ELEN0062 : Introduction to machine learning
Project 2 report: Var and variance analysis

Aldeghi Florian(s183157) Baguette Brice (s181482)
Dasnois Louis (s181779)

September 28, 2023

1 Analytical derivations

1.1 Bayes model and residual error in classification

a. The Bayes model for the zero-one error loss can be written as:

$$h_b(x_1, x_2) = \arg\min_{y'} E_y\{\mathbb{1}(y \neq y')|x_1, x_2\} \tag{1}$$

We know that:

$$E_y\{\mathbb{1}(y \neq y')|x_1, x_2\} = P(Y \neq y'|x_1, x_2) \tag{2}$$

Since y can only take two values, the expression is minimized when $P(Y \neq y'|x_1, x_2) \leq 1/2$, which is the same as $P(Y = y'|x_1, x_2) \geq 1/2$. Assuming these probabilities vary continuously in x_1 and x_2 (which they do, seeing the expressions below), the boundary happens where both classes are equally probable. The boundary equation is thus $P(Y = 1|x_1, x_2) = 1/2$. Let us calculate this probability. Through Bayes' theorem, we get:

$$\begin{aligned} P(Y = 1|\mathbf{X} = \mathbf{x}) &= \frac{f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 1)P(Y = 1)}{f_{\mathbf{X}}(\mathbf{X} = \mathbf{x})} \\ &= \frac{f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 1)P(Y = 1)}{f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 1)P(Y = 1) + f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 0)P(Y = 0)} \\ &= \frac{f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 1)}{f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 1) + \alpha f_{\mathbf{X}|Y}(\mathbf{X} = \mathbf{x}|Y = 0)} = \frac{1}{2} \end{aligned} \tag{3}$$

assuming the negative class is α times more likely than the positive class. The PDF of a k-dimensional multivariate Gaussian distribution centered in $\boldsymbol{\mu}$ with a covariance matrix $\boldsymbol{\Sigma}$ is given by:

$$(2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{4}$$

In the case of a circular Gaussian distribution with covariance matrix $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, this reduces to:

$$\frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right) \tag{5}$$

The samples of a given class being distributed this way, we can plug this PDF into equation 3, which gives:

$$\frac{\exp\left(-\frac{1}{2\sigma^2}\left((x_1 + 1.5)^2 + (x_2 + 1.5)^2\right)\right)}{\exp\left(-\frac{1}{2\sigma^2}\left((x_1 + 1.5)^2 + (x_2 + 1.5)^2\right)\right) + \alpha \exp\left(-\frac{1}{2\sigma^2}\left((x_1 - 1.5)^2 + (x_2 - 1.5)^2\right)\right)} = \frac{1}{2} \tag{6}$$

This is equivalent to:

$$\exp\left(-\frac{1}{2\sigma^2}\left((x_1+1.5)^2+(x_2+1.5)^2\right)\right)=\alpha\exp\left(-\frac{1}{2\sigma^2}\left((x_1-1.5)^2+(x_2-1.5)^2\right)\right) \tag{7}$$

Taking the logarithm on both sides and using the logarithm product rule, we arrive to:

$$-\frac{1}{2\sigma^2}\left((x_1+1.5)^2+(x_2+1.5)^2\right)=\ln\alpha-\frac{1}{2\sigma^2}\left((x_1-1.5)^2+(x_2-1.5)^2\right) \tag{8}$$

$$\Leftrightarrow (x_1-1.5)^2+(x_2-1.5)^2-(x_1+1.5)^2-(x_2+1.5)^2=2\sigma^2\ln\alpha \tag{9}$$

We have:

$$\begin{aligned}(x_i-1.5)^2-(x_i+1.5)^2&=((x_i-1.5)+(x_i+1.5))((x_i-1.5)-(x_i+1.5))\\&=-6x_i\end{aligned} \tag{10}$$

So finally, our boundary equation looks like:

$$x_1+x_2=-\frac{\sigma^2\ln\alpha}{3} \tag{11}$$

which represents a line, as we could have expected. In our case, with $\alpha=3$, the Bayes model is thus:

$$h_b(x_1,x_2)=\begin{cases}0 & \text{if } x_1+x_2\geq-\frac{\sigma^2\ln 3}{3}\\1 & \text{otherwise}\end{cases} \tag{12}$$

b. For any ratio α between the negative and the positive class, the Bayes model becomes:

$$h_b(x_1,x_2)=\begin{cases}0 & \text{if } x_1+x_2\geq-\frac{\sigma^2\ln\alpha}{3}\\1 & \text{otherwise}\end{cases} \tag{13}$$

It stays essentially the same, except the boundary gets translated towards the center of the distribution of the class which becomes less prevalent. The boundaries stay parallel, and in the case where both classes are equally likely ($\alpha=1$), the boundary is the line that is equidistant to the two centers.

c. We have:

$$E_{x_1,x_2,y}\{\mathbb{1}(y\neq h_b(x_1,x_2))\}=E_y\{E_{x_1,x_2|y}\{\mathbb{1}(y\neq h_b(x_1,x_2))|y\}\} \tag{14}$$

We know that, thanks to the Bayes model in equation 12, that:

$$E_{x_1,x_2|y}\{\mathbb{1}(y\neq h_b(x_1,x_2))|y=1\}=P\left(X_1+X_2\geq-\frac{\sigma^2\ln 3}{3}\middle|Y=1\right) \tag{15}$$

$$E_{x_1,x_2|y}\{\mathbb{1}(y\neq h_b(x_1,x_2))|y=0\}=P\left(X_1+X_2<-\frac{\sigma^2\ln 3}{3}\middle|Y=0\right) \tag{16}$$

These probabilities can be computed using an integral and the PDFs of the distributions (see equation 5):

$$P\left(X_1+X_2\geq-\frac{\sigma^2\ln 3}{3}\middle|Y=1\right)=\iint_{x_1+x_2\geq-\frac{\sigma^2\ln 3}{3}}\frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\left((x_1+1.5)^2+(x_2+1.5)^2\right)\right)dx_1\,dx_2 \tag{17}$$

$$P\left(X_1+X_2<-\frac{\sigma^2\ln 3}{3}\middle|Y=0\right)=\iint_{x_1+x_2<-\frac{\sigma^2\ln 3}{3}}\frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\left((x_1-1.5)^2+(x_2-1.5)^2\right)\right)dx_1\,dx_2 \tag{18}$$

For the first integral, we use the change of variable:

$$\begin{cases} x_1 = r\cos\theta - 1.5 \\ x_2 = r\sin\theta - 1.5 \end{cases} \tag{19}$$

with a Jacobian r, leading to:

$$\frac{1}{2\pi\sigma^2}\iint_{\sin\theta+\cos\theta\geq\frac{9-\sigma^2\ln 3}{3r}} re^{-\frac{r^2}{2\sigma^2}}\,dr\,d\theta \tag{20}$$

Let's try to solve the condition for θ:

$$\sin\theta+\cos\theta \geq \frac{9-\sigma^2\ln 3}{3r} \tag{21}$$

$$\Leftrightarrow \sin\theta+\sin(\pi/2-\theta) \geq \frac{9-\sigma^2\ln 3}{3r} \tag{22}$$

$$\Leftrightarrow 2\sin(\pi/4)\cos(\pi/4-\theta) \geq \frac{9-\sigma^2\ln 3}{3r} \tag{23}$$

$$\Leftrightarrow \cos(\pi/4-\theta) \geq \frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r} \tag{24}$$

This is always true if the right hand side is smaller than -1 and always false if it is bigger than 1. If it is between -1 and 1, we have:

$$-\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r} \leq \frac{\pi}{4}-\theta \leq \arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r} \tag{25}$$

$$\Leftrightarrow \frac{\pi}{4}-\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r} \leq \theta \leq \arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}+\frac{\pi}{4} \tag{26}$$

The integral becomes:

$$\begin{aligned} &\frac{1}{2\pi\sigma^2}\int_0^{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}} re^{-\frac{r^2}{2\sigma^2}}\,dr\int_0^{2\pi}d\theta \\ &+\frac{1}{2\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}}\,dr\int_{\frac{\pi}{4}-\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}}^{\frac{\pi}{4}+\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}} d\theta \\ &= -e^{-\frac{r^2}{2\sigma^2}}\Big|_0^{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}+\frac{1}{\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}}\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr \\ &= 1-e^{-\frac{(\sigma^2\sqrt{2}\ln 3-9\sqrt{2})^2}{72\sigma^2}}+\frac{1}{\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}}\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr \end{aligned} \tag{27}$$

if $9-\sigma^2\ln 3<0$ and:

$$\begin{aligned} &\frac{1}{2\pi\sigma^2}\int_0^{\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6}} re^{-\frac{r^2}{2\sigma^2}}\,dr\int_0^{0}d\theta \\ &+\frac{1}{2\pi\sigma^2}\int_{\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}}\,dr\int_{\frac{\pi}{4}-\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}}^{\frac{\pi}{4}+\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}} d\theta \\ &=\frac{1}{\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}}\arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr \end{aligned} \tag{28}$$

if $9-\sigma^2\ln 3\geq 0$. Following a similar procedure for the second integral with this variable change:

$$\begin{cases} x_1 = r\cos\theta + 1.5 \\ x_2 = r\sin\theta + 1.5 \end{cases} \tag{29}$$

we get:

$$1 - e^{-\frac{(9\sqrt{2}+\sigma^2\sqrt{2}\ln 3)^2}{72\sigma^2}} + \frac{1}{\pi\sigma^2}\int_{\frac{9\sqrt{2}+\sigma^2\sqrt{2}\ln 3}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}} \arccos\frac{-9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr \tag{30}$$

Finally, putting all of this back into our first equation, we have:

$$\begin{aligned} E_{x_1,x_2,y}\{\mathbb{1}(y \neq h_b(x_1,x_2))\} &= P(Y=1)E_{x_1,x_2|y}\{\mathbb{1}(y \neq h_b(x_1,x_2))|y=1\} \\ &+P(Y=0)E_{x_1,x_2|y}\{\mathbb{1}(y \neq h_b(x_1,x_2))|y=0\} \\ = 1 + \frac{1}{4}\left(-e^{-\frac{(\sigma^2\sqrt{2}\ln 3-9\sqrt{2})^2}{72\sigma^2}} + \frac{1}{\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}} \arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr\right) \\ +\frac{3}{4}\left(-e^{-\frac{(9\sqrt{2}+\sigma^2\sqrt{2}\ln 3)^2}{72\sigma^2}} + \frac{1}{\pi\sigma^2}\int_{\frac{9\sqrt{2}+\sigma^2\sqrt{2}\ln 3}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}} \arccos\frac{-9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr\right) \end{aligned} \tag{31}$$

if $9 - \sigma^2\ln 3 < 0$ or:

$$\begin{aligned} &\frac{1}{4}\left(\frac{1}{\pi\sigma^2}\int_{\frac{\sigma^2\sqrt{2}\ln 3-9\sqrt{2}}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}} \arccos\frac{9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr\right) \\ +\frac{3}{4}&\left(1 - e^{-\frac{(9\sqrt{2}+\sigma^2\sqrt{2}\ln 3)^2}{72\sigma^2}} + \frac{1}{\pi\sigma^2}\int_{\frac{9\sqrt{2}+\sigma^2\sqrt{2}\ln 3}{6}}^{+\infty} re^{-\frac{r^2}{2\sigma^2}} \arccos\frac{-9\sqrt{2}-\sigma^2\sqrt{2}\ln 3}{6r}\,dr\right) \end{aligned} \tag{32}$$

if $9 - \sigma^2\ln 3 \geq 0$.

d. Generating 10 different data sets of 1000 samples and taking the mean for the error rate, we get 0.9246, for 10000 samples we have a value of 0.9245 and for 100 samples, 0.917 for the error rate.

1.2 Var and variance in regression

a. We take Bayes model

$$y' = E_{y|x}\{y\} = E_{y|x}\{a*x+\epsilon\} = E_{y|x}\{a*x\} + E_{y|x}\{\epsilon\} = a*E_{y|x}\{x\} = a*x \tag{33}$$

We have a Bayes model $y' = a*x$
The residual error of this Bayes model is

$$E_{y|x}\{(y - E_{y|x}\{y\})^2\} = E_{y|x}\{(a*x+\epsilon-a*x)^2\} = E_{y|x}\{\epsilon^2\} = \sigma^2 \tag{34}$$

b. The mean squared Var is
$Var^2 = (E_{y|x}\{y\} - E_{LS}\{\hat{y}\})^2 = (a*x - E_{LS}\{\mu\})^2$
Knowing that x has a uniform distribution on [0, 1], the mean of μ on all learning sample is the mean of $y = a*x+\epsilon = 0.5*a$. Hence, we have :

$$Var^2 = (a*(x-0.5))^2 \tag{35}$$

For the variance, we have :
$\mu : \frac{\delta}{\delta\mu}\frac{1}{N} * \sum_{LS}(y_i-\mu)^2 = 0 => \mu = \frac{1}{N}\sum_{LS}\{y_i\}$
$var_{LS}\{\hat{y}\} = var_{LS}\{\frac{1}{N}\sum y_i\} = \frac{1}{N^2}var_{LS}\{\sum_i y_i\}$
Since variables are independents, we can invert sum and var:

$$var_{LS}\{\hat{y}\} = \frac{1}{N}var_y\{y\} = \frac{\sigma^2}{N} \tag{36}$$

c. σ does not influence the Var at all since the mean of $\sigma = 0$. Otherwise, the Var is proportional to the a squared. Actually, the learning algorithm has a nul slope while the true model has a slope of a. Greater a, greater the difference between true model and this one obtained by learning algorithm.
It's totally different for the variance. In fact, the variable a has no influence on the variance of the model of the learning algorithm. No matter the slope, as we are looking the mean, that has no influence. Otherwise, σ is directly related to the variance since it represents the variation between sample with same x.

2 Empirical analysis

a. We generate a finite number of y for a x_0 that is given, then we compute the variance over the y that we get and we obtain residual error.

For the Var, we evaluate the mean of y in the same way that we did for the residual error. Then we evaluate the mean of $\hat{y}$ by computing the LS over different data sets and finally we evaluate the squared of the difference between those two means:

$$\text{Var}^2 = (E(y) - E(\hat{y}))^2$$

Finally for the variance, we re-use $\hat{y}$ and we compute the variance such as

$$\text{var}_{\text{LS}}(\hat{y}) = E_{LS}\{(\hat{y} - E_{LS}(\hat{y}))^2\}$$

b. We now consider pairs (x, y), we use protocol that is described at the point *2.a* as we consider every x from the pairs as x_0 and then we calculate the mean values over the pairs for each item.

c. We only use a finite number of samples for our estimates, so we can use our finite number of samples as if they were generated sequentially. The only difference is that we can no longer choose an input x_0 for which we generate output values y. Instead, we have to assume that for two values of x that are close to each other, the corresponding y distributions are similar. Then we can simulate the generation of new values of y for a given x_0 by taking (x, y) pairs in our finite sample for which x is close to x_0.

d. In first place, we'll have to set a number of samples arbitrarily, we choose $nSamples = 1000$, we found it as a good trade between efficiency and accuracy for our evaluations.
In second place, we had to choose two model, one linear regression and one non-linear regression. For the linear model, we choose a Ridge regression model because it offers us the possibility to modify an α term for complexity performance evaluation at point *2.e*. On the other side, we choose a KNN regression for the non-linear model. Since it is a model that we used a lot in the course and that we understand well, we thought it was the best choice.

- Firstly, let's plot the residual error which is $var(y)$.

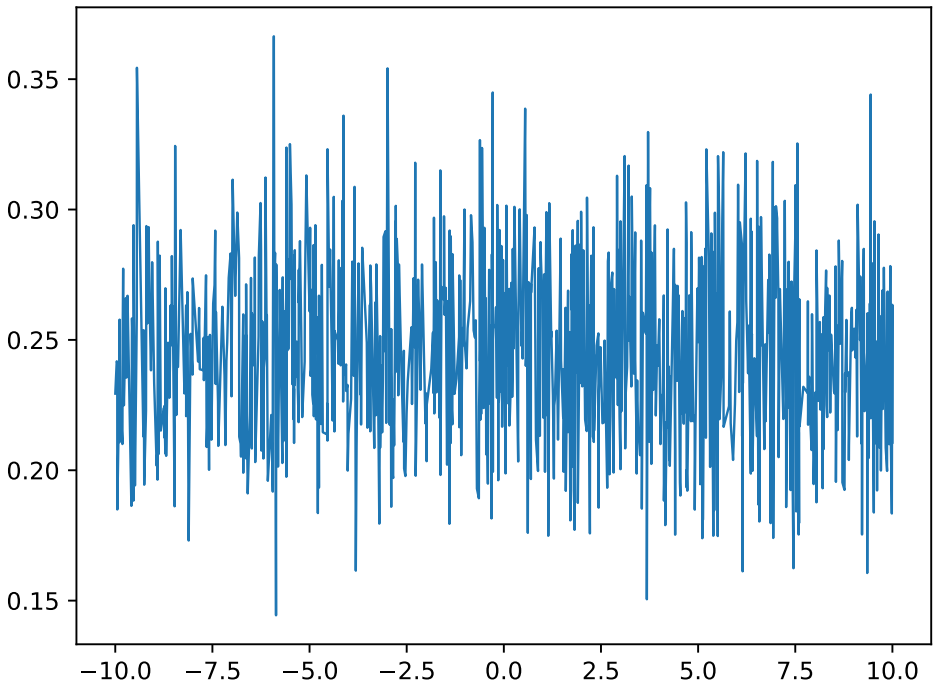
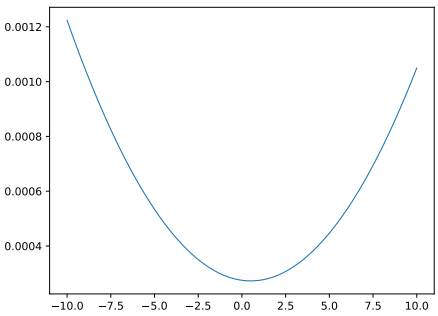


Figure 1: Residual error

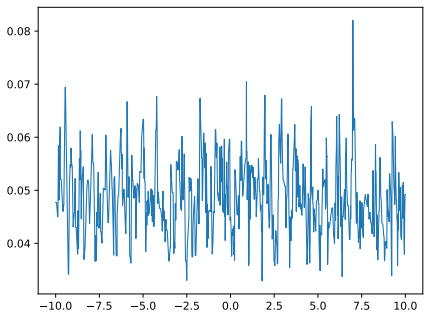
We observe that the residual error oscillate around 0.25 which is the variance of the Normal distribution of the noise variable ϵ.

Now, we'll have to differentiate the linear model from the non-linear model since the value of the training sample are used to evaluate the Var and the variance.

- For the Var, we observe a big difference between the linear model and the non linear, this means that there's a tendency to fit more the learning sample in the non-linear model

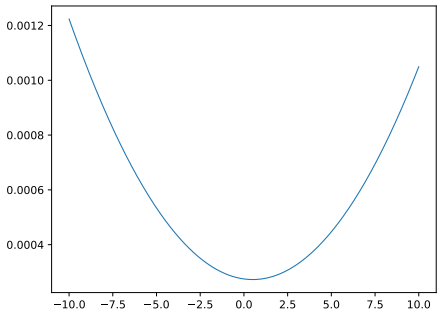


(a) Var of the linear model

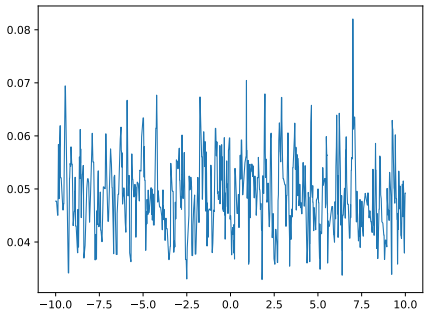


(b) Var of the non-linear model

- Finally for the variance, we observer a higher variance in the non linear model, it makes sense since there's a trade off between variance and Var

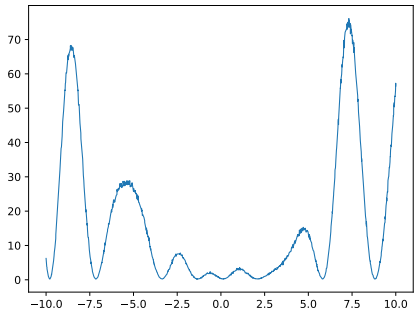


(a) Variance of the linear model

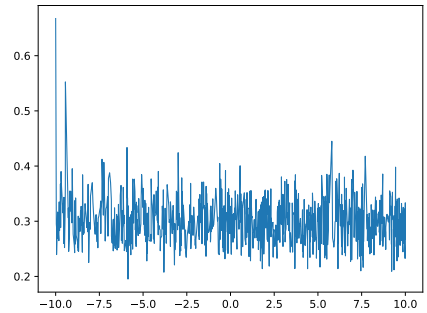


(b) Variance of the non-linear model

- Finally we have the expected error which is the sum of the residual error, the Var and the variance.



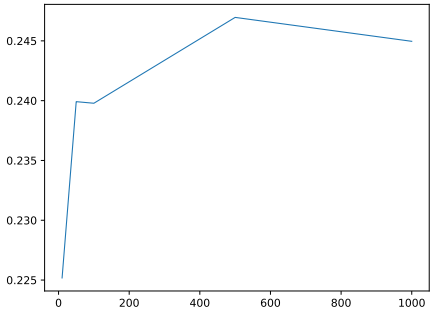
(a) Expected error of the linear model



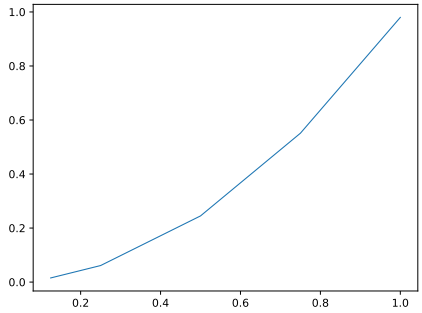
(b) Expected error of the non-linear model

e. For this point, we will use the tuple $[10, 50, 100, 500, 1000]$ for the number of samples, $[0.125, 0.25, 0.5, 0.75, 1]$ for the standard deviation, $[0.25, 0.5, 0.75, 1, 2]$ for alpha parameters in Ridge regression and $[1, 2, 5, 10, 50]$ for nNeighbors parameters for the KNN regression.

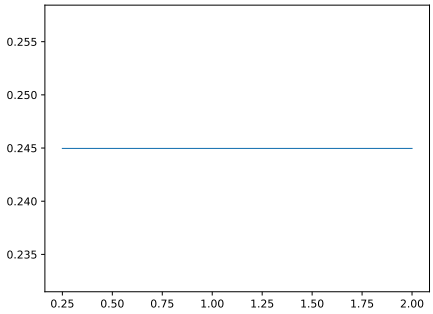
(a) Since the residual error is independent from the model, we can see that the variation of the complexity, i.e. nNeighbors and alpha parameters, doesn't change the residual error. Obviously, the more samples we have, the more the variance of y is small and converge to the standard deviation of the noise. When we augment the standard deviation, the residual error follows its since the 2 values are directly links in y formula.



(a) Residual error with variation of number of samples

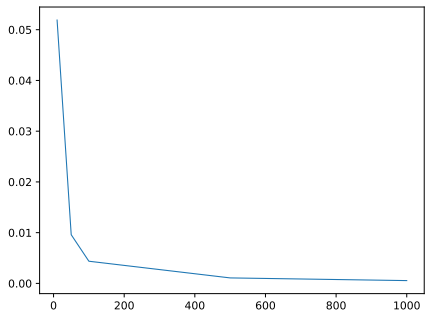


(b) Residual error with variation of number of standard deviation

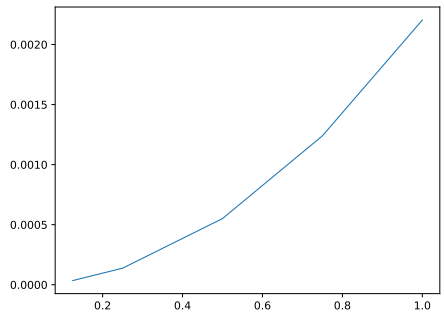


(c) Residual error with variation of number of complexity

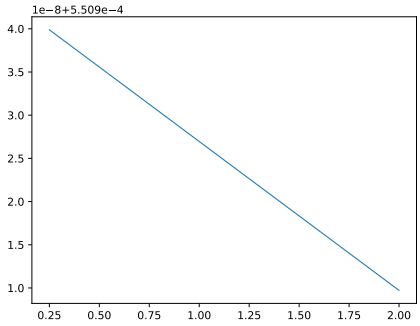
(b) For the linear model, the Var increase pretty fast with number of samples and then stabilizes. The Var decreases when the standard deviation increase. It's the opposite for the complexity



(a) Linear model Var with variation of number of samples

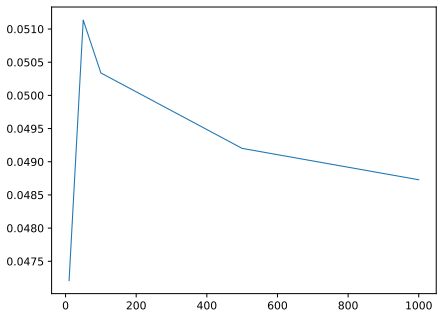


(b) Linear model Var with variation of number of standard deviation

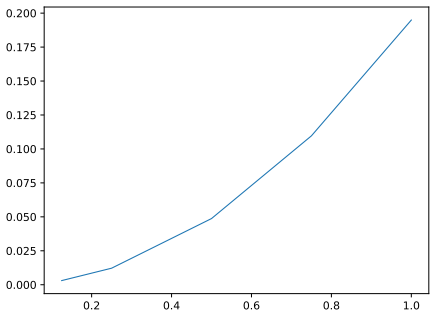


(c) Linear model Var with variation of number of complexity

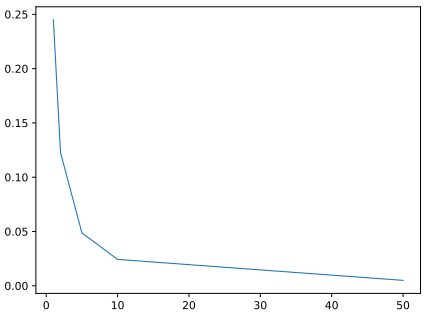
For the non linear model, the Var decreases and the stabilizes with number of samples, basically the opposite of the linear model. Then Var increases with both standard deviation and complexity.



(a) Non linear model Var with variation of number of samples

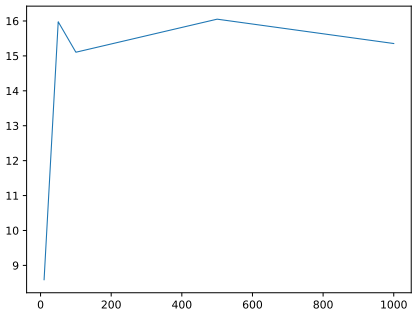


(b) Non linear model Var with variation of number of standard deviation

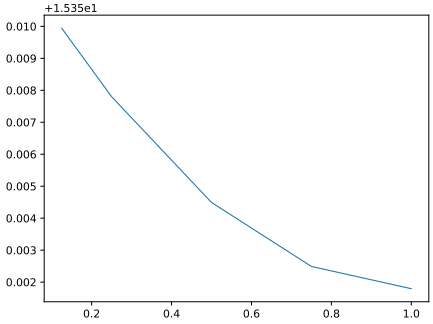


(c) Non linear model Var with variation of number of complexity

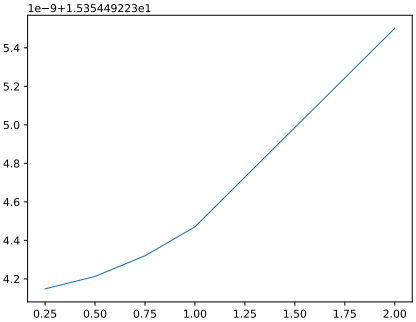
(c) Like we discuss before there's a trade off between variance and bias, so the comportment of the bias is the total opposite of the comportment of the variance



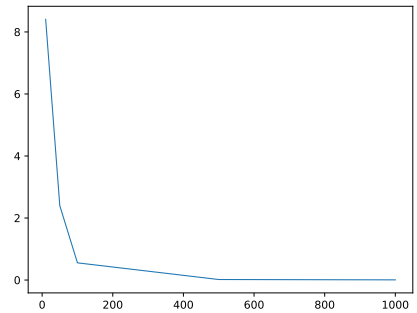
(a) Linear model bias with variation of number of samples



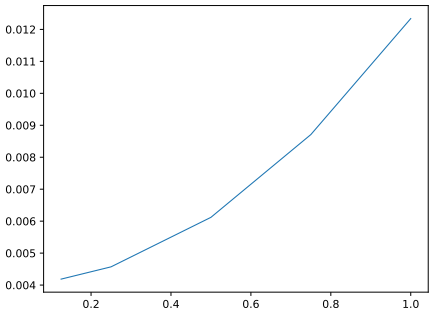
(b) Linear model bias with variation of number of standard deviation



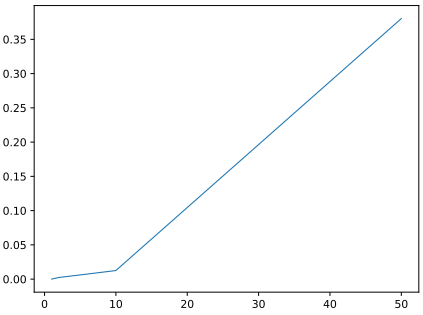
(c) Linear model bias with variation of number of complexity



(d) Non linear model bias with variation of number of samples



(e) Non linear model bias with variation of number of standard deviation



(f) Non linear model bias with variation of number of complexity

(d) The squared error is based on all previous items, it's a combination of all previous points.