Faculty of Applied Sciences

Introduction to Machine Learning

ELEN062-1

Project 2:
Bias and variance analysis

Authors :
Lewin Sacha - s181947
Delcour Florian - s181063
Makedonsky Aliocha - s171197

Instructors :
Geurts Pierre
Wehenkel Louis

November 2021

1 Analytical derivations

1.1 Bayes model and residual error in classification

(a) The Bayes classifier is a probabilistic model which minimizes the generalisation error, for a given $\boldsymbol{x}$. It estimates the class y of an observation with the value of the vector of parameters $\boldsymbol{x}$.

$$E_{y|\boldsymbol{x}}\left\{\mathcal{L}\left(y, h_B(\boldsymbol{x})\right)\right\}$$
$$h_B(\boldsymbol{x}) = \arg\max_c P(y = c \mid \boldsymbol{x})$$

The probabilities are obtained with the Bayes theorem :

$$p(y|\boldsymbol{x}) = \frac{p(y, \boldsymbol{x})}{p(\boldsymbol{x})}$$

With the naive assumption used by the Bayes model (each parameter is independent of the others), the conditional probabilities can be re-written as the following joint probabilities :

$$\begin{aligned} p(x_1, ..., x_n, y) &= p(x_1|y) \cdot ... \cdot p(x_n|y) \cdot p(y) \\ &= p(y) \prod_{i=1}^{n} p(x_i|y) \end{aligned} \tag{1}$$

$p(\boldsymbol{x})$ being a constant we finally have

$$\hat{y} = \operatorname*{argmax}_y p(y) \prod_{i=1}^{n} p(x_i|y)$$

Applied to our case, we have the following development.
The 2 parameters x_1 and x_2 are not correlated thanks to the covariance matrix which is diagonal. Then assuming the naive assumption, which tells that the parameters are independent, is justified. There are 2 outputs y possible, $y = 0$ for the negative class and $y = +1$ for the positive one. Since we have circular Gaussian distributions, we have the following conditional probabilities :

- y = 0 :

$$p(x_1|y = 0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1 - 1.5)^2}{2\sigma^2}}$$
$$p(x_2|y = 0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2 - 1.5)^2}{2\sigma^2}}$$

- y = 1 :

$$p(x_1|y = 1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1 + 1.5)^2}{2\sigma^2}}$$
$$p(x_2|y = 1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2 + 1.5)^2}{2\sigma^2}}$$

And we have $p(y=0) = 3 * p(y=1) = 0.75$
Finally we have the joint probabilities, with $\boldsymbol{x} = (x_1, x_2)$:

- y = 0 :

$$\begin{aligned} p(y=0|\boldsymbol{x}) &= \frac{1}{p(x)} \cdot p(y=0) \cdot p(x_1|y=0) \cdot p(x_2|y=0) \\ &= \frac{1}{p(x)} \cdot 0.75 \cdot \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1-1.5)^2+(x_2-1.5)^2}{2\sigma^2}} \end{aligned} \tag{2}$$

- y = 1 :

$$\begin{aligned} p(y=1|\boldsymbol{x}) &= \frac{1}{p(x)} \cdot p(y=1) \cdot p(x_1|y=1) \cdot p(x_2|y=1) \\ &= \frac{1}{p(x)} \cdot 0.25 \cdot \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1+1.5)^2+(x_2+1.5)^2}{2\sigma^2}} \end{aligned} \tag{3}$$

Then we need to solve an inequation to derive a relation between the two parameters x_1 and x_2 that will allow us to predict the class of the sample. To have the negative class predicted, we have the following inequation :

$$\begin{aligned} & p(y=0|\boldsymbol{x}) > p(y=1|\boldsymbol{x}) \\ \Leftrightarrow \quad & \frac{0.75}{p(x)} \cdot \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1-1.5)^2+(x_2-1.5)^2}{2\sigma^2}} > \frac{0.25}{p(x)} \cdot \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1+1.5)^2+(x_2+1.5)^2}{2\sigma^2}} \\ \Leftrightarrow \quad & 3 \cdot e^{-\frac{(x_1-1.5)^2+(x_2-1.5)^2}{2\sigma^2}} > e^{-\frac{(x_1+1.5)^2+(x_2+1.5)^2}{2\sigma^2}} \\ \Leftrightarrow \quad & \ln(3) - \frac{(x_1-1.5)^2+(x_2-1.5)^2}{2\sigma^2} > -\frac{(x_1+1.5)^2+(x_2+1.5)^2}{2\sigma^2} \\ \Leftrightarrow \quad & \ln(3) > \frac{-6x_1-6x_2}{2\sigma^2} \\ \Leftrightarrow \quad & x_1 > -\frac{\sigma^2 \ln(3)}{3} - x_2 \end{aligned} \tag{4}$$

The Bayes model can thus be written as :

$$h_b(x_1, x_2) = \begin{cases} 0 & \text{if} \quad x_1 > -\dfrac{\sigma^2 \ln(3)}{3} - x_2 \\ 1 & \text{otherwise} \end{cases}$$

(b) If negative samples become even more likely, we have a bigger argument for the $\ln()$ so x_1 can extend to smaller values to predict the negative class so it will be more often predicted, and positive class is even less likely to be predicted which is logic. If negative samples become less likely, we simple have the contrary.

(c) Here is the definition of the zero-one error loss :

$$\mathbb{1}(y, \hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases}$$

The residual error can then be computed as the sum of the probabilities that the model fails to predict correctly the class.

$$\begin{aligned} \text{Err} = &\iint_{\mathbb{R}} P(\boldsymbol{x}|y=1)P(y=1)\mathbb{1}(1, h_b(x_1, x_2))dx_1dx_2 \\ &+ \iint_{\mathbb{R}} P(\boldsymbol{x}|y=0)P(y=0)\mathbb{1}(0, h_b(x_1, x_2))dx_1dx_2 \end{aligned} \tag{5}$$

Applying the Bayes model with the zero-one error loss to each term, we obtain :

$$\begin{aligned} \text{Err} = &\int_{-\infty}^{+\infty}\int_{k-x_2}^{+\infty} P(\boldsymbol{x}|y=1)P(y=1)dx_1dx_2 \\ &+ \int_{-\infty}^{+\infty}\int_{-\infty}^{k-x_2} P(\boldsymbol{x}|y=0)P(y=0)dx_1dx_2 \end{aligned} \tag{6}$$

with $k = -\dfrac{\sigma^2 \ln(3)}{3}$ for the readability. And we know that :

$$\begin{aligned} P(\boldsymbol{x}|y=1) &= \frac{1}{2\pi\sigma^2} \cdot e^{-\dfrac{(x_1+1.5)^2 + (x_2+1.5)^2}{2\sigma^2}} \\ P(y=1) &= 0.25 \\ P(\boldsymbol{x}|y=0) &= \frac{1}{2\pi\sigma^2} \cdot e^{-\dfrac{(x_1-1.5)^2 + (x_2-1.5)^2}{2\sigma^2}} \\ P(y=0) &= 0.75 \end{aligned} \tag{7}$$

(d) Using `Matlab` to solve the integrals, we get for the first term an error of 0.0453, and for the second one an error of 0.0307, which makes a total residual error of 0.0759. Now using `python` to verify empirically the results, we obtain a residual error close to the result that we obtained with the analytical function, as you can see in the Table 1. If we increase N, we get a residual error closer to the one obtained with the analytical function (approximately, there is still a part of randomness).

The error is small, which means that our Bayes model estimates correctly the classes of the samples.

	$N = 10^3$	$N = 10^4$	$N = 10^5$
Residual error	0.072	0.0756	0.0759

Table 1: Residual error for different sampling size

1.2 Bias and variance in regression

(a) The Bayes model tries to minimize the expectation of the squared error between the true y and the predicted $\hat{y}$. Since we want to minimize it, we will derive it with regard to the estimated $\hat{y}$ and cancel this derivative :

$$\begin{aligned} & \frac{\partial}{\partial \hat{y}} E_y\{(y-\hat{y})^2\} = 0 \\ \Leftrightarrow \quad & \hat{y} = E_y\{y\} \end{aligned} \tag{8}$$

We know that $y = ax + \epsilon$, so we have :

$$\begin{aligned} E_y\{y\} &= E_y\{ax+\epsilon\} \\ &= E_x\{ax\} + E_\epsilon\{\epsilon\} \\ &= aE_x\{x\} + 0 \\ &= a \cdot 0.5 \end{aligned} \tag{9}$$

since x is drawn uniformly on [0,1], and the mean of the normal distribution followed by ϵ is 0.
The bias/variance decomposition gives us the residual error as the variance of y ($\text{var}_y\{y\}$). Thanks to the following properties :

- $\text{var}\{aX\} = a^2\text{var}\{X\}$
- $\text{var}\{X+Y\} = \text{var}\{X\} + \text{var}\{Y\} + 2\text{cov}(X,Y)$
- $\text{var}\{X\} = \frac{1}{12}$ if X is an uniformly distributed random variable in [0;1]

Which lead to the below development of the variance of y :

$$\begin{aligned} E_y\left\{\left(y - E_y\{y\}\right)^2\right\} = \text{var}_y\{y\} &= \text{var}_y\{ax+\epsilon\} \\ &= \text{var}\{ax\} + \text{var}\{\epsilon\} + 2\text{cov}(ax,\epsilon) \\ &= \text{var}\{ax\} + \text{var}\{\epsilon\} \quad \text{x and } \epsilon \text{ are independent} \\ &= a^2\text{var}\{x\} + \text{var}\{\epsilon\} \\ &= \frac{a^2}{12} + \sigma^2 \end{aligned} \tag{10}$$

which is the expression of the residual error.

(b) The mean squared bias is expressed as :

$$\text{bias}^2 = (E_y\{y\} - E_{LS}\{\hat{y}\})^2$$

where $E_y\{y\} = a \cdot 0.5$ and $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ for one learning sample. Thus we have

$$E_{LS}\{\hat{y}\} = E_y\{y\} = a \cdot 0.5$$

so the mean squared bias is 0.
We already know $\text{var}_y\{y\}$, so we can express the variance of the learning sample as :

$$\begin{aligned} \text{var}_{LS}\{\hat{y}\} &= \frac{1}{N}\text{var}_y\{y\} \\ &= \frac{1}{N}(\frac{a^2}{12} + \sigma^2) \end{aligned} \tag{11}$$

(c) As we can see with the expressions of the bias and the variance, a and σ have no impact on the bias, but the variance directly depends on the square of those variables, so bigger they are, bigger the variance is. It shows that bigger the variance of the noise, bigger the variance over the learning samples, which is logic. But a^2 is divided by 12 so its influence is smaller than the one of σ^2.

2 Empirical analysis

(a) Let us consider our generator that can infinitely generate (random) y values given a value of x, and a value x_0 for which we want to estimate the residual error, squared bias, and variance.
First, we can estimate the residual error by pushing x_0 through our black box generator N_{test} (which can be arbitrarily big) times, and then we take as estimate the **empirical** variance over these values of y.

$$\text{var}_y\{y\} \approx \frac{1}{N_{\text{test}}} \sum_{i=0}^{N_{\text{test}}} (y_i - \bar{y})^2$$

Then, for the squared bias and variance, we generate N_{test} values of $\hat{y}$ at $x = x_0$. For that purpose, we repeat N_{test} the following procedure:

- Take N_{learn} random pairs (x, y) to get a learning set.
- Fit the model on the generated set.
- Predict $\hat{y}$ by pushing x_0 through the prediction model obtained.

Let's note that to generate each learning set, we take N_{learn} values of x and compute the function + noise at each value to get y. This procedure gives a list of N_{test} values of $\hat{y}$ at $x = x_0$.
We can estimate the squared bias by taking the square of the empirical mean of y values minus the empirical mean of $\hat{y}$ values.

$$\text{bias}^2 = \left(\frac{1}{N_{\text{test}}} \sum_{i=0}^{N_{\text{test}}} y_i - \frac{1}{N_{\text{test}}} \sum_{i=0}^{N_{\text{test}}} \hat{y}_i \right)^2$$

Finally, variance can be estimated by taking the empirical variance of the $\hat{y}$ values.

$$\text{var}_{LS}\{\hat{y}\} \approx \frac{1}{N_{\text{test}}} \sum_{i=0}^{N_{\text{test}}} (\hat{y}_i - \bar{\hat{y}})^2$$

(b) To compute mean values over the possible values of x, we can simply take N_{mean} values of x and for each of them do the protocol mentioned above. Then, we can take the empirical mean of the values.

(c) They are obviously not appropriate anymore, considering the method described above.

For residual error, it is no longer possible to have a large set of values of y at a same value of x.

For bias and variance, it is not possible anymore to generate the many data sets to fit then predict at x_0.

For all values, the main issue is thus not to be able to focus on one x and have as many values of y and $\hat{y}$ as we need at that particular x.

(d) Applying the protocol described before with $N_{\text{learn}} = 100$ and $N_{\text{test}} = 500$ gives the following plots. The linear model used is simply linear regression, and the non-linear one is decision trees;

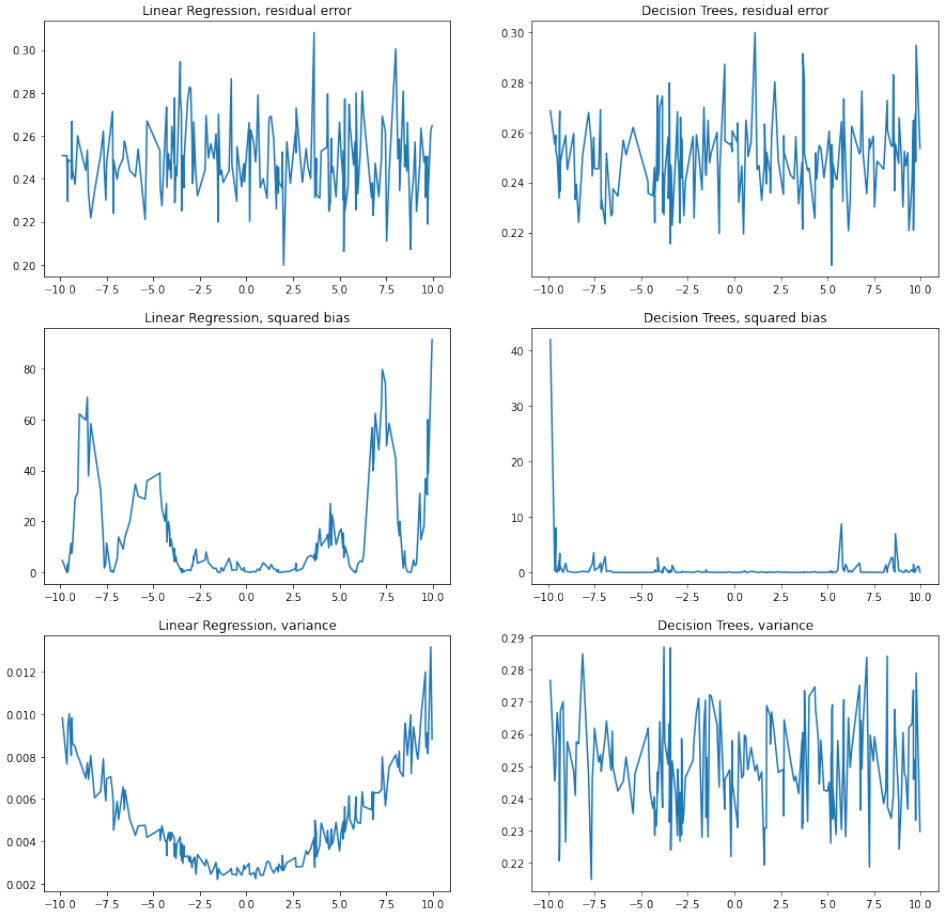

Figure 1: Performance of both models

As can be seen on the plots, the linear model has very high bias, and very low variance, whereas the bias of the non-linear model is a lot lower but has higher variance. This is a great example of the bias-variance trade-off. The linear model

indeed underfits a lot, as can be seen below, and this is caused by the low variance and high bias. Indeed, the high bias is expected because the real curve is not a line at all, and therefore at the extrema, it is very far from the prediction.
The non linear model overfits, this is due to high bias and low variance.
Below our shown predictions for the two selected models. The decision trees model was fit with a max depth of 6 for showcase. As expected by looking at expected error plots, we can see that the non linear model obviously approximates the curve way better almost everywhere.

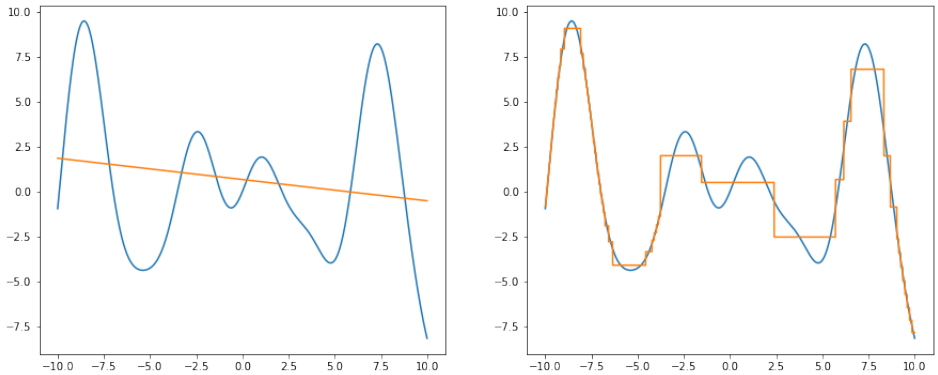

Figure 2: Predictions for each model

(e) For the linear model, here are the trends for each error term as a function of the learning set size.

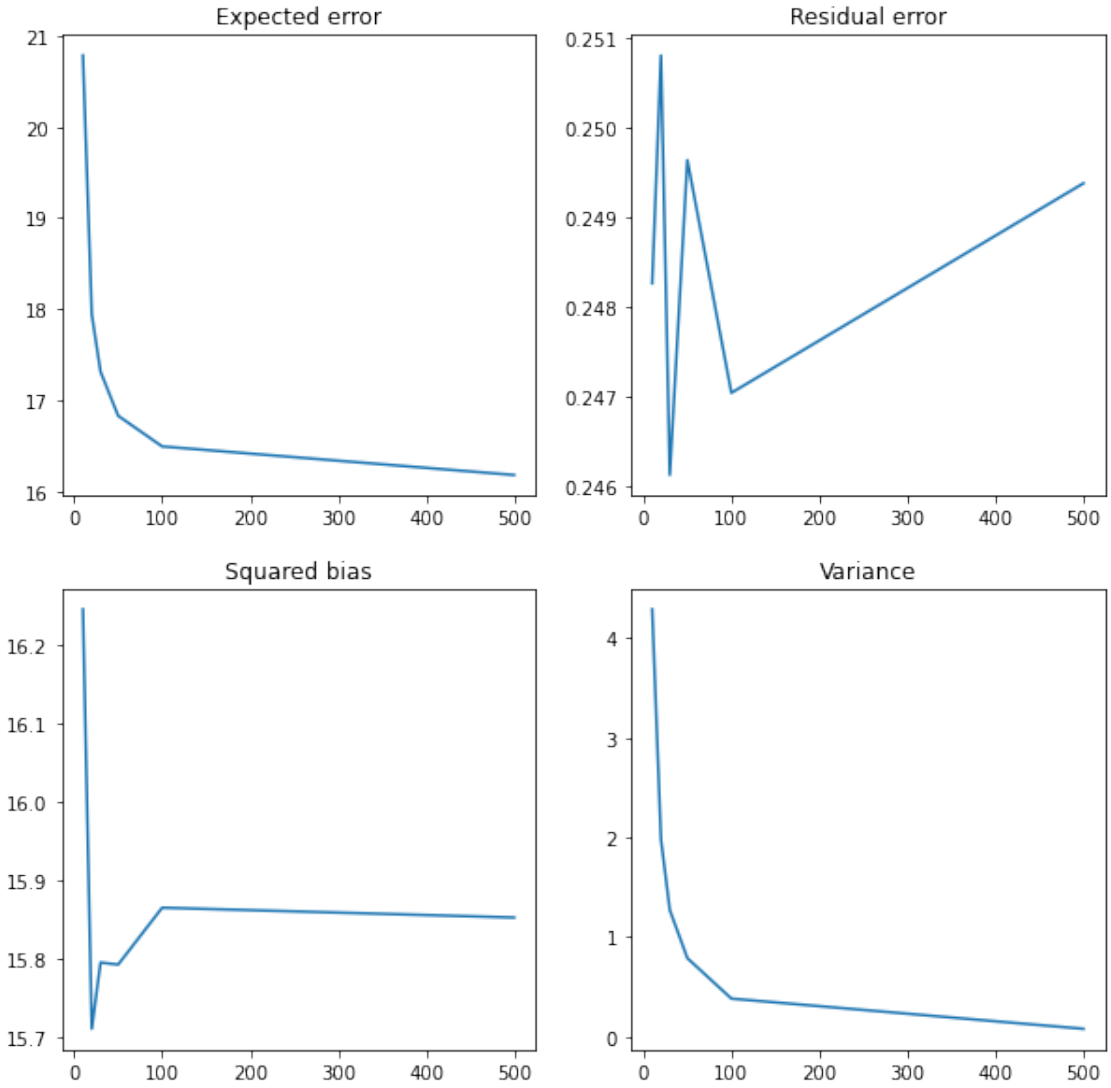

Figure 3: Linear model errors as a function of the LS size

Below is the same but for decision trees.

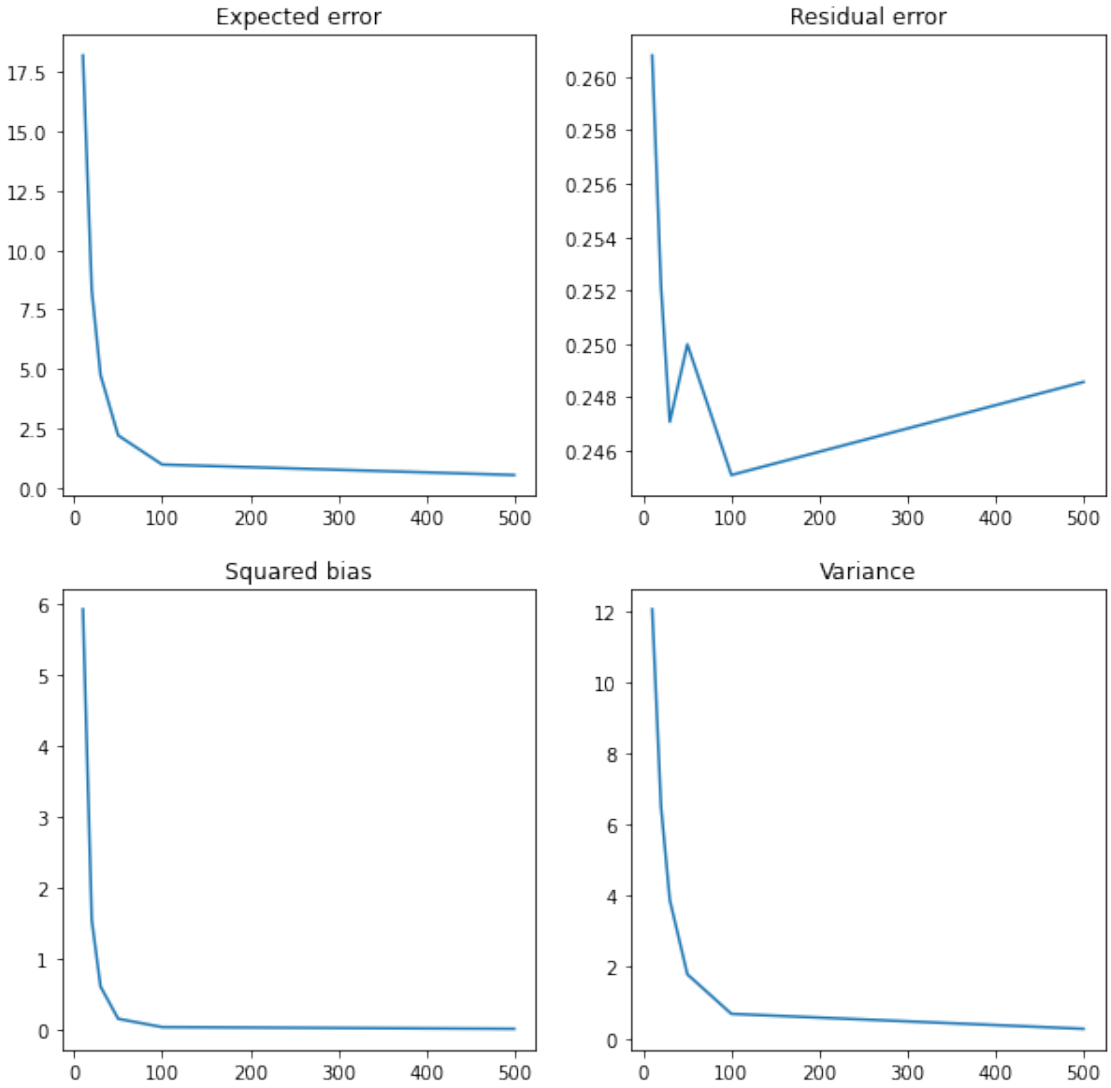

Figure 4: Decision trees errors as a function of the LS size

Now, here are the trends for each error as a function of the complexity of the decision trees, so by tuning the maximum depth parameter.

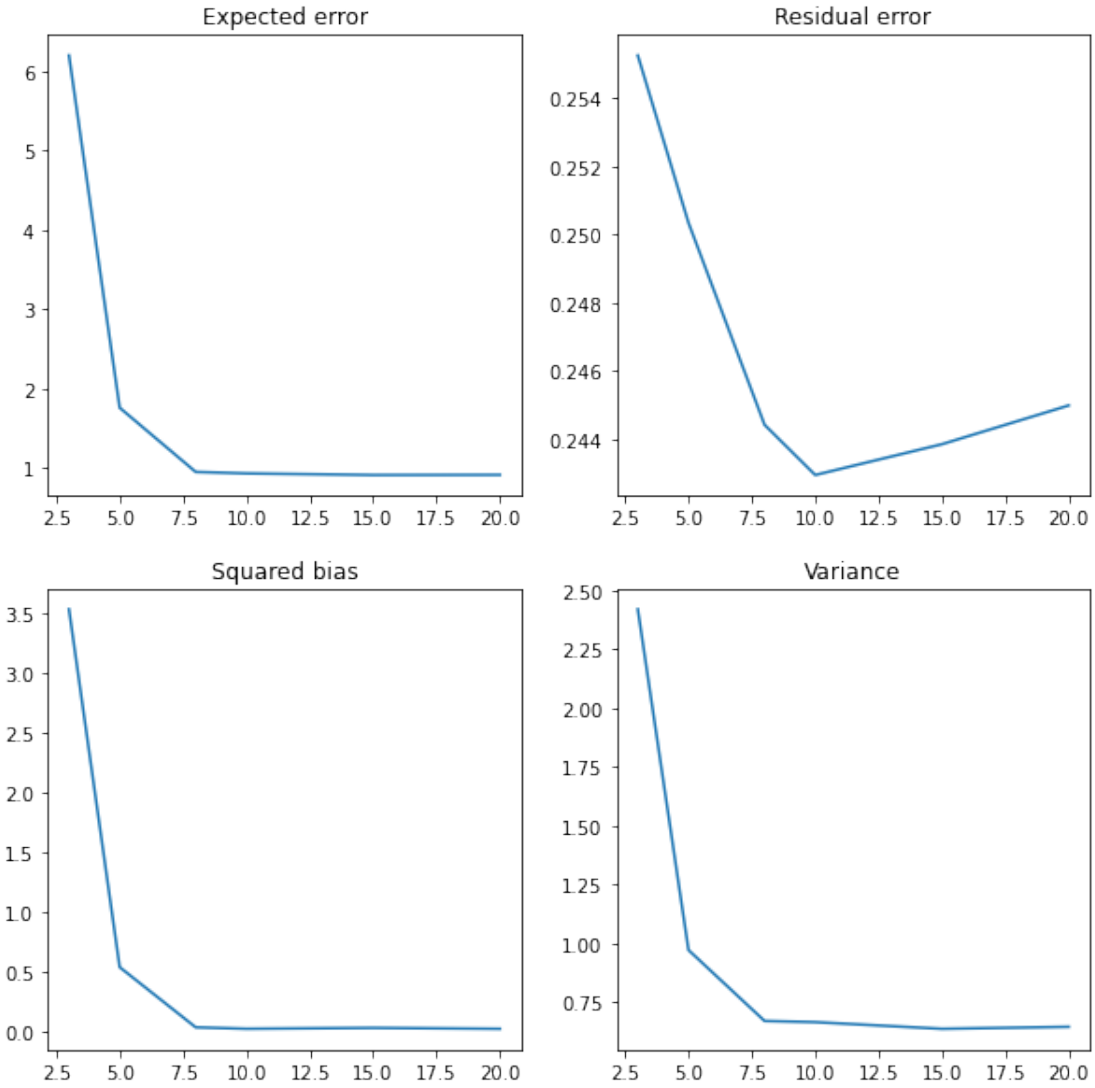

Figure 5: Decision trees model errors as a function of the max depth