Project 2 - Bias and variance analysis

Valentin Absalon S214256-Maziane Yassine S184384

September 13, 2023

1 Analytical derivation

1.1 Bayes model and residual error in classification

1.1.1 Bayes Model Formulation

In this section, we wish to find an analytical formulation for the Bayes model $h_b(x_1, x_2)$, that is, a model that will classify a point (x_1, x_2) as belonging to the positive or negative class upon some condition that we will derive. We expect that condition to depend both upon x_1 and x_2. The Bayes model $h_b(x_1, x_2)$ is defined as

$$h_b(x_1, x_2) = \underset{y \in Y}{\operatorname{argmax}} P(y|x_1, x_2) \tag{1}$$

that can be developed using Bayes Theorem

$$h_b(x_1, x_2) = \underset{y \in Y}{\operatorname{argmax}} \frac{P(y) * P(x_1, x_2|y)}{P(x_1, x_2)} \tag{2}$$

and as $P(x_1, x_2)$ is independent of y

$$h_b(x_1, x_2) = \underset{y \in Y}{\operatorname{argmax}} P(y) * P(x_1, x_2|y) \tag{3}$$

where Y=-1,+1, where $y = -1$ denotes the negative class and $y = +1$ denotes the positive class. We thus obtain

$$h_b(x_1, x_2) = \begin{cases} +1 & \text{if } P(y=+1) * P(x_1, x_2|y=+1) > P(y=-1) * P(x_1, x_2|y=-1) \\ -1 & otherwise \end{cases} \tag{4}$$

Let's see under which conditions on x_1 and x_2 the latter condition will be respected.
Since P(y)'s are respectively known as $P(y = +1) = \frac{1}{4}$ and $P(y = -1) = \frac{3}{4}$, we must develop $P(x_1, x_2|y)$ which follows a multi dimensional normal distribution. We know its probability density function is given by Where Σ represents the variance-covariance matrix, μ the mean vector and N, the number of dimensions which is equal to 2 in our case.
We must now distinguish both cases corresponding to y=+1 and y=-1.
Regarding **y=+1** whose distribution is centered at $(-1.5, -1.5)$ we get

$$f(x_1, x_2|y = +1) = \frac{1}{2\pi\sigma^2} exp(\frac{-1}{2}((\frac{x_1 + 1.5}{\sigma})^2 + (\frac{x_2 + 1.5}{\sigma})^2)) \tag{5}$$

Regarding **y=-1** whose distribution is centered at $(1.5, 1.5)$ we get

$$f(x_1, x_2|y = +1) = \frac{1}{2\pi\sigma^2} exp(\frac{-1}{2}((\frac{x_1 - 1.5}{\sigma})^2 + (\frac{x_2 + -.5}{\sigma})^2)) \tag{6}$$

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right]$$

Figure 1: multi dimensional normal distribution

We now have to solve for x_1 and x_2

$$P(y=+1)*P(x_1,x_2|y=+1) > P(y=-1)*P(x_1,x_2|y=-1) \tag{7}$$

Which is developed thanks to 6 and 5.

$$\frac{1}{4}*\frac{1}{2\pi\sigma^2}*e^{\frac{-1}{2}\left(\left(\frac{x_1+1.5}{\sigma}\right)^2+\left(\frac{x_2+1.5}{\sigma}\right)^2\right)} > \frac{3}{4}*\frac{1}{2\pi\sigma^2}*e^{\frac{-1}{2}\left(\left(\frac{x_1-1.5}{\sigma}\right)^2+\left(\frac{x_2-1.5}{\sigma}\right)^2\right)} \tag{8}$$

Simplifying and taking the natural logarithm we get

$$\frac{-1}{2}\left(\left(\frac{x_1+1.5}{\sigma}\right)^2+\left(\frac{x_2+1.5}{\sigma}\right)^2\right) > ln(3)+\frac{-1}{2}\left(\left(\frac{x_1-1.5}{\sigma}\right)^2+\left(\frac{x_2-1.5}{\sigma}\right)^2\right) \tag{9}$$

Simplifying again we get

$$\left(\left(\frac{x_1+1.5}{\sigma}\right)^2+\left(\frac{x_2+1.5}{\sigma}\right)^2\right) < -2*ln(3)+\left(\left(\frac{x_1-1.5}{\sigma}\right)^2+\left(\frac{x_2-1.5}{\sigma}\right)^2\right) \tag{10}$$

Or

$$(x_1+1.5)^2+(x_2+1.5)^2 < -2\sigma^2 ln(3)+(x_1-1.5)^2+(x_2-1.5)^2 \tag{11}$$

From there we may get

$$6x_1+6x_2 < -2\sigma^2 ln(3) \tag{12}$$

And our condition finally is

$$x_1+x_2 < \frac{-\sigma^2 ln(3)}{3} \tag{13}$$

Therefore, our Bayes model writes

$$h_b(x_1,x_2) = \begin{cases} +1 & \text{if } x_1+x_2 < \frac{-\sigma^2 ln(3)}{3} \\ -1 & otherwise \end{cases} \tag{14}$$

And the corresponding zero-one error loss function writes

$$L_{0/1}(x_1,x_2,y) = \begin{cases} 0 & \text{if } h_b(x_1,x_2)=y \\ +1 & otherwise \end{cases} \tag{15}$$

1.1.2 Bayes dependence on the ratio

We will now suppose that the negative examples are m times more likely than positive ones in order to be as general as possible. As the ratio does not affect the conditional probabilities, the only change in the Bayes model will be noticed in $P(y=+1)$ and $P(y=-1)$ in equation 4. We observe in equation 7 that the ratio only appears in the natural logarithm. We may show that for a ratio m to 1, the Bayes model may be expressed as

$$h_b(x_1,x_2) = \begin{cases} +1 & \text{if } x_1+x_2 < \frac{-2\sigma^2 ln(m)}{3} \\ -1 & otherwise \end{cases} \tag{16}$$

As the ratio m increases, the right hand side, i.e the threshold, decreases and the probability that a point (x_1,x_2) is smaller than the threshold decreases as well. Therefore as the ratio m increases, the Bayes model will classify less and less points as belonging to the positive class and this is what we want. Indeed if m increases, there are more and more negative samples and less positive samples so it is totally logic that we classify less points as belonging the the positive class.
It might be useful to look at the Bayes condition for different values of the ratio m. We observe on figure 2 that the plane is divided in 2 equal parts, furthermore we may notice that the decision boundary is the mediator of the two distributions' centers. We realize on figures 3 and 4 that as m increases, the decision boundary moves to the south-west but it remains parallel to the latter mediator.

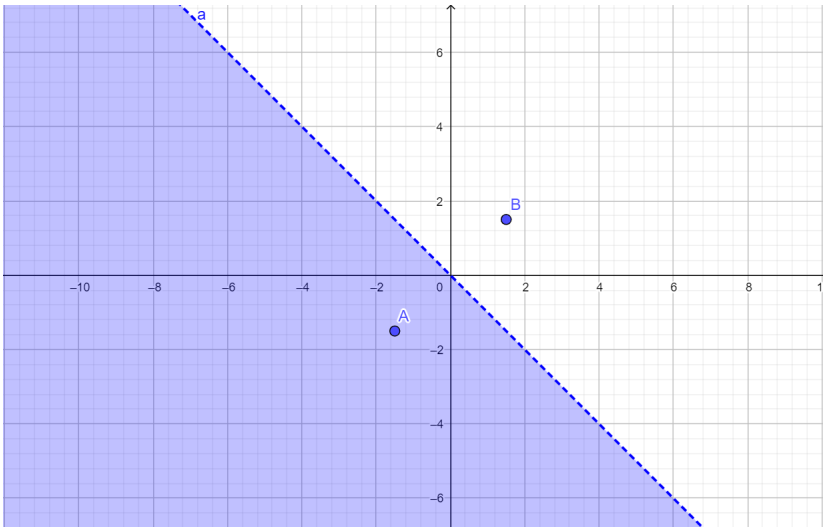

Figure 2: Bayes Model for a ratio of 1

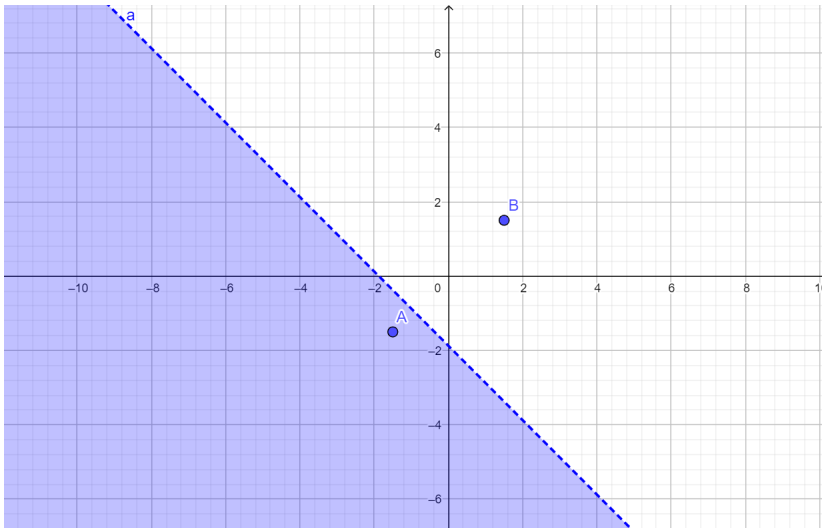

Figure 3: Bayes Model for a ratio of 3

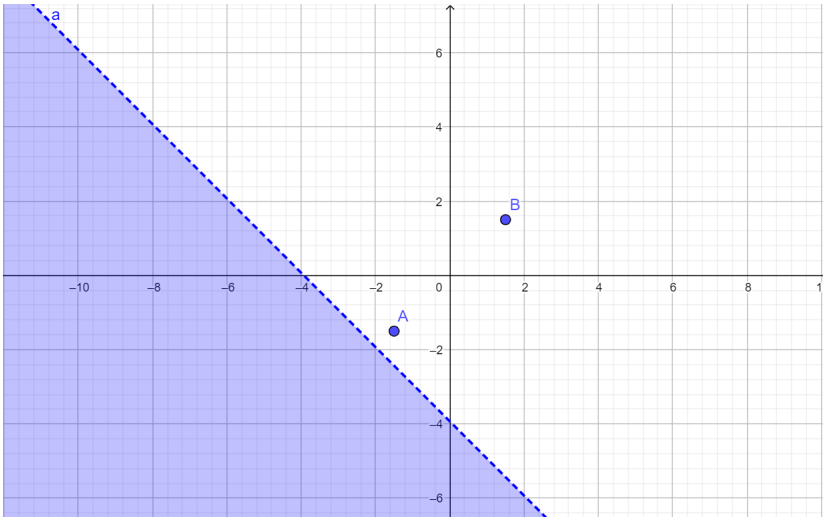

Figure 4: Bayes Model for a ratio of 10

1.1.3 Residual error

In this section, we will try to formulate analytically the residual error i.e. the generalization error $E_{x_1,x_2,y}\{1(y \neq h_b(x_1,x_2)\}$.
(Ask for the probability) In our context, there are 2 possible ways of miss classifying:

1. The Bayes model predicts $h_b(x_1,x_2) = +1$ while $y = -1$
2. The Bayes model predicts $h_b(x_1,x_2) = -1$ while $y = +1$

We have,

$$E_{x_1,x_2,y}1(y \neq h_b(x_1,x_2) = P(y=+1, h_b(x_1,x_2)=-1) + P(y=-1, h_b(x_1,x_2)=+1) \tag{17}$$

We then use the definition of conditional probabilities

$$P(y,h_b) = P(y)P(h_b|y) \tag{18}$$

Equation 17 becomes

$$E_{x_1,x_2,y}1(y \neq h_b(x_1,x_2) = P(y=+1)P(h_b(x_1,x_2)=-1|y=+1)+P(y=-1)P(h_b(x_1,x_2)=+1|y=-1) \tag{19}$$

Since P(y=+1)=$\frac{1}{4}$ and P(y=-1)=$\frac{3}{4}$ we get

$$E_{x_1,x_2,y}1(y \neq h_b(x_1,x_2) = \frac{1}{4}P(h_b(x_1,x_2)=-1|y=+1) + \frac{3}{4}P(h_b(x_1,x_2)=+1|y=-1) \tag{20}$$

Which we will write

$$E_{x_1,x_2,y}1(y \neq h_b(x_1,x_2) = \frac{1}{4}P_A + \frac{3}{4}P_B \tag{21}$$

for readability purposes. We may compute P_A as

$$P_A = P(h_b(x_1,x_2)=-1|y=+1) \tag{22}$$

$$P_A = P(x_1+x_2 > \frac{-2\sigma^2 ln(3)}{3}|y=+1) \tag{23}$$

$$P_A = P(x_2 > \frac{-2\sigma^2 ln(3)}{3} - x_1|y=+1) \tag{24}$$

Since we know the distribution of x_1 and x_2 given y, we may compute the latter probability by means of integration. We then get

$$P_A = \int_{-\infty}^{+\infty} \int_{\frac{-2\sigma^2 ln(3)}{3}-x_1}^{+\infty} f(x_1,x_2|y=+1)dx_2dx_1 \tag{25}$$

By plugging 5 we get

$$P_A = \int_{-\infty}^{+\infty} \int_{\frac{-2\sigma^2 ln(3)}{3}-x_1}^{+\infty} \frac{1}{2\pi\sigma^2} exp(\frac{-1}{2}((\frac{x_1+1.5}{\sigma})^2 + (\frac{x_2+1.5}{\sigma})^2))dx_2dx_1 \tag{26}$$

We may compute P_B in a similar way except that we want

$$P_B = P(x_1+x_2 < \frac{-2\sigma^2 ln(3)}{3}|y=-1) \tag{27}$$

By plugging 6 we get

$$P_B = \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-2\sigma^2 ln(3)}{3}-x_1} \frac{1}{2\pi\sigma^2} exp(\frac{-1}{2}((\frac{x_1-1.5}{\sigma})^2 + (\frac{x_2-1.5}{\sigma})^2))dx_2dx_1 \tag{28}$$

We have now a developed analytical formulation for the residual error.

1.1.4 Analytical and empirical estimation of the error

Using $\sigma = 1.6$, we may compute P_A and P_B however they are defined by integrals whose solutions can not be expressed by closed-form solutions. We therefore decided to compute these integrals numerically thanks to sciypy integrate.
We obtain a generalization error of 7.59 %.
We now want to compute the generalization error empirically. To do so, we generate a data set and study N samples from it. For each sample, we compare the outcome of our Bayes model to the true class of the considered sample. This way we will count the number of failures. We will repeat this process 500 times for each value of N considered. The values of N we studied and their corresponding error rate may be observed on table

N	25	100	200	500	1000	2000	3000
Error in %	7.14	7.52	7.43	7.58	7.60	7.56	7.56

For N=3000, the error is of 7.57 % while the "true", omitting numerical integration approximation, error is of 7.59 %. This is a very good approximation as it is only far from the true value by 0.26 % in relative error.

1.2 Bias and variance in regression

1.2.1 Bayes model and its residual error

We know that the Bayes model $h_b(x)$ may be written as

$$h_b(x) = E_{Y|X}[y] \tag{29}$$

Where Y and X are both random variables.We may realize that $Y|X$ is actually a normal distribution. Indeed as X takes the fixed value x, the distribution of $Y|x$ writes $P_{Y|X}(y|x) \sim N(a * x, \sigma^2)$. We may therefore write the Bayes as :

$$h_b(x) = a * x \tag{30}$$

We then compute the residual error as

$$RE = E_X[(y - h_b(x)^2] = E_X[(ax + \epsilon - ax)^2] = E_X[\epsilon^2] \tag{31}$$

Since

$$E_X[\epsilon^2] = V_X[\epsilon] + E_X[\epsilon]^2 = V_X[\epsilon] \tag{32}$$

We find that

$$RE = \sigma^2 \tag{33}$$

An alternative method would have been to deduce

$$RE = \sigma^2 \tag{34}$$

simply based on the distribution of $Y|X$.

1.2.2 Mean squared bias and mean variance

The mean squared bias MSB may be computed as

$$MSB = E_X[bias(x)^2] = E_X[(h_b(x) - E_{LS}[\mu])^2] \tag{35}$$

Developping the square and given expectation linearity

$$MSB = E_X[a^2x^2] + E_X[E_{LS}[\mu]^2] + E_X[-2ax * [E_{LS}[\mu]] \tag{36}$$

We may compute each term individually for readability purposes, the first term is

$$E_X[a^2x^2] = \int_0^1 a^2x^2 = a^2/3 \tag{37}$$

The second term is $E_X[E_{LS}[\mu]^2]$. Since the optimal μ for a data set is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{38}$$

The proof may be sent by mail if asked, maybe put it in the annex.
It yields,

$$E_{LS}[\mu] = \frac{1}{N}\sum_{i=1}^{N} E_{LS}[y_i] \tag{39}$$

But we know that as N grows, the expectation of y_i of the learning set tends to the expectation of y_i of the whole population. Hence, we write

$$E_{LS}[\mu] = \frac{1}{N}\sum_{i=1}^{N} E_y[y] \tag{40}$$

Since $E_y[y] = \frac{a}{2}$, we have

$$E_{LS}[\mu] = \frac{a}{2} \tag{41}$$

So that the whole second term writes

$$E_X[E_{LS}[\mu]^2] = \frac{a^2}{4} \tag{42}$$

The final and third term may be computed as follows:

$$E_X[-2ax * [E_{LS}[\mu]] = E_X[-2ax * \frac{a}{2}] = -a^2 E_X[x] \tag{43}$$

Since $E_X[x] = \frac{1}{2}$, the third term writes

$$E_X[-2ax * [E_{LS}[\mu]] = \frac{-a^2}{2} \tag{44}$$

All in all, the MSB is

$$MSB = \frac{a^2}{3} + \frac{a^2}{4} - \frac{a^2}{2} = \frac{a^2}{12} \tag{45}$$

Which is our final result for the mean squared bias.
Similar developments may now be performed regarding the mean variance MV.

$$MV = E_X[E_{LS}[(\mu - E_{LS}[\mu])^2] \tag{46}$$

Which is nothing but

$$MV = E_X[V_{LS}[\mu]] \tag{47}$$

Which may be re expressed thanks to 38 as:

$$MV = E_X[V_{LS}[\frac{1}{N}\sum_{i=1}^{N} y_i]] \tag{48}$$

Since V(ax)=$a^2V(x)$ and $V(a+b) = V(a) + V(b)$ if a and b are independant which is the case for the $y_i's$

$$MV = E_X[\frac{1}{N^2}\sum_{i=1}^{N} V_{LS}[y_i]] \tag{49}$$

We know assume that the variance of the y's over the learning set converges to the variance of y's over the whole population as N grows. This may be exprimed as

$$V_{LS}[y_i] = V_Y[y_i] \tag{50}$$

Hence, we get

$$MV = E_X[\frac{1}{N^2}\sum_{i=1}^{N} V_y[y_i]] \tag{51}$$

Where

$$V_Y[y_i] = a^2 * V_Y[x] + V_y[\epsilon] = \frac{a^2}{12} + \sigma^2 \tag{52}$$

Injecting the latter result in the latest form of MV yields,

$$MV = E_X[\frac{1}{N^2}(N(\frac{a^2}{12} + \sigma^2))] \tag{53}$$

So that our variance is

$$MV = \frac{1}{N}(\frac{a^2}{12} + \sigma^2) \tag{54}$$

Which is our final result.

1.2.3 Impact of a and σ on bias and variance

Let us first study the bias. We observe that the bias is a quadratic function of the slope a. We could have expected the result to be highly dependant on the slope. Indeed, as the slope increases, the difference between the bayes model $h_b(x) = ax$ and $f_{LS} = \mu = \frac{a}{2}$ get larger and larger especially at points far from $x = \frac{1}{2}$. As the slope increases it is harder and harder for the model output by the learning algorithm to be close to the Bayes mode. To extend this thought, as N gets larger and larger, the bayes model and f_{LS} tend to get more and more orthogonal to each other.
Since one little picture says more than a long speech, we may look at the figures 6 and 5

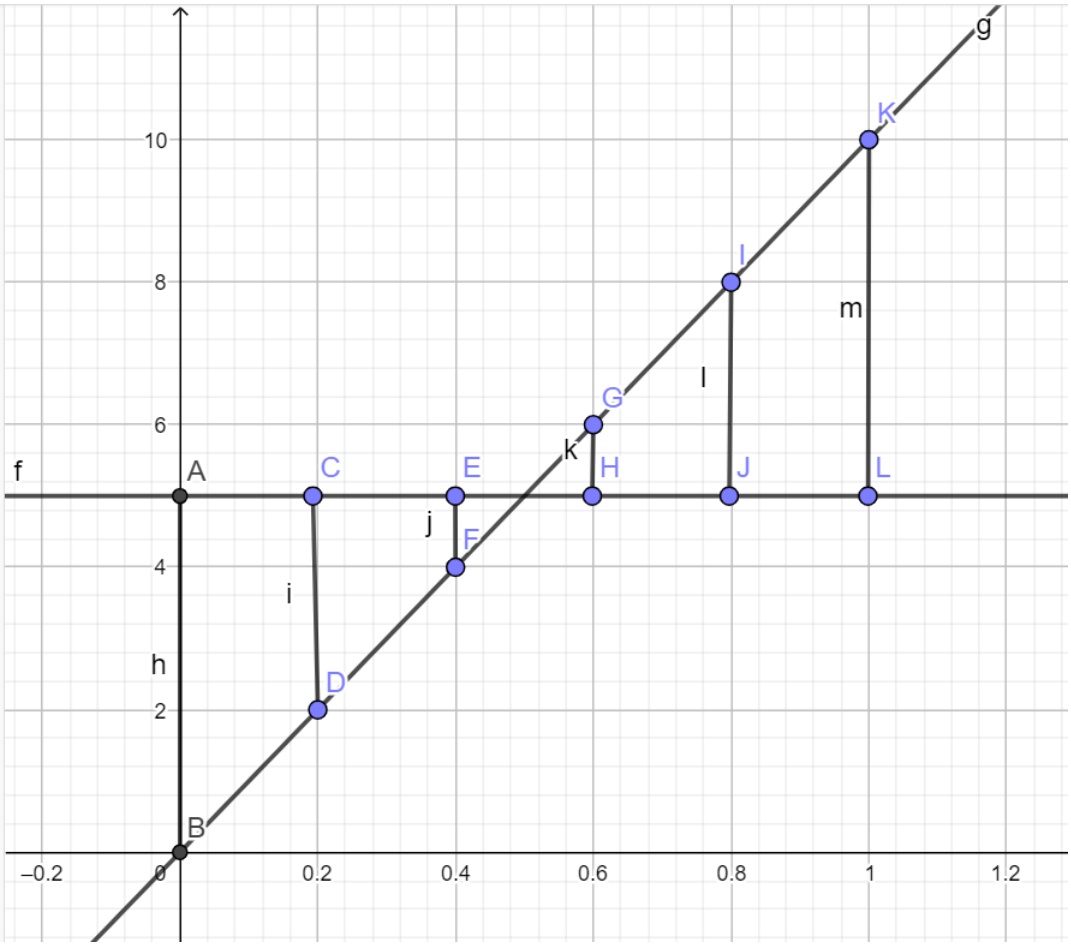

Figure 5: Bias with a large slope

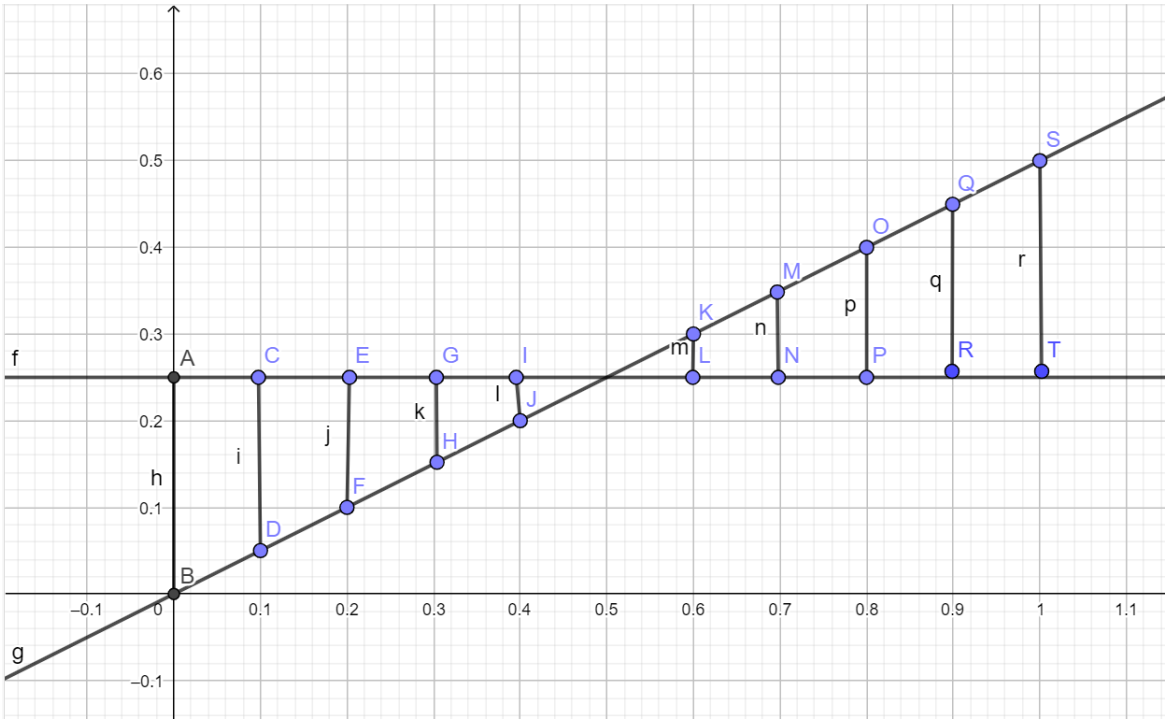

Figure 6: Bias with a small slope

We observe that for most x's, the difference between the bayes model and the constant model output by the learning algorithm is far larger for the large slope. Notice the y axis scale has changed! We should also notice that the mean squared bias is null for a null slope.
Finally, we notice the bias does not depend on the learning set size N which is expected, usually the bias depends on the complexity which depends itself on N. In this case, the model is extremely simple and its complexity does not depend on N so that the bias does not depend on N either.
Regarding the mean variance, we first observe that the noise it depends on all three parameters N, a and σ. At first we notice that its decreasing with respect to N. Indeed, as N increases, μ tends to converge towards the true mean of the population.
We may also notice that the mean variance is quadratic with the slope a. We may "justify" this dependence in the following way : Let LS_1 and LS_2 be two learning sets of same size, surely there are little chances that the same x_i is present in both LS. Let (x_A, y_A) be a pair in LS_1 and consider $(x_B = x_A + \delta, y_B)$ a pair in LS_2 with delta a small quantity. If a is small, then y_A will be close to y_b however if a is large it will not be the case. Alternatively, from the true distribution, we know that $y_A = a * x_A + \epsilon_A$ and $y_B = a * x_B + \epsilon_B$. For simplicity reasons, we will consider that $\epsilon_A = \epsilon_B$ which is probably not the case. Thus we have $y_A - y_B = -a * \delta$ which is proportional to the slope a. Therefore the mean y_1 of LS_1 and the mean y_2 of LS_2 will differ more and more as "a" grows. Hence, for larger "a", different data sets will output models (μ_1, μ_2) which are further and further away from one another, thus leading to a higher variance.
Finally we observe that the mean variance is quadratic with the noise, this is a well know result in machine learning. As the model learnt by the algorithm depends on the y values of the learning set with the y values being dependant on the noise, we may realize that for a different data set the y values will be very different hence leading to a very different μ. Considering the noise on the y values to be very large allows to be more easily convinced.

2 Empirical analysis

2.1 Protocol

We are given a learning sample LS = (x1, y1), . . . ,(xN , yN) of N pairs to train a model and $\hat{y}(x)$ the function learned from LS. We know that for a given point xo and for the given supervised learning

algorithm, we have :

The squared bias formula :

$$(E_{y|x}[y] - E_{LS}[\hat{y}])^2 \tag{55}$$

The variance formula :

$$V_{LS}[\hat{y}(x)] = (E_{LS}[y] - E_{LS}[\hat{y}])^2 \tag{56}$$

The noise formula :

$$E_{y|x}[(y - E_{y|x}[y])^2] \tag{57}$$

So to obtain this differents formulas, we have to do this following steps :

1. First create a large data set LS of size N .
2. Select each pair $(x_i, y_i) \in LS$ such that $x_i = x_0$ to create the set LS'.
3. Compute the mean and variance of LS' with respect to y's as x's belonging to LS' are identical. The mean corresponds to $E_{y|x_0}(y)$ and the variance corresponds to $V_{y|x_0}(y)$ the noise .
4. Split LS into k subsets that will be denoted as LS_i, i=1,2,3,...,k.
5. Let $\hat{y}_i$ be the model output by the supervised algorithm applied on LS_i.By taking the mean on all k subsets of $\hat{y}_i(x_0)$ we get $E_{LS}\hat{y}(x_0)$, computing the variance in the following way yields $V_{LS}\hat{y}(x_0)$.
6. The squared bias may be computed by squaring the difference of the means : $bias^2 = (E_{LS}\hat{y}(x_0) - E_{y|x_0}(y))^2$

2.2 Protocol to estimate the mean values of the residual error, the squared bias and the variance of the learning algorithm

For all $x_i \in LS$, we compute the 3 quantities. We now have N residual errors, N squared biases and N variances corresponding to different x_i. We then simply take the mean of each quantity to get the mean values desired. Since the random variable X is uniformly distributed, the mean of the quantities for all x is equivalent to $E_X[]$ of the quantities.

2.3 Finite learning set

The protocols may or may not be appropriate anymore. On one hand, as stated in question a), our protocols work with finite learning set of size N even if N may be arbitrarily large. Notice that as N tends to infinity, our Bayes model tends to the true $y(x_0)$. On the other hand, a considerable issue will rise if N is not large enough, we need N to be large enough so that a given x_0 has several pairs (x_0, y_i) in the learning set. Furthermore, a small N also means that the models will be trained on very small data sets thus leading to unreliable errors, biases and variances, they will also be more sensible to noise. These are the main reasons we need N to be large. if it is, it is very unlikely that the latter problems will appear. A problem that will rise for sure is that if N is finite, the Bayes model will never converge to the true mean $y(x_0)$.
To answer the question, the protocols will be appropriate if N is large enough.

2.4 Our protocol

2.4.1 Plot protocol for linear model

We generate a dataset of size N= 10000. We have the shape on this following figure :

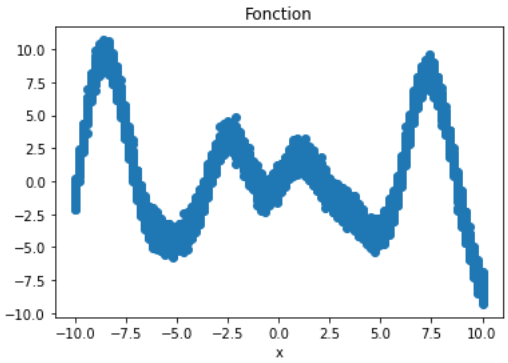

Figure 7: Shape of the fonction

We will observe all different plot for a size of 10000 and a subset equal to 20. For the linear regression method, we use linear regression from sklearn. We observe that the noise oscillate between 0.16 and 0.32, so that correspond of the result that we have to obtain. The squared bias and the expector error are similar due to the low value of the noise. We notice an oscillation with the squared bias and the expected error that may due to the impact of the cosinus and sin of our fonction. We observe an high error when x is between -10 and -7.5 and x between -10 and -7.5. Moreover, we notice a kind of symmetry for our expected error and squared bias. This is may due to the symmetry of our function. The variance is low so our regressor train doesn't vary a lot. We observe a hyperbol shape. The value is very high in the extremities and low in the center.

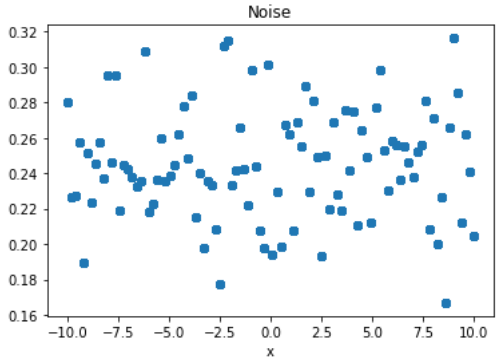

Figure 8: Noise

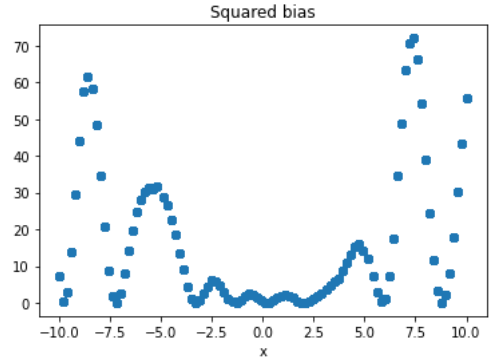

Figure 9: Squared bias

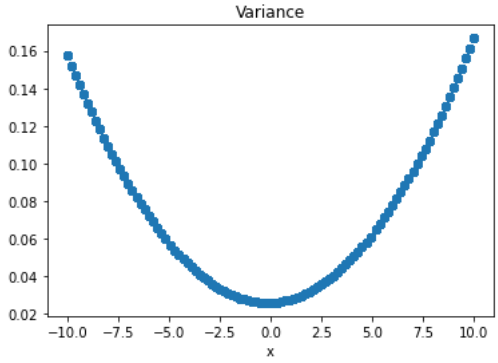

Figure 10: Noise

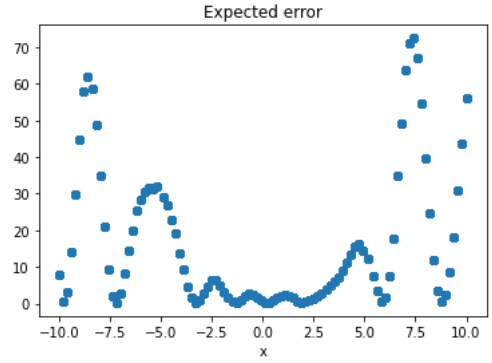

Figure 11: Squared bias

2.4.2 Plot protocol for non-linear model

For the linear regression method, we use non-linear regression from sklearn with k neighbors = 3. We observe that the noise oscillate in the same way. This is what we would expect. Nevertheless, we observe that the squared bias and the expected error vary so much and are not similar. We notice that oscillations that we observe previously are less prevalent. We observe that squared bias has low values so less impact in the expected error. The noise has in this case as much impact as the variance. The shape of the expected error is closer to the noise than the squared bias. This is the opposite of the linear model. We observe that the model is underfitting, we have a low bias and high variance. So the linear model is better than the non-linear model.

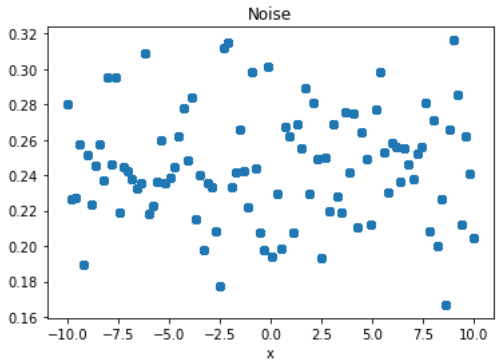

Figure 12: Noise

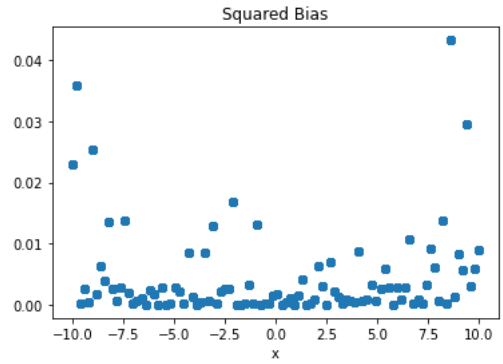

Figure 13: Squared bias

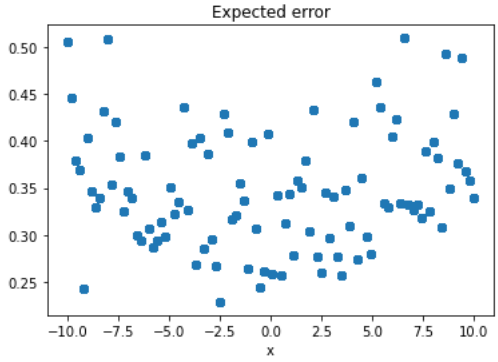

Figure 14: Expected error

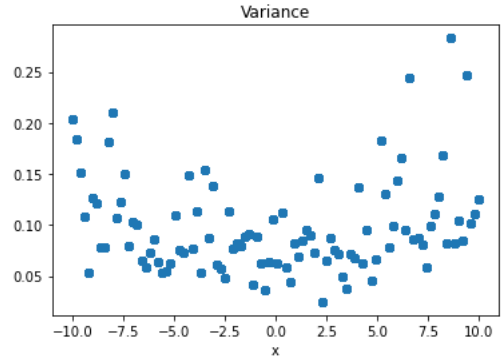

Figure 15: Variance