Therefore, now the noise is also included in the kernel function, described before, which is used to calculate both mean and variance.

Below, it can be seen two plots which express different samples of the posterior, with and without noise.



(a) Sampling without noise (b) Sampling with noise

Figure 12: Samples of posterior with and without noise

As it can be seen in the first plot, the samples go trough the known points and the further they are from them, the more the samples approximate to the mean. On the other hand, in the second one, due to the noise addition, the samples do not go through all known data points. The same happens to the original function without noise, represented in blue; due to the noise, it does not go trough all data points.

If we compare these plots to those obtained in the previous question, it can be seen that due to that now there are data points that provide information, the likelihood defines the appearance of the samples. Before, as there was no more information but the prior, the samples were much more random.

Finally, according to the fact of adding a diagonal covariance matrix to the squared exponential, this is translated into adding noise. The more the noise, the greater the variance. Therefore, the samples will not be that close to the mean, they will vary a lot around it instead.

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

2.1 Theory

Question 12:

The prior $p(\mathbf{X}) = \mathcal{N}(0, \mathbf{I})$ defines a normal distribution in which its dimensions show no correlation. Therefore, it is a good prior to choose since it establishes no preference over the data.

Question 13:

The expression to marginalize is the following:

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

This corresponds to a linear Gaussian model, whose marginal distribution is given by:

$$p(\mathbf{Y}|\mathbf{W}) = \mathcal{N}(\mu, \mathbf{C})$$

This can be proved by the following, knowing that X follows a normal distribution $\mathcal{N}(0, I)$:

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{W}\mathbf{X} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$

$$cov[\mathbf{Y}] = \mathbb{E}[(\mathbf{W}\mathbf{X} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{X} + \boldsymbol{\epsilon})^T] \rightarrow \mathbb{E}[\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \rightarrow \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Question 14:

- For the maximum likelihood, being $\Sigma = \sigma^2\mathbf{I}$:

$$\mathcal{L}(\mathbf{W}) = argmax_{\mathbf{W}} \sum_{i=1}^{N} -\frac{1}{2}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T\Sigma^{-1}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)$$

  This expression takes into account the initial data to maximize the parameters which will describe its distribution. Nevertheless, the lack of regularization term can suppose the generation of an overfitted model when the amount of data is large.

- For the maximum-a-posteriori:

$$\mathcal{L}(\mathbf{W}) = argmax_{\mathbf{W}} \sum_{i=1}^{N} -\frac{1}{2}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T \Sigma^{-1}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i) + \frac{1}{2\sigma_{\mathbf{W}}^2}\mathbf{W}^T\mathbf{W}$$

For this expression, the probability of W is also taken into account when trying to predict the values of the next targets according to the new observations. This finds the model which has the higher probability of generating the dataset. Moreover, the addition of the second term allows regularization, mechanism that allows to prevent overfitting.

- For the Type-II maximum likelihood:

$$\mathcal{L}(\mathbf{W}) = argmax_{\mathbf{W}} -\frac{ND}{2}ln(2\pi) - \frac{N}{2}ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})$$

In this case, the expression is marginalized over $\mathbf{X}$, is equal to the expression obtained in the previous question. Nevertheless, and equally to the first equation, the lack of regularization term can generate overfitting. Even though, it is a good idea to use it since it allows to maximize without relating to the latent variables.

Now, for the second equation:

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

These two terms are equal because of the denominator. As the denominator is marginalized over $\mathbf{W}$, the expression will be expressed without $\mathbf{W}$ and will act as a constant that will not affect to the value of $\mathbf{W}$. This is why it can be removed.

Question 15:

Taking into account the Type-II maximum likelihood derived in the previous exercise, it will only be necessary multiplying it by minus one.

$$\mathcal{L}(\mathbf{W}) = -log(p(\mathbf{Y}|\mathbf{W})) = \frac{ND}{2}ln(2\pi) - \frac{N}{2}ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu}) =$$

$$= \frac{N}{2}\{Dln(2\pi) + ln|\mathbf{C}| + Tr(\mathbf{C}^{-1}\mathbf{S})\}$$

Where S is the data covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^T$.

Once the objective function has been computed, the derivative over $\mathbf{W}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = -N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W})$$

Note: see [1] for more information.

2.2 Practical

Question 16:

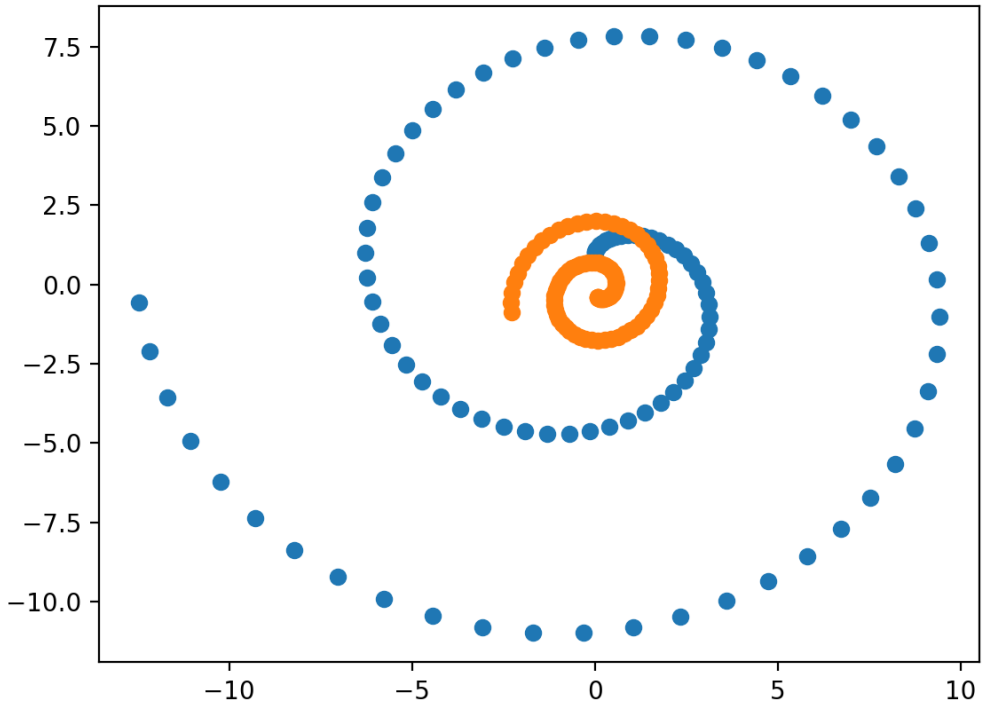
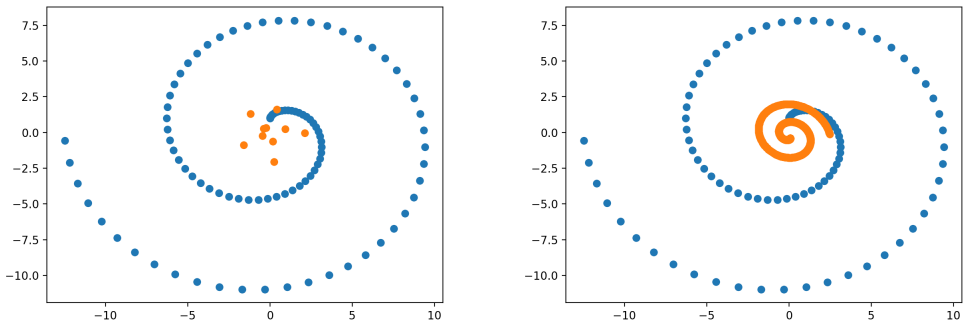


Figure 13: Learned model 100 points

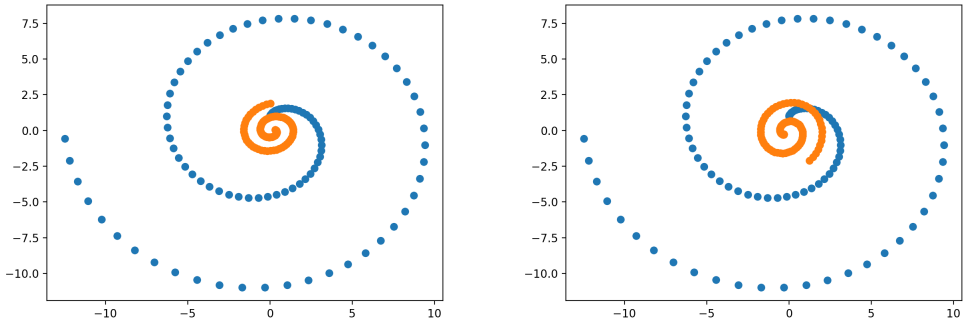
In the plot it can be seen first, in blue, the different original points and, in orange, the predicted ones by the learned model. The spiral described by the predicted points is much smaller than the one conformed by the original points. Its direction is different due to the optimization algorithm because there are several configurations of parameters in the algorithm which can provide a valid result.

Note: See [2] section 12.2.1 for more information.



(a) Learned model 10 points (b) Learned model 500 points

Figure 14: Learned models with different amounts of data points



(a) Learned model 100 points $\sigma^2 = 0.01$ (b) Learned model 100 points $\sigma^2 = 10$

Figure 15: Learned models with different values of σ^2

Nevertheless, when the number of points is changed, as well as the value of σ^2, there are small variations in the shape defined by the retrieved points from the model. With more points, the shape will be better defined, but the result describes the same figure. Only when the value of points is extremely small is not possible to devise the shape of the data. The variations in σ^2 do not affect the final output because when it changes, the weights $\mathbf{W}$ also change proportionally to compensate for this variation.

3 The Evidence $p(\mathcal{D})$

3.1 Theory

Question 17:

Because it gives uniform probability, i.e. all elements of the data domain are equally probable. Therefore, it is not a good model because it does not provide relevant information about the data. On the other hand, it is a good model because it is simple to define since it does not have free parameters, and complex in the way that assigns different behaviours the same probability [3].

Question 18:

As explained in the previous question, the first model assigns equal probability to all elements of the data domain, without taking into account any characteristic of the data. Nevertheless, the second model takes into account the first dimension of the data x_1, weighted by the parameter θ_1^1; therefore, this model takes into account some information provided by the data set and is, then, more flexible than the first model since it will adapt to the inputs instead of assigning an equal probability.

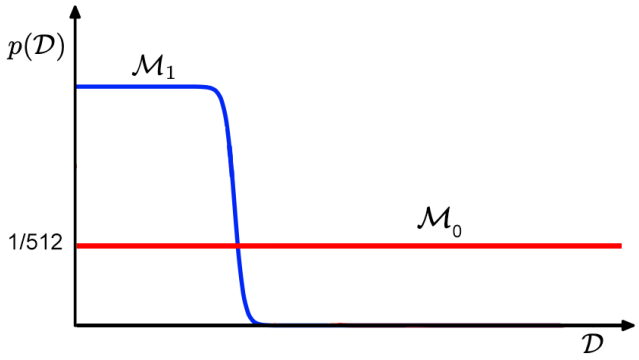


Figure 16: Probability according to the data domain

In the figure above, it can be seen how the probability $p(\mathcal{D})$ varies according to the complexity. The more complex the data domain is, the lower is the probability provided by $\mathcal{M}_1$; this is due to the model complexity. When the dataset is too complex, the model can not represent it properly. Nevertheless, as $\mathcal{M}_0$ has a fixed probability, it will not change independently of the model complexity.

Question 19:

The main difference between these two models is the addition of an parameter θ_3^3 in the last one, which represents the bias parameter. This makes this model more complex, since more parameters and taken into account, therefore the model will be more sensitive to data changes.

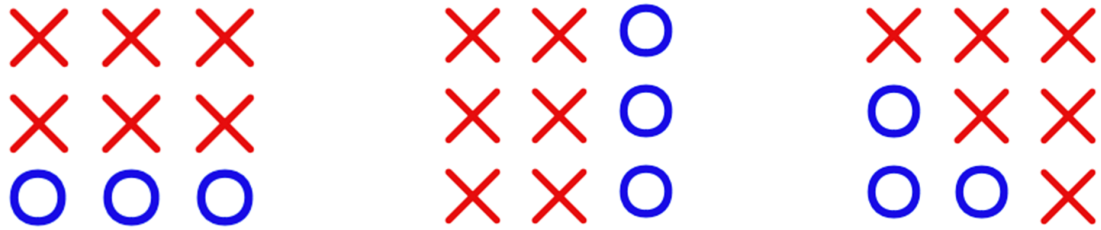


Figure 17: Labeled data sets

According to the labeled data sets showed above, for the first one, The $\mathcal{M}_1$ is the simplest model, since it can only define vertical boundaries (only considers x_1), $\mathcal{M}_2$ can define vertical, horizontal ones and diagonal ones through the origin, and $\mathcal{M}_3$ can do the same $\mathcal{M}_2$ does but adding and offset too, so this is the most complex one. Regarding the figure above, the first set can be represented by $\mathcal{M}_2$ and $\mathcal{M}_3$, the second one by any model and the last one only by $\mathcal{M}_3$.

Now, according to flexibility and restriction of the models:

Let's define a flexible model as the one with the capacity of adapting to different data sets; and a restricted model as the one with less parameters.

- $\mathcal{M}_0$: This model is the less flexible of all of them. It will assign the same probabilities independently of the input characteristics. Therefore, this is the more restrictive model too, since it has no parameters.
- $\mathcal{M}_1$: This model is flexible only attending to the horizontal coordinates x_1 of the input data. This is also a restrictive model since it uses just one parameter.
- $\mathcal{M}_2$: This model is more flexible since it takes into account both horizontal and vertical coordinates of data. This model is less restrictive than the previous one since it adds another parameter.
- $\mathcal{M}_3$: This is the most flexible model since it represents the most complex of all those defined. It is also the least restricted one as it is the one with more parameters.

Question 20:

Marginalization allows to directly express the probability of obtaining some data regarding the model. This provides the probability density of the model, and the model which represents the maximum likelihood (the biggest probability is where we will see the data) is the one which should be chosen.

Question 21:

Choosing mean equal to zero implies that the prior is centered at the origin. The fact that the variance $\Sigma = \sigma^2\mathbf{I}$ implies, first, that the dimensions are independent (no correlation), and second, that the uncertainty is high, since $\sigma^2 = 1000$, therefore, it is a big area in which the data points can be located.

3.2 Practical

Question 22:

The evidence according to each dataset obtained can be seen in the following plot:

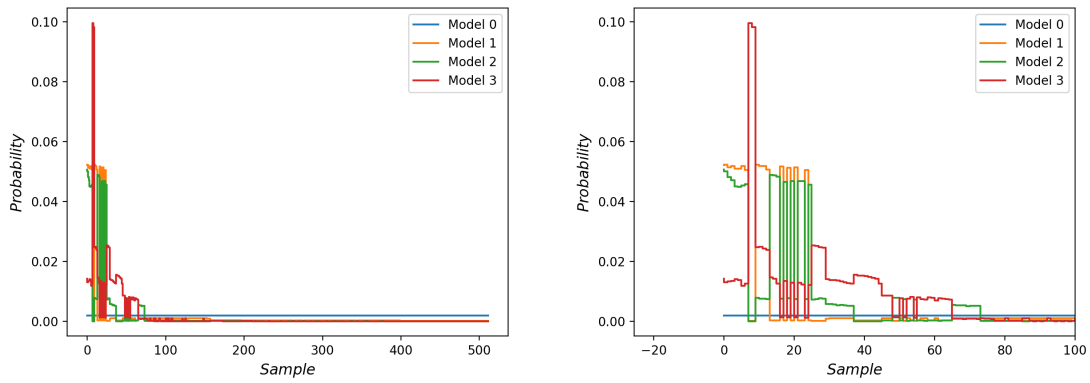
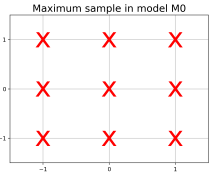


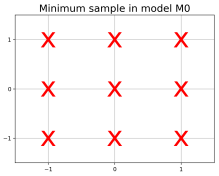
Figure 18: Evidence over dataset (right figure is only 100 first ones)

As it can be seen, the highest probability is given by *Model 3*, what makes sense because it is the most complex one. This model contains three parameters and can describe more boundaries since it takes into account horizontal and vertical coordinates and also owns a free parameter. *Model 1* and *Model 2* probability varies according to the samples. The second one is more complex since it has one extra parameter and takes into account both coordinates of the elements x_1 and x_2, whereas the first one only takes into account x_1. Depending on how the sample elements are distributed, the first or the second model will classify it better: if the model can be classified

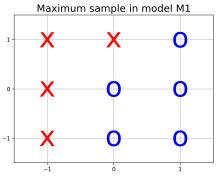
only considering x_1, the first one will do it better since it is the most simple one (Occam's razor); otherwise, the second model will perform the task better. *Model 0* has a constant probability of $\frac{1}{512}$ since it does not depend on any parameter.

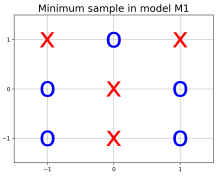
Question 23:



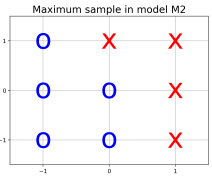


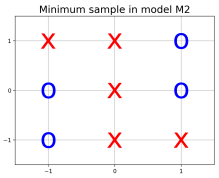
(a) Maximum/Minimum*Model 0*



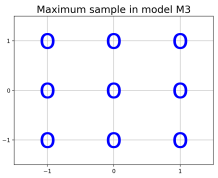


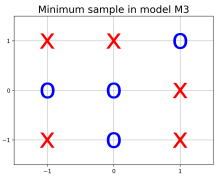
(b) Maximum/Minimum *Model 1*





(c) Maximum/Minimum*Model 2*





(d) Maximum/Minimum *Model 3*

The figures above show the sample that provided the maximum and minimum probability in a model respectively.

- According to *Model 0*, it makes sense that the best and worse probability are generated by the same sample since the probability is constant over all the samples.
- *Model 1* can only define vertical lines, so it makes sense that the best probability is achieved in one partially-separable by vertical lines and that the worst is found in one impossible to do so.
- Regarding *Model 2*, it takes into account both vertical and horizontal positions, so it is logical that the maximum sample can be separated by them and that the minimum results impossible to split.
- Finally, regarding *Model 3*, the best probability is found in a sample full of circles, thus this model is the only one which can classify it correctly due to its third parameter. The minimum one is found in a sample that is impossible to separate by linear boundaries.

Question 24:

The parameters $\boldsymbol{\theta}$ of the model are randomly generated from a normal distribution. Changing the characteristics of the distribution produces the following:

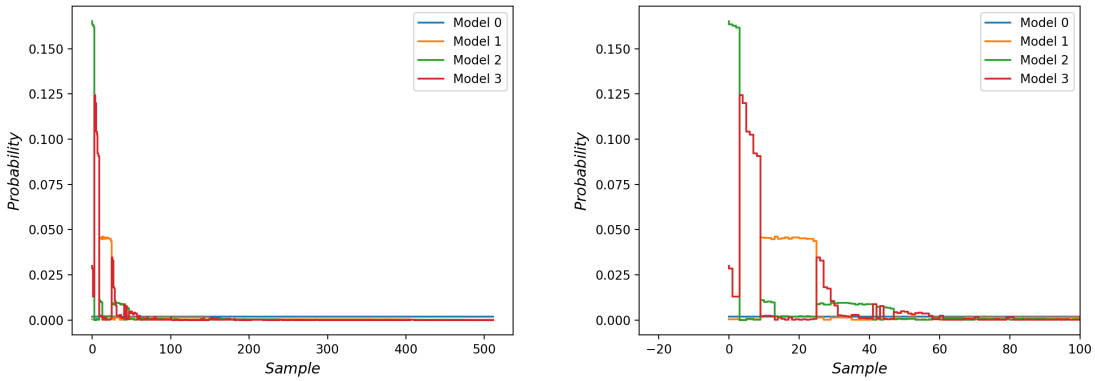


Figure 20: Evidence over dataset with non-diagonal matrix

For the case of non-diagonal covariance matrix the following values have been used:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 39.40273397 & 72.52896578 & 59.89545561 \\ 72.52896578 & 146.26307346 & 130.77208632 \\ 59.89545561 & 130.77208632 & 150.35368032 \end{bmatrix}$$

As it can be seen, when using a non-diagonal (but semi-definite) diagonal matrix, the obtained plot is quite different to the original one. It presents some peaks in the distribution and higher maximum probabilities. This can be consequence of the positive correlation between dimensions, which denotes a great dependence.

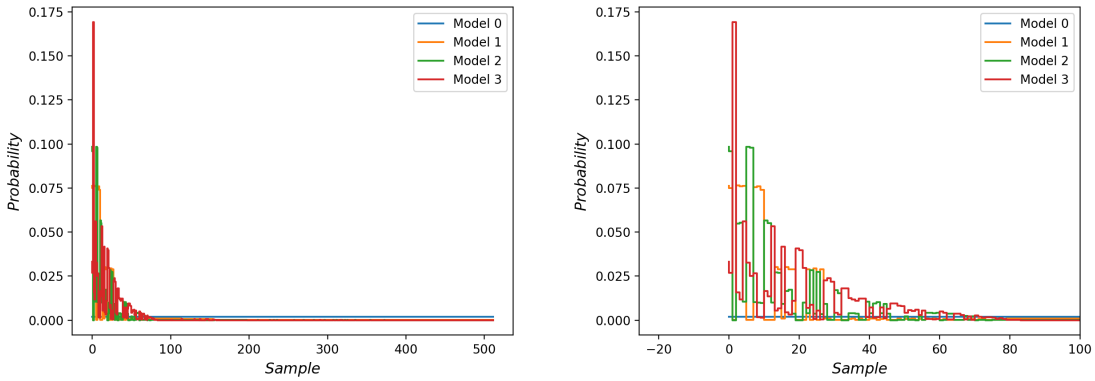


Figure 21: Evidence over dataset with $\mu = 5$

When the mean is equal to five, it is possible to achieve higher probabilities with some data

sets, but it can be seen that there are more peaks in the distribution. This is due to the fact that not centering the mean to zero makes having higher probabilities in those samples which have a value close to 5, but reducing the probability in those which have not. This is why in this case, the plot is much less flat in some points than the original one.

References

[1] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[2] C. M. Bishop, *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[3] I. M. Z. Ghahramani, "A note on the evidence and bayesian occam's razor", Gatsby Unit Technical Report GCNU-TR 2005–003, Tech. Rep., 2005.