University of Liège
MSc in Data Science and Engineering
Academic year 2020-2021

ELEN0062 - Introduction to Machine Learning

Project 2 - Bias and variance analysis

Author:
BAUDINET Charles – s164489
MALAY THIBAULT – s164812

Instructors:
PROF. L.WEHENKEL
PROF. P. GEURTS
ANTONIO SUTERA

LIÈGE
Novembre 2020

Introduction

In this project, we will try to better understand the notion of bias and variance. We will visualize the concept of Bias-variance trade-off through different examples. The first part will be very theoretical and we will then use these results (especially for Ridge) in the second part of the project.

1 Analytical derivations

1.1 Analytical formulation of the Bayes model

The Bayes model is given by

$$h_B(x_0, x_1) = \arg\max_c P(y = c|x_0, x_1)$$

Since we only have 2 possible outputs ($y \in \{-1, 1\}$), we can rewrite it as:

$$h_B(x_0, x_1) = \begin{cases} 1 & \text{If } P(y = 1|x_0, x_1) > P(y = -1|x_0, x_1) \\ -1 & \text{Otherwise} \end{cases} \tag{1}$$

The Bayes theorem tells us that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

The condition from the equation [1] in order to predict 1 is,

$$P(y = 1|x_0, x_1) > P(y = -1|x_0, x_1)$$

which can be transformed using the Bayes theorem [2] to,

$$\frac{P(x_0, x_1)|y = 1)P(y = 1)}{P(x_0, x_1)} > \frac{P(x_0, x_1|y = -1)P(y = -1)}{P(x_0, x_1)} \tag{3}$$

Since there is an equal probability of each class, $P(y = 1) = P(y = -1) = 0.5$, we obtain:

$$P(x_0, x_1|y = 1) > P(x_0, x_1|y = -1) \tag{4}$$

These two probability distributions are known since x_0^i and x_1^i are drawn from the multivariate Gaussian distribution,

$$P(\boldsymbol{x}|y = 1) = \frac{1}{2\pi\sqrt{\det \Sigma_+}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \mu)^T \Sigma_+^{-1} (\boldsymbol{x} - \mu)\right] \tag{5}$$

with

$$\Sigma_+ = \begin{bmatrix} 1 & \rho_+^i \\ \rho_+^i & 1 \end{bmatrix} \text{ and } x_0 \text{ and } x_1 \text{ are centered at the origin } (\mu = [0\ 0]).$$

Σ_- for $P(\boldsymbol{x}|y = -1)$ is defined in the same way except that $\rho_-^i = -\rho_+^i$

The determinant of Σ is $1 - \rho^2$ and has to be positive. Therefore, $|\rho|$, which is positive, has to bellow to]0,1]. We can observe that the determinant of the two probability distributions will be

equal since we take the square of ρ.

We can therefore transform, using the condition [4] and the probability distribution [5] of each class, to this probability condition:

$$\exp\left[-\frac{\mathbf{x}^T\Sigma_+^{-1}\mathbf{x}}{2}\right] > \exp\left[-\frac{\mathbf{x}^T\Sigma_-^{-1}\mathbf{x}}{2}\right]$$

We can take the `ln` in both side to obtain,

$$\mathbf{x}^T\Sigma_+^{-1}\mathbf{x} < \mathbf{x}^T\Sigma_-^{-1}\mathbf{x} \tag{6}$$

$$\Leftrightarrow \mathbf{x}^T[\Sigma_+ - \Sigma_-]\mathbf{x} < 0 \tag{7}$$

$$\Leftrightarrow \mathbf{x}^T\Sigma^*\mathbf{x} < 0 \tag{8}$$

Where $\Sigma^* = \Sigma_+ - \Sigma_-$

After some trivial developments and using the fact that $\rho_-^i = -\rho_+^i$, we obtain that

$$\Sigma^* = \frac{1}{1-\rho^{+2}}\begin{bmatrix} 0 & -2\rho_+^i \\ -2\rho_+^i & 0 \end{bmatrix}$$

The inequation [8] can be now expressed as follow,

$$\frac{-2}{1-\rho^{+2}}\mathbf{x}^T\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\mathbf{x} < 0$$

$$\Leftrightarrow \begin{bmatrix} x_2 x_1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} > 0$$

After developing this matrix multiplication, we obtain that the condition [4] can be rewrite to a simpler form as below

$$x_0 x_1 > 0 \tag{9}$$

Using [9], we can rewrite our Bayes model [1] as follow

$$h_B(x_0, x_1) = \begin{cases} 1 & \text{if } x_0 x_1 > 0, \\ -1 & \text{if } x_0 x_1 \leq 0. \end{cases} \tag{10}$$

This result is not surprising. Indeed, when y = 1, we create some kind of positive link (positive correlation) which will result in a same sign. If $x_O > 0$ then x_1 should also be most of the time > 0 and vice versa. Since we create the opposite construction for y = -1, it is logical we obtain such opposite results.

1.2 Analytical formulation of the residual error

By definition of the residual error, we have

$$E_{x_0,x_1,y}\{1(y \neq h_B(x_0, x_1)\}$$

$$= \iint P(\underline{x}|y=-1)\cdot P(y=-1)\cdot Loss(-1, h_b(x_0,x_1))\partial x_0 \partial x_1$$
$$+ \iint P(\underline{x}|y=1)\cdot P(y=1)\cdot Loss(1, h_b(x_0,x_1))\partial x_0 \partial x_1$$

where P(y = -1) = P(y = 1) = 0.5 and the Loss function is defined as follow

$$Loss = \begin{cases} 0 & \text{If } h_b(x_0,x_1) = y \\ 1 & \text{Otherwise} \end{cases} \tag{11}$$

These two integrals can therefore be separate according to the Loss function.

$$\begin{aligned}
=&\frac{1}{2}\int_0^{+\infty}\int_0^{+\infty}\frac{1}{2\pi\sqrt{1-\rho_-^2}}\exp\left[-\frac{1}{2\left(1-\rho_-\right)^2}\left(x_0^2-2x_0x_1\rho_-+x_1^2\right)\right]\partial x_0\partial x_1\\
&+\frac{1}{2}\int_{-\infty}^{0}\int_{-\infty}^{0}\frac{1}{2\pi\sqrt{1-\rho_-^2}}\exp\left[-\frac{1}{2\left(1-\rho_-\right)^2}\left(x_0^2-2x_0x_1\rho_-+x_1^2\right)\right]\partial x_0\partial x_1\\
&+\frac{1}{2}\int_0^{+\infty}\int_{-\infty}^{0}\frac{1}{2\pi\sqrt{1-\rho_+^2}}\exp\left[-\frac{1}{2\left(1-\rho_+\right)^2}\left(x_0^2-2x_0x_1\rho_++x_1^2\right)\right]\partial x_0\partial x_1\\
&+\frac{1}{2}\int_{-\infty}^{0}\int_0^{+\infty}\frac{1}{2\pi\sqrt{1-\rho_+^2}}\exp\left[-\frac{1}{2\left(1-\rho_+\right)^2}\left(x_0^2-2x_0x_1\rho_++x_1^2\right)\right]\partial x_0\partial x_1
\end{aligned}$$

By using the function `nquad` of the Python package `scipy.integrate` and replacing in the expression $\rho_- = -\rho_+ = 0.75$, we obtain 0.23005 as result of this sum of integrals.

If we generate a sample of points drawn from the multivariate Gaussian distribution and then estimate the outputs using [10], we can compute an error of 0.23075 which is very close to our analytical computation.

1.3 Best w_R assuming X is orthogonal

In the ordinary least-square, we want the argmin of a classical mean squared error,

$$\hat{w}_{OLS} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - x_i^T w)^2$$

$$\begin{aligned}
&\Rightarrow -x^T y - x^T y + 2x^T x w = 0\\
&\Leftrightarrow x^T x w = x^T y\\
&\Leftrightarrow \hat{w}_{OLS} = (x^T x)^{-1} x^T y
\end{aligned}$$

But in practice we often use the Ridge (and Lasso) regression. The Ridge regression belongs to the class of regression tools that use L2 regularization. The L2 penalty is used to penalize the square of the coefficients. This is a very powerful tool used to simplify the model we are dealing with. The only weakness of the L2 regularization is the inefficient to obtain a sparse coefficient vector.

Let's try to find the best Ridge estimator for our model using a similar Loss function than the ordinary least-square with the L2 penalty

$$\hat{w}_{ridge} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - x_i^T w)^2 + \lambda w^t w$$

Differentiating with respect to w and equals it to 0 will lead us to this next equality:

$$\begin{aligned}
&\Rightarrow -2x^T y + 2x^T x w + 2\lambda w = 0 \\
&\Leftrightarrow -x^T y + x^T x w + \lambda w = 0 \\
&\Leftrightarrow (x^T x + \lambda) w = x^T y \\
&\Leftrightarrow w = \frac{x^T y}{(x^T x + \lambda)}
\end{aligned}$$

where $x^T y = (x^T x)^{-1} x^T y = \hat{w}_{OLS}$ since x is orthogonal (i.e $x^T x = \mathbb{1}$)

Therefore, the best Ridge regression can be expressed in this form,

$$\hat{w} = \frac{\hat{w}_{OLS}}{1+\lambda}$$

Which is the minimum since MSE is a convex function.

1.4 Bias and Variance discussion

1.4.1 Relationships between the bias and variance of the OLS and the Ridge

- **Bias**:

$$\begin{aligned}
&Bias_{Ridge} = \mathbb{E}_y[y] - \mathbb{E}_{LS}[\hat{y}_{ridge}] \\
&\Leftrightarrow Bias_{Ridge} = \mathbb{E}_y[y] - \frac{1}{1+\lambda}\mathbb{E}_{LS}\left[x^T w_{OLS}\right] \\
&\Leftrightarrow Bias_{Ridge} = \mathbb{E}_y[y] - \mathbb{E}_{LS}\left[x^T w_{OLS}\right] + \frac{\lambda}{1+\lambda}\mathbb{E}_{LS}\left[x^T w_{OLS}\right] \\
&\Leftrightarrow Bias_{Ridge} = \mathbb{E}_y[y] - \mathbb{E}_{LS}\left[\hat{y}_{OLS}\right] + \frac{\lambda}{1+\lambda}\mathbb{E}_{LS}\left[x^T w_{OLS}\right] \\
&\Leftrightarrow Bias_{Ridge} = Bias_{OLS} + \frac{\lambda}{1+\lambda}\mathbb{E}_{LS}\left[x^T w_{OLS}\right]
\end{aligned}$$

- **Variance**:

$$
\begin{aligned}
&Variance_{Ridge} = \mathbb{E}_{LS}\left[\left(\hat{y}_{Ridge} - \mathbb{E}_{LS}\left[\hat{y}_{Ridge}\right]\right)^2\right] \\
\Leftrightarrow\ &Variance_{Ridge} = \mathbb{E}_{LS}\left[\left(\frac{x^T w_{OLS}}{1+\lambda} - \mathbb{E}_{LS}\left[\frac{x^T w_{OLS}}{1+\lambda}\right]\right)^2\right] \\
\Leftrightarrow\ &Variance_{Ridge} = \mathbb{E}_{LS}\left[\frac{1}{(1+\lambda)^2}\left(x^T w_{OLS} - \mathbb{E}_{LS}\left[x^T w_{OLS}\right]\right)^2\right] \\
\Leftrightarrow\ &Variance_{Ridge} = \frac{1}{(1+\lambda)^2}\mathbb{E}_{LS}\left[\left(x^T w_{OLS} - \mathbb{E}_{LS}\left[x^T w_{OLS}\right]\right)^2\right] \\
\Leftrightarrow\ &Variance_{Ridge} = \frac{1}{(1+\lambda)^2} Variance_{OLS}
\end{aligned}
$$

1.4.2 Discussion on λ on bias and variance

First, we can observe that if $\lambda = 0$, the Ridge Bias and Variance are equivalent to the ordinary least-square. This was of course expected since the two minimizing problems are equivalent with $\lambda = 0$. Second, the variance should decrease drastically when λ increases and the Bias increases with respect to λ.

2 Empirical analyses

In this second part, we are dealing with a regression problem where $y = f(x) + \epsilon$ where ϵ follows $\mathcal{N}(0,1)$ and x follows $\mathcal{U}(0,2)$.

2.1 Analytical expressions of the residual error, squared bias, variance and expected error at a given point x_0

The residual error correspond to the difference between the observed and predicted output of our model.

$$\mathbb{E}_{y|x_0}\left[\left(y - \mathbb{E}_{y|x_0}[y]\right)^2\right]$$

The square bias measures the error between the Bayes model and the average model.

$$\left(\mathbb{E}_{y|x_0}[y] - \mathbb{E}_{LS}[\hat{y}]\right)^2$$

The variance quantifies the variation of my y from one learning to another.

$$\mathbb{E}_{LS}\left[\left(\hat{y} - \mathbb{E}_{LS}\left[\hat{y}\right]\right)^2\right]$$

And finally, the expected error is the sum of the three first measures.

2.2 Experimental protocol

In order to estimate the measures defined below, we will proceed as follow:

First, we will create n_ls (= 100) learning samples of N (= 30) points in order to train our linear regression model. All these models will then be used to predict the same new fixed sample of size 500. The predictions for each sample will help us to construct our Bias, Variance and expected error computation.

- **Value of the Bayes model and the residual error**: In order to estimate this value, we associate for each x a number n_y of y_{pred} using the noisy function. We have to compute the difference between the noisy y associates to one x the Bayes model. Then, the residual error is obtained by taking the mean over all y of the squared difference for a fixed x.
- **The squared Bias**: For this one, we will first compute the expectation over all y associate to one x. Then, we have to compute the mean of our predictions for each Learning sample. The difference of these first two terms is the Bias which has to be taken to the square to obtain the squared bias.
- **The variance**: In order to compute the variance, we will, for each learning sample, take the difference between our predictions associated to one x and the mean overall learning samples of all predictions associated to this x. We then take the square of this difference. The variance is obtained by taking the mean, for a fixed x, of all learning samples.
- **The expected error**: We just have to sum the first three measures in order to compute the expected error.

2.3 Estimation and plot of the Bayes model and the residual error

The theoretical value of the residual error is given by

$$\begin{aligned}
&\mathbf{E}\left[\left(y - h_b(x)\right)^2\right] \\
&= \mathbf{E}\left[\left(f(x) + \epsilon - f(x)\right)^2\right] \\
&= \mathbf{E}\left[epsilon^2\right] \\
&= \mathbf{Var}\left[epsilon\right] - \mathbf{E}\left[\epsilon\right]^2 \\
&= 0.1^2 + 0 = 0.1^2
\end{aligned}$$

When we plot the real values with the Bayes model, we obtain this graph:

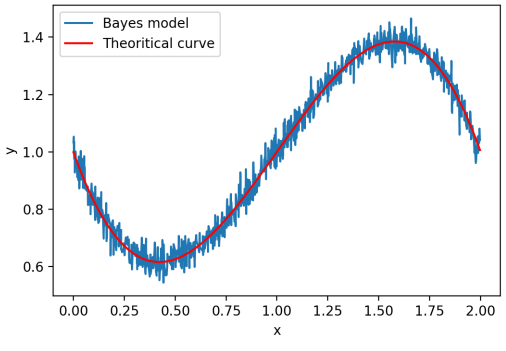

Figure 1: Difference between Analytical and Bayes model

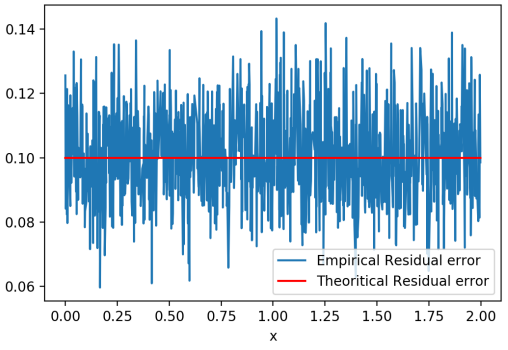

Figure 2: Residual error

2.4 Empirical estimation of Bias^2, Variance and the Expected Error

In this simulation, since $f(x) = \sum_{i=1}^{m} a_i \cdot x^i$, the higher m is, the higher the complexity of our function will be. For m = 5 for instance, a small change in the data set will probably result in a

completely different output. The function will probably fit in a better way our data but will be, we suppose, very sensitive to the noise (in other words, it will overfit our data). On the other hand, with m = 0 or 1, we expect quite a bad model in terms of bias but insensible to some variation in our data set. All plots below will try to answer these problems with some visualizations.

It can be seen through these plots that the higher m is, and so the complexity, the smaller our squared bias will be. It was expected and these plots confirm our feelings about it. Since the complexity increases, we will have more degrees of freedom to fit our data.

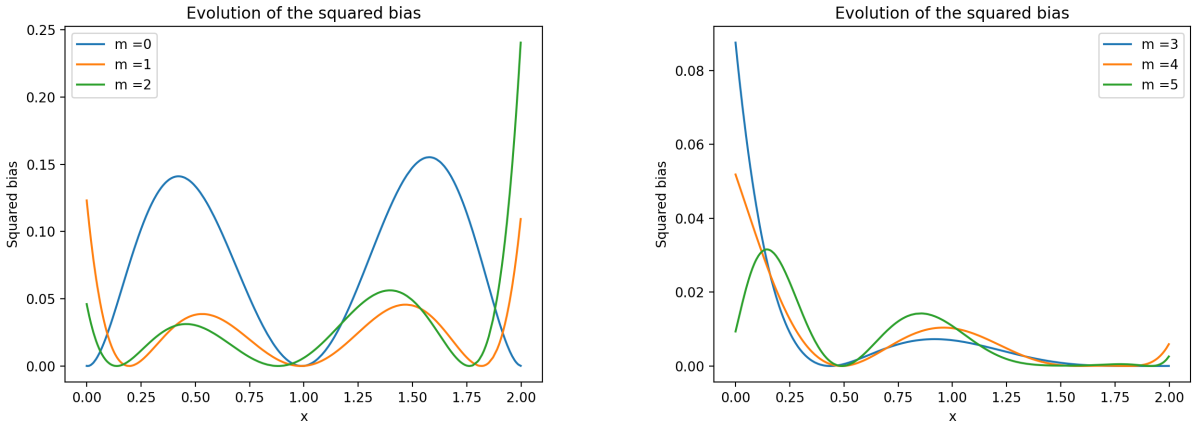

Figure 3: Evolution of the Squared bias w.r.t x

Let's try to explain one of our results (for m = 2 for instance). In this case, the Figure 4 represents the mean of all predictions made for m = 2 in orange, the real function in blue and our squared bias in green. We can see through this graph that when our mean is far away (e.g 0 and 0.5) from the real function the bias increases and vice versa (e.g 1 and 1.75).

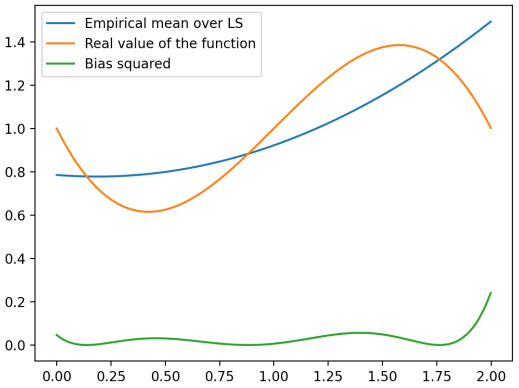

Figure 4: Bias explanation for m = 2

On Figure 5, the results are clear. The higher the complexity, the bigger the variance is. For m = 0, the variance is almost nonexistent. For m bigger than 2, the variance seems to be very sensitive to the learning sample. Notice that the scales are not the same between the two graphs.

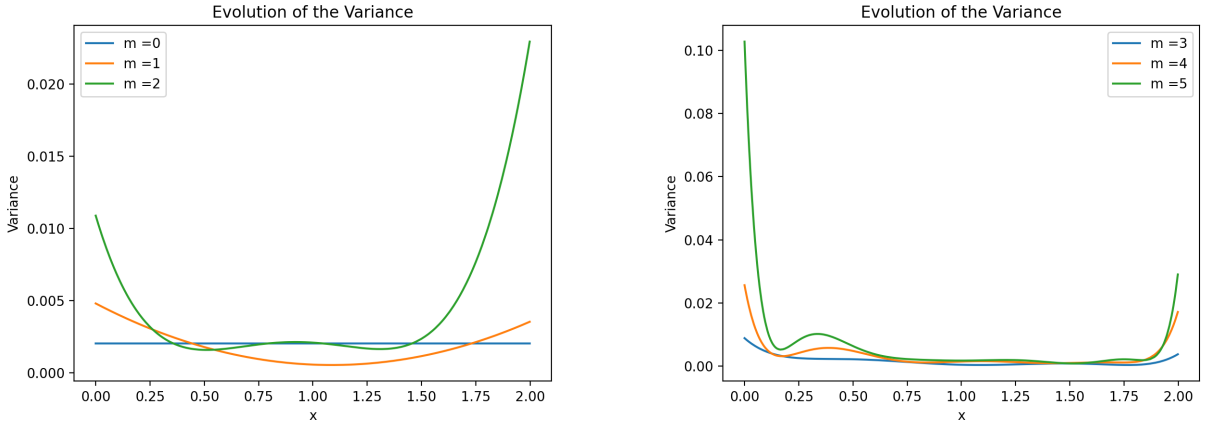

Figure 5: Evolution of the Variance w.r.t x

In order to explain our results, let's focus on m = 5. We will plot all predictions made by each learning samples.

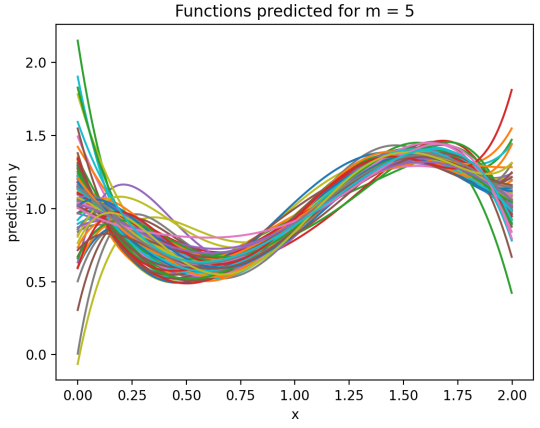

Figure 6: All LS predictions for m = 5

Compared with the variance obtained on Figure 2.4, we can easily explain the results. Indeed, for x close to 0, the predictions obtained on Figure 6 are very dispersed. It will result in a high variance for this 'point'. The predictions are quite close in the interval until they reach the end, x = 2 (e.g 1.75). Here, the dispersion is present but smaller than for x = 0. On Figure 2.4, the variance for x = 2 is thus smaller than it was in x = 0. Our results seem reliable together. Of course, all variances can be explained using the same approach.

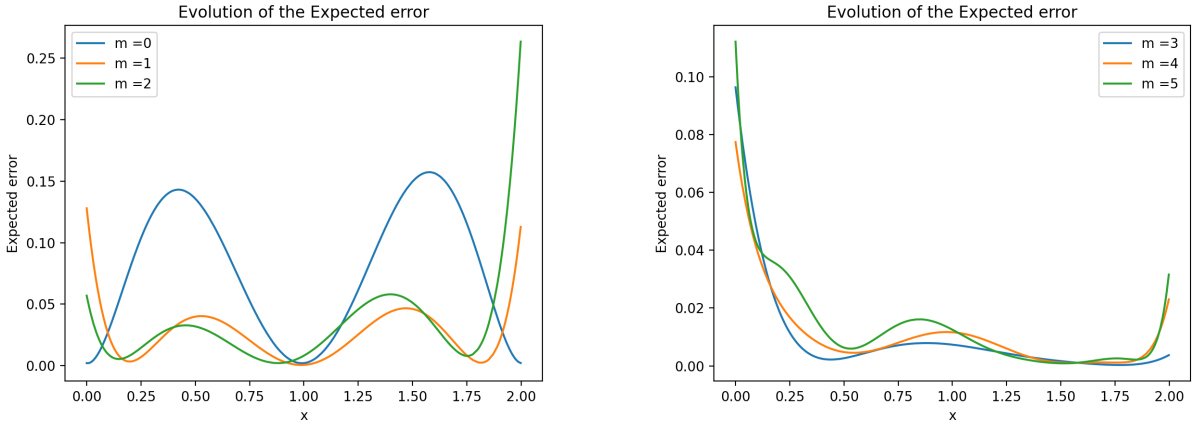

Figure 7: Evolution of the Expected error w.r.t x

The expected error is nothing else than the sum of the two firsts measures explained below.

2.5 Mean values of previous quantities

We can see through this plot that the means for each bias and variance are very close to the theoretical expectation. Since m is synonym of complexity, increasing m will result in a smaller Bias and a bigger variance. We have to find our best model that tries to make some trade-off between theses two measurements. The expected error will be the key to choose which model will be kept. Since the expected error is minimized for m = 3, we decide to choose this model to fit the real function defined in the statement. In a sense, it is not surprising that m = 3 fits is the best trade-off since the function we want to fit is a third polynomial degree.

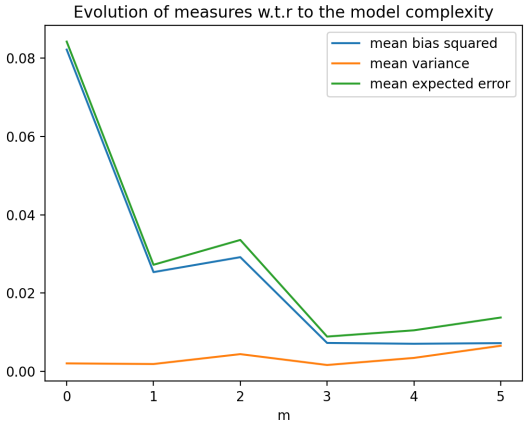

Figure 8: Caption

2.6 Ridge regression

First, we can notice that the Ridge bias is exactly the same as the OLS bias for $\lambda = 0$. It was of course expected. Second, we can see that increasing λ will not drastically change our bias results.

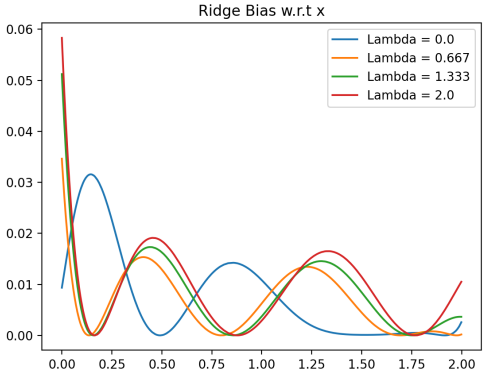

Figure 9: Caption

We can observe here that increasing λ will result in a huger variance decreasing. This is logical since we constraint our model and so limited the coefficients. We do not have to choose a big λ to see a significant impact on the variance. This was also expected by the theoretical formulation.

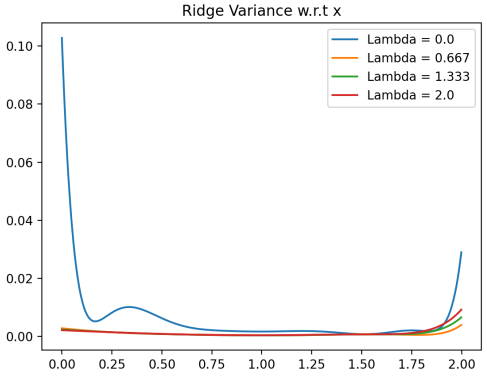

Figure 10: Caption

The expected error is much better when λ is different from 0 due to the variance decreasing. However, the impact with the Bias is difficult to analyze. We have to plot the mean for each λ is order to see which λ is the better.

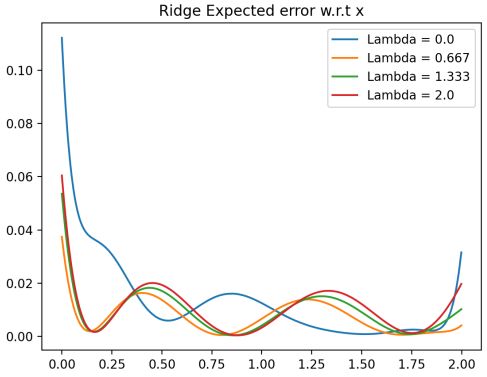

Figure 11: Caption

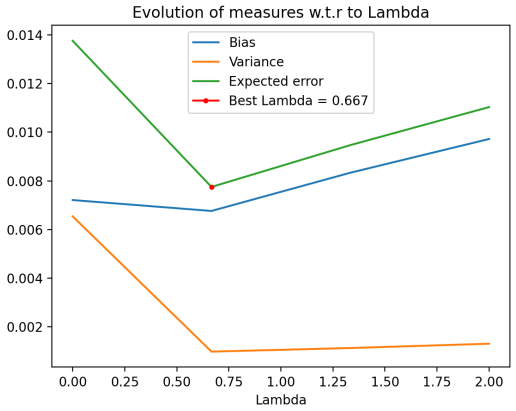

Figure 12: Caption

It can be seen here that the best trade-off for the ridge regression seems to be reached in λ = 0.666. This is of course dependant of the number of λ we have computed which is, by the way, not so big in this case.