
# Introduction to machine learning
-
# *Project 2: Bias and variance analysis*

Deflandre Sophie - s160767
Gómez Herrera María Andrea Liliana - s198387

Master 2 Biomedical engineering
Master 1 Data Science

*Academic year 2020-2021*

# 1 Analytical derivations

## 1.1 Bayes model and residual error in classification

### a. Analytical formulation of the Bayes model

Considering a binary classification problem with an output $y \in \{-1, +1\}$ and two real input variables $x_0$ and $x_1$. Each sample $x^i = (x_0^i, x_1^i)$ is generated by selecting its class $y^i$ at random with an equal probability for each class, and then drawing their values from a multivariate Gaussian distribution

$$\begin{pmatrix} x_0^i \\ x_1^i \end{pmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^i \\ \rho^i & 1 \end{bmatrix} \right)$$

where $\rho^i$ is class-dependent, *i.e.*, $\rho^i = \rho^+ > 0$ if $y^i = +1$ and $\rho^i = \rho^- = -\rho^+$ if $y^i = -1$

As we have a classification problem, the prediction made by the model is the one that maximize the probability $\mathbf{P}(y|x)$. In case of a binary response, if $\mathbf{P}(y = 1|x) > \mathbf{P}(y = -1|x)$ then the Bayes classifier, h(x), is equal to 1 and otherwise is equal to -1.

We have to calculate $\mathbf{P}(Y = y|\overline{X} = \overline{x}) = \frac{\mathbf{P}(\overline{X}=\overline{x}|Y=y)\mathbf{P}(Y=y)}{\mathbf{P}(\overline{X}=\overline{x})}$,
with $y \in \{1, -1\}, \overline{X} = (x_0, x_1), \overline{x}^i = (x_0^i, x_1^i)$.

We notice that for our hypothesis:

$$\mathbf{P}(Y = y) = \begin{cases} \mathbf{P}(Y = +1) = 0.5 \\ \mathbf{P}(Y = -1) = 0.5 \end{cases}$$

$$\mathbf{P}(\overline{X} = \overline{x}|Y = y) = \begin{cases} f_1(x_0, x_1) & if\ y = +1 \\ f_2(x_0, x_1) & if\ y = -1 \end{cases}$$

With $f_1(x_0, x_1)$ the joint density of a bivariate normal distribution of parameters $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma_1 = \begin{pmatrix} 1 & \rho^+ \\ \rho^+ & 1 \end{pmatrix}$. And $f_2(x_0, x_1)$ a bivariate normal distribution of parameters $\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 1 & -\rho^+ \\ -\rho^+ & 1 \end{pmatrix}$.

Now,

$$\begin{aligned} \mathbf{P}(\overline{X} = \overline{x}) &= \mathbf{P}(\overline{X} = \overline{X}|y = 1)\mathbf{P}(y = 1) + \mathbf{P}(\overline{X} = \overline{X}|y = -1)\mathbf{P}(y = -1) \\ &= \mathrm{P}\ (\mathrm{y} = 1)\ \mathrm{f}_1(x_0, x_1) + \mathbf{P}(y = -1) f_2(x_0, x_1) \end{aligned}$$

$$h(\overline{x}) = \begin{cases} 1 & if \quad P(1|x) > p(-1|x) \\ \\ -1 & otherwise \end{cases}$$

We have that, for $h(\overline{x}) = 1$,

$$P(y = 1|\overline{x}) > p(y = -1|\overline{x}) \iff \frac{\mathbf{P}(\overline{X} = \overline{x}|Y = 1)\mathbf{P}(Y = 1)}{\mathbf{P}(\overline{X} = \overline{x})} > \frac{\mathbf{P}(\overline{X} = \overline{x}|Y = -1)\mathbf{P}(Y = -1)}{\mathbf{P}(\overline{X} = \overline{x})}$$
$$\iff \mathbf{P}(\overline{X} = \overline{x}|Y = 1) > \mathbf{P}(\overline{X} = \overline{x}|Y = -1) \tag{1}$$

Because we know that $P(y = -1) = P(y = 1) = 0.5$ As the probability $P(\overline{X} = \overline{x}|Y = y)$ is given by a bivariate normal distribution, we can express $f_1$ and $f_2$, previously defined, as:

$$\frac{1}{2\pi|\Sigma_1|^{\frac{1}{2}}} e^{\frac{-(\overline{x}^T\Sigma_1^{-1}\overline{x})}{2}} > \frac{1}{2\pi|\Sigma_2|^{\frac{1}{2}}} e^{\frac{-(\overline{x}^T\Sigma_2^{-1}\overline{x})}{2}} \iff e^{\frac{-(\overline{x}^T\Sigma_1^{-1}\overline{x})}{2}} > e^{\frac{-(\overline{x}^T\Sigma_2^{-1}\overline{x})}{2}} \iff (\overline{x}^T\Sigma_1^{-1}\overline{x}) < (\overline{x}^T\Sigma_2^{-1}\overline{x})$$
$$\iff x_0(x_0 - \rho x_1) + x_1(x_1 - \rho x_0) < x_0(x_0 + \rho x_1) + x_1(x_1 + \rho x_0) \iff 0 < x_0 * x_1 \tag{2}$$

We can rewrite $h(\overline{x})$ as:

$$h(\overline{x}) = \begin{cases} 1 & if \quad 0 < x_0 * x_1 \\ \\ -1 & otherwise \end{cases}$$

We can verify this condition by looking at the data distribution when generating random samples following the bivariante Gaussian.

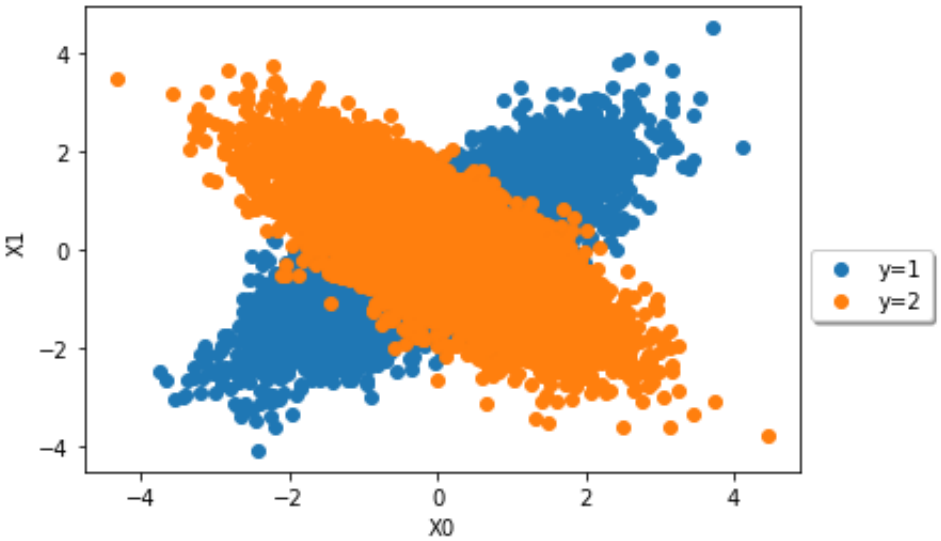


Figure 1: Generated samples

**b. Analytical formulation of the residual error**

Derive an analytical formulation of the residual error,i.e., the generalization error of the Bayes model:

$$E_{x_0,x_1,y}\left\{1\left(y \neq h_b\left(x_0,x_1\right)\right)\right\}$$

We have that,

$$\mathbf{E} = \int_{-\infty}^{\infty}[\mathbf{P}(\overline{x}, y=-1)\mathbf{1}(h_b(\overline{x})=1)\mathbf{P}(\overline{x}, y=1)\mathbf{1}(h_b(\overline{x})=-1)]dx$$

From $\mathbf{P}(\overline{x}, y) = \mathbf{P}(y)\mathbf{P}(\overline{x}|y)$, we can write:

$$\mathbf{E} = \frac{1}{2}(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[\frac{1}{2\pi\sqrt{1-(-\rho^+)2}}e^{-\frac{1}{2}(x_0^2-2\rho^+x_0x_1+x_1^2)}\mathbf{1}(h_b=1)$$

$$+\frac{1}{2\pi\sqrt{1-(\rho^+)2}}e^{-\frac{1}{2}(x_0^2+2\rho^+x_0x_1+x_1^2)}\mathbf{1}(h=-1)]dx_0dx_1)$$

with $\rho^+ = 0.75$

To compute the value of this integral we need to separate it in four part to meet the condition on $h_b$.

$$\mathbf{E} = \frac{1}{4\pi\sqrt{1-(\rho^+)^2}}[\ (\int_{-\infty}^{0}\int_{-\infty}^{0}e^{-\frac{1}{2}(x_0^2-2\rho^+x_0x_1+x_1^2)}dx_0dx_1 + \int_{0}^{+\infty}\int_{0}^{+\infty}e^{-\frac{1}{2}(x_0^2-2\rho^+x_0x_1+x_1^2)}dx_0dx_1)$$

$$+(\int_{-\infty}^{0}\int_{0}^{+\infty}e^{-\frac{1}{2}(x_0^2+2\rho^+x_0x_1+x_1^2)}dx_0dx_1 + \int_{0}^{+\infty}\int_{-\infty}^{0}e^{-\frac{1}{2}(x_0^2+2\rho^+x_0x_1+x_1^2)}dx_0dx_1)]$$

Finally, by considering values of $x_0$ and $x_1$ between -5 and 5 (see figure 1), the error can be computed using the matlab code *error.m*. The error we found with this estimation is 1.756. Obviously, it isn't correct by the way there is an error in our expression but we didn't find it.

Let's now see empirically what should be the error. We plotted both bivariate Gaussians and looked at the area under their intersection (see figure 2. This corresponds to the error. These Gaussians are given by :

$$f_1(x_0,x_1) = \frac{1}{2\pi\sqrt{1-(0.75)^2}}e^{-\frac{1}{2}(x_0^2-2(0.75)x_0x_1+x_1^2)}\ ,\ f_2(x_0,x_1) = \frac{1}{2\pi\sqrt{1-(-0.75)^2}}e^{-\frac{1}{2}(x_0^2+2(0.75)x_0x_1+x_1^2)}$$

To obtain this error, we generated several large data sets and calculated the zero-error of each one. After 1000 simulations we obtained that the average error is 0.4900.

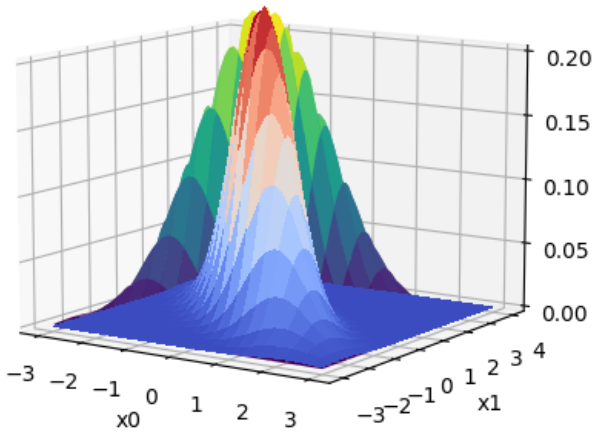


Figure 2: Illustration of both gaussians

## 1.2 Bias and variance of ridge regression

### a.Ordinary least-square and Ridge regression

Assuming that X is orthogonal, show that

$$w_R = \frac{w_{OLS}}{1+\lambda}$$

*Proof.* We have that $w_{OLS} = (X^T X)^{-1} X^T Y$ and the problem of the Ridge regression is reduced to find a $w_R$ such that minimizes the next function:
$RR(w_R) = (Y - Xw_R)^T (Y - Xw_R) + \lambda ||w_R||^2$, $\lambda > 0$ and $||w_R||^2 = \sum w_i^2$
$RR(w_R) = (Y - Xw_R)^T (Y - Xw_R) + \lambda w_R^T w_R$
$\frac{\partial RR}{\partial w_R} = -2X^T(Y - Xw_R) + 2\lambda w_R$

Now, we will calculate the minimum:

$$\begin{aligned} \frac{\partial RR}{\partial w_R} = 0 &\iff -X(Y - Xw_R) - \lambda w_R = 0 \\ &\iff X^T Y - (X^T X) w_R - \lambda w_R = 0 \\ &\iff (X^T X) w_R + \lambda w_R = X^T Y \\ &\iff (X^T X + \lambda I d_p) w_R = X^T Y \end{aligned} \tag{3}$$

If $X^T X + \lambda I d_p$ is invertible, then $w_R = (X^T X + \lambda I d_p)^{-1} X^T Y$.
P.D. $X^T X + \lambda I d_p$ is invertible.

*Proof.* $X^TX \in M_{pxp}(\mathbf{R})$ and is symmetric, therefore $X^TX = Q^TDQ$ with Q and orthogonal matrix and D a diagonal matrix.
$D = diag\{\lambda_1, ...\lambda_p\}$ where $\lambda_i$ is a proper value of $X^TX$. Then:

$$det(X^TX) = det(D) = \lambda_1\lambda_2...\lambda_p$$

We have that:

$$X^TX + \lambda Id_p = Q^TDQ + \lambda Q^TId_pQ = Q^T(D + \lambda Id_p)Q$$

Then,

$$det(X^TX + \lambda Id_p) = det(D + \lambda Id_p) = (\lambda_1 + \lambda)(\lambda_2 + \lambda)...(\lambda_p + \lambda)$$
$$D + \lambda Id_p = diag\{\lambda_1 + \lambda, \lambda_2 + \lambda, ..., \lambda_p + \lambda\}$$

As $\lambda > 0$ thus $\lambda_j + \lambda > 0, \forall j = 1, ..., p$, from the fact that $X^TX$ is positive semidefined we have that $\lambda_j \geq 0$.
Therefore $(X^TX + \lambda Id_p) > 0$ and from this $X^TX + \lambda Id$ is invertible. □

We conclude that $w_R = (X^TX + \lambda Id_p)^{-1}X^TY$ and as X is orthoghonal,

$$w_R = \frac{w_{OLS}}{1+\lambda}$$

□

**b. Relationships between bias and variance**

We compute the Bias and Variance of $x^Tw_{OLS}$ and $\frac{x^Tw_{OLS}}{1+\lambda}$, assuming that we use $y\hat{(}x) = \frac{x^Tw_{OLS}}{1+\lambda}$.

$$\begin{aligned}
var(y\hat{(}x)) &= var(\frac{x^Tw_{OLS}}{1+\lambda}) \\
&= \frac{1}{(1+\lambda)^2}var(x^Tw_{OLS}) \\
&= \frac{1}{(1+\lambda)^2}\mathbf{E}[x^Tw_{OLS} - x^T\mu]^2 \\
&= \frac{1}{(1+\lambda)^2}\mathbf{E}[x^T(w_{OLS} - \mu)]^2 = \frac{1}{(1+\lambda)^2}\mathbf{E}[x^T(w_{OLS} - \mu)x^T(w_{OLS} - \mu)] \\
&= \frac{1}{(1+\lambda)^2}\mathbf{E}[x^T(w_{OLS} - \mu)(w_{OLS} - \mu)^Tx] \\
&= \frac{1}{(1+\lambda)^2}x^T\mathbf{E}[(w_{OLS} - \mu)(w_{OLS} - \mu)^T]x \\
&= \frac{1}{(1+\lambda)^2}x^T\Sigma x
\end{aligned} \tag{4}$$

Where $\mu$ is $Xw$ and $\Sigma = \sigma^2Id$ given the normality of the residuals.

$$
\begin{aligned}
bias^2(y\hat{(x)}) &= (\mathbf{E}_y(y(x)) - \mathbf{E}_{LS}(y\hat{(x)}))^2 \\
&= (\mathbf{E}_y(f(x) + \epsilon) - \mathbf{E}_{LS}(\frac{x^T w_{OLS}}{1+\lambda}))^2 \\
&= (\mathbf{E}_y(f(x)) - \frac{1}{1+\lambda}\mathbf{E}(x^T w_{OLS}))^2 \\
&= (f(x) - \frac{x^T}{1+\lambda}\mathbf{E}(w_{OLS}))^2 \\
&= (f(x) - \frac{x^T w}{1+\lambda})^2
\end{aligned}
\tag{5}
$$

Let's now compute the variance and bias of $x^T w_{OLS}$:

Because of the equation (2) we have that $Var(x^T w_{OLS}) = x^T \Sigma x$ with $\Sigma = \sigma^2 Id$

And we have that $Bias(x^T w_{OLS}) = (f(x) - x^T w)^2$

Finally, we can observe that the variance of $x^T w_{OLS}$ and $\frac{x^T w_{OLS}}{1+\lambda}$ only differs by a constant and the smaller the value of $\lambda$ the bigger the variance, given that $\lambda > 0$. For the bias, the relation between $\lambda$ and this parameter is the opposite, when the value of $\lambda$ increases, the value of the bias increases nevertheless, this increase is not proportional.

On the basis of these formulas, we can explain the impact of $\lambda$ on bias and variance. Indeed, we notice that the value of $\lambda$ is inversely proportional tho the value of the variance. The size of this parameter is a Bias-Variance trade-off decision between the fit versus the size of the coefficients. If the value of $\lambda = 0$ we would have a simple regression and if $\lambda \to \infty$ all the coefficients of the regression would get closer to zero, ie , we would have a model with just the intercept. In a Ridge regression the mean square error, the sum of variance and bias is minimized (depending on the value of $\lambda$), and becomes lower than for the full least squares estimate.

# 2 Empirical analysis

**a. Analytical expressions at $x_0$**

The residual error is given by :
$noise(x_0) = Var_{y|x_0}\{f(x_o) + \epsilon\} = E_{y|x_0}\{(f(x_0) + \epsilon - E_{y|x_0}\{f(x_0) + \epsilon\})^2\}$,
Knowing the definition of the esperance of a scalar and the linear property, we can reduce this expression to :
$noise(x_0) = E_{y|x_0}\{(f(x_0) + \epsilon - f(x_0))^2\} = E_{y|x_0}\{\epsilon^2\} = \sigma^2$

The squared bias is given by:
$bias^2(x_0) = (E_{y|x_0}\{f(x_0) + \epsilon\} - E_{LS}\{\hat{y}(x_0)\})^2 = (f(x_0) - E_{LS}\{\hat{y}(x_0)\})^2$

The variance is given by:
$variance(x_0) = E_{LS}\{(\hat{y}(x_0) - E_{LS}\{\hat{y}(x_0)\})^2\}$

Finally, the expected error is given by:
$E = \sigma^2 + (f(x_0) - E_{LS}\{\hat{y}(x_0)\})^2 + E_{LS}\{(\hat{y}(x_0) - E_{LS}\{\hat{y}(x_0)\})^2\}$

### b. Experimental protocol

1. Generate a huge (in theory infinity) number of learning sample sets each of size N assumed to be huge (also infinity in theory) thanks to the function and noise distribution that are known.
2. Creating a new set with all (x,y) samples for which $x = x_0$.
3. Computing an output vector with all the output values of this set.
4. Computing the residual error by taking the variance of the output vector created in the previous step and the Bayes model by taking the mean.
5. Training the learning algorithm on each learning sample.
6. Computing $\hat{y}$ with $x = x_0$ for each trained model and make a vector with them.
7. Computing squared bias by taking the difference between the Bayes model and the mean of the $\hat{y}$ vector computed in the previous step.
8. Computing the variance of the learning algorithm by taking the variance of the $\hat{y}$ vector.
9. Computing the expected error by taking the sum of the residual error, squared bias and variance.

### c. Bayes model and residual error

To estimate the Bayes model and the residual error, we need to follow the steps one to four of the procedure. Indeed, the Bayes model is made by taking the mean of a huge numbers of output vectors. The differences between each output vectors is determined by the amount of noise. This noise follow a normal distribution of mean 0 and standard deviation $\sqrt{0.1}$ meaning that output values for a same input x differ from each other due to a noise value taken randomly following the normale law with the random.normale function of numpy.
By the way, the Bayes model computed on an infinity numbers of output vectors is a perfect estimation of the real output. In figure 3 is shown the Bayes model that we obtain with 5000 samples and 2000 values of the input x compared to the function f(x) without noise (the perfect output). Like we can see the Bayes model approximates well the function even if we have taken 1000 samples and not an infinite number (not possible in practise).

In figure 4 is shown the residual error. The analytical value found is $\sigma^2 = 0.1$ and as we can see the values are around 0.1 in the plot. It is not perfectly 0.1 because we have not taken an infinite number of samples but if we increase the number of samples, the variance is closer to 0.1.

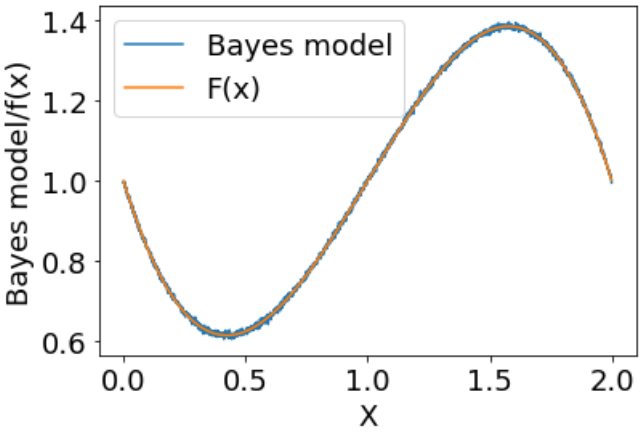


Figure 3: Bayes model

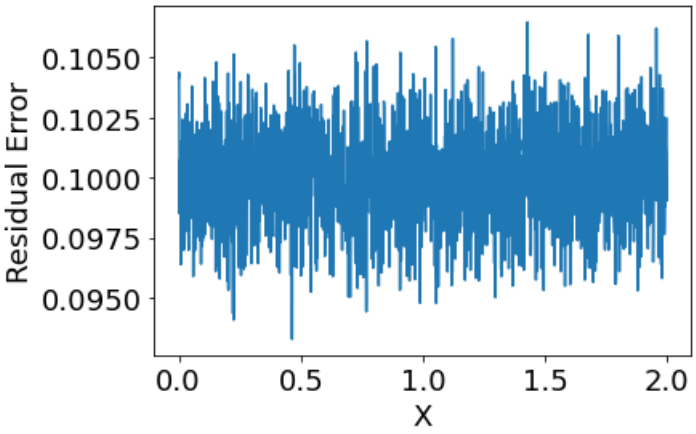


Figure 4: Residual error

### d. Bias, Variance and Expected Error

In this section are analysed the bias, variance and expected error of the learning algorithm using the method LinearRegression from scikitlearn. To compute these quantites the procedure has been followed and f(x) has been used instead of $E_y\{y\}$. Moreover, as it isn't possible to take an infinity of learning samples, we have chosen to take 1000 learning sample sets.

Lets's now discuss the impact of the linear regression degree. Firstly, we notice that with a degree of 0 the bias is the higher for x=0.5 and 1.75 but is null for x=0 and 1. With a degree of 1 or two the bias is the higher for x = 0. and 1.75 but also for x=0 but the bias is still null for x=1. With higher degree the bias is always null.

According to the variance, it seems higher with higher degrees especially for x=0. However, for x=1, degrees following each other has similar variance (0 with 1, 2 with 3,...). For x=0.5 and 1.75, it is a bit more complicated, the variance still increases with degree but some variance are really close to each other or even equal.

Obvisouly, as the expected error is the sum of the residual error, squared bias and variance, it increases with the degree too. Degree 0 seems really bad especially for x=0.5 and 1.75. According to degree 1 and 2 they are really bad for x=0 and 2.

According to these observations it is obvious that the best degrees to estimate the function are 4 and 5.

### e. Error in function of the degree

In this section, we look at the evolution of the error with the number of degrees by taking the mean of each error (bias, variance, expected error) over all inputs x. By the way, the results can be expressed in function of the degree but do not depend of x anymore.

First of all, the bias is null with a degree higher or equal than three meaning that the estimated model perfectly fits the real output. On the contrary, the variance between learning sample is 0 with a degree of 0 and increase with increasing degree. Meaning that with higher

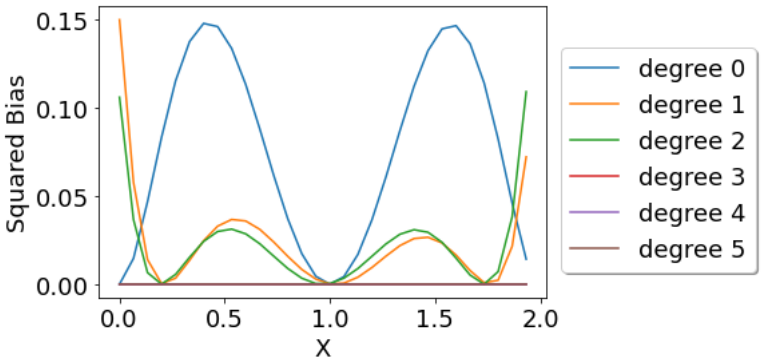


Figure 5: Squared Bias

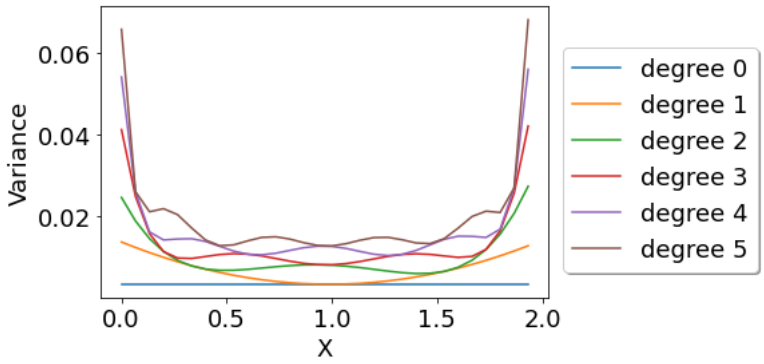


Figure 6: Variance

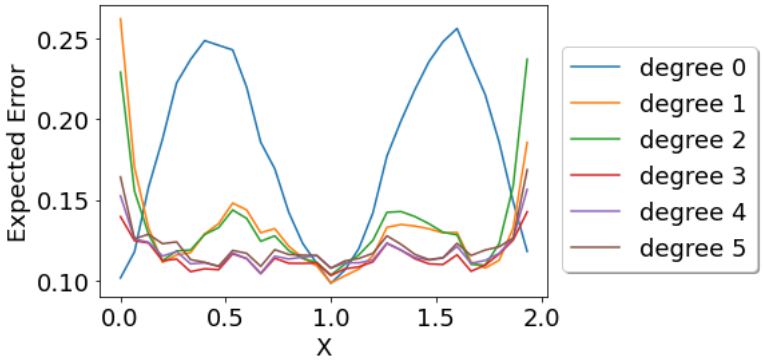


Figure 7: Expected error

degree the error will depend more on the learning sample sets used.

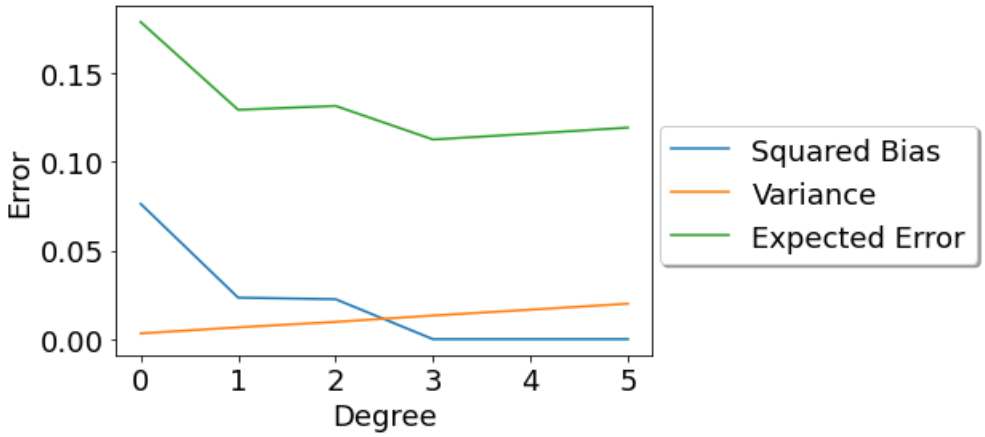


Figure 8: Error in function of the degree

### f. Ridge regression

In this last section will be observed the errors in function of the regularisation parameter by using a ridge regression instead of the ordinary least square regression as used before.

First, in figure 9 are shown the evolution of the squared bias, variance and expected error

in function of the input x. We can already see that the value of the regularisation level doesn't seem to influence the errors when it is higher than 0.

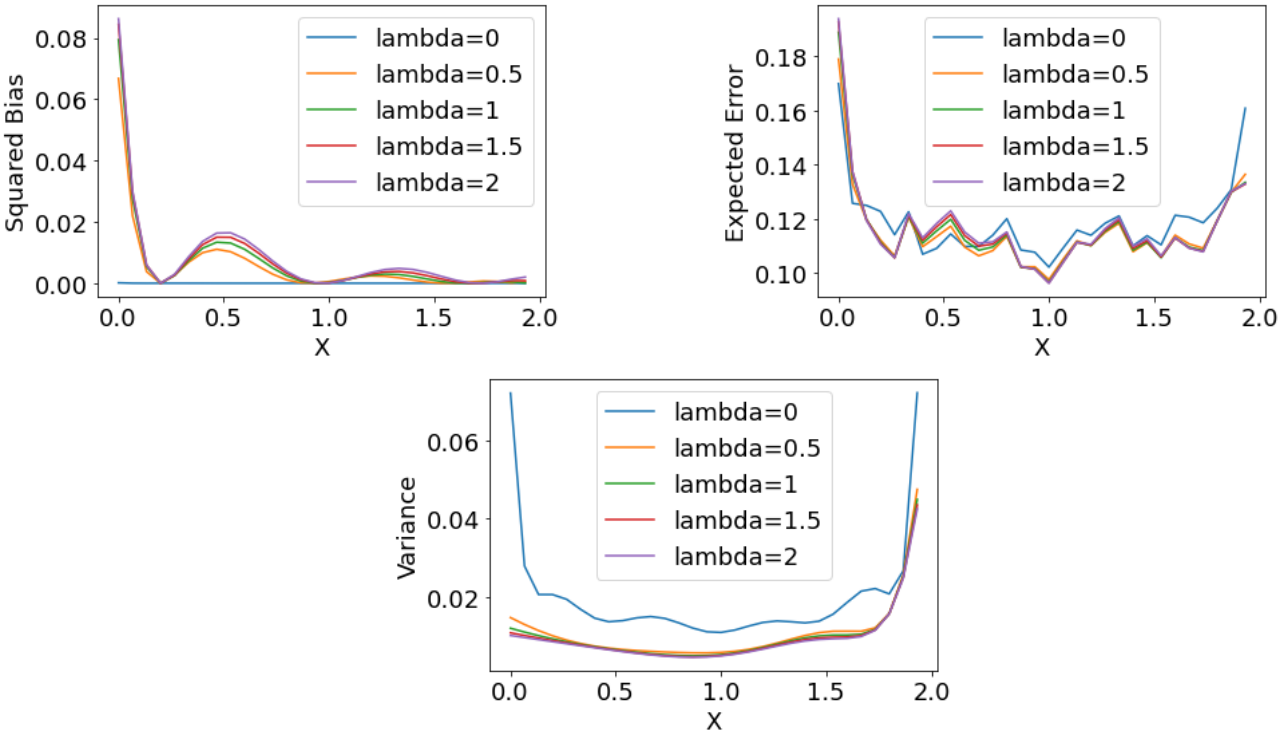


Figure 9: Errors

In figure 10 is plotted the evolution of the errors in function of the regularisation level. In this case the variance decreases when $\lambda$ is higher than 0 but the variance increases with $\lambda$. This means that the model is better estimated with $\lambda = 0$. But as in almost every case, the dependency of the learning sample sets will increase, meaning that we have a compromise between variance and bias.
Finally, the expected error is almost not influenced by $\lambda$

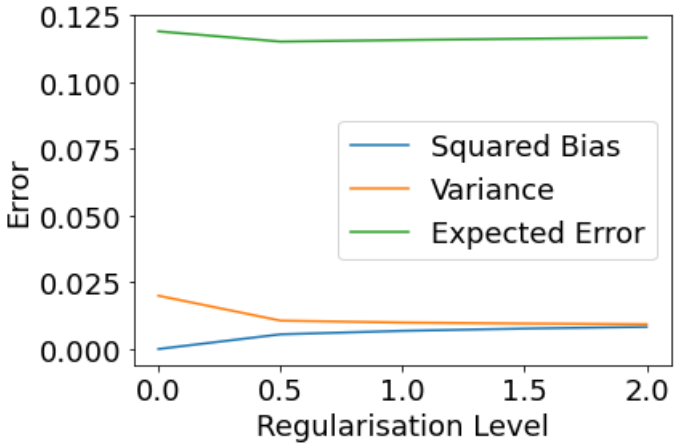


Figure 10: Error in function of the regularization