FTML Project

Corentin Pion and Youri Colera

17 June 2022

1 Exercise 1

1.1 Question 1

Settings:
- Input space [0;10]
- Output space : R/0
- X follow an uniform distribution
- loss function is squared error l(x,y) = (x-y)(x-y)
- Y follow an exponential distribution Exp(c+X) with c superior or equal 1

We have seen in the lectures that the Bayes estimator with the squared loss function is equal to :

$$f^*(x) = E[Y|X = x]$$

We now that

$$E[Exp(\theta)] = \frac{1}{\theta}$$

so :

$$f^*(x) = E[Y|X = x] = \frac{1}{\theta + c}$$

Now that we have our estimator we can use it to calculate our risk

$$R^* = E[l(Y, f^*(X))]$$

$$R^* = E[E_y[(Y - f^*(X))^2|X]]$$

we can see

$$Var(X) = E[(X - E[X])^2]$$

$$E_y[(Y - f^*(X))^2|X] = E_y[(Y - E_y[Y| = x])^2|X] = E[Var(Y|X)]$$

The variance of the the Distributed Exp(theta) is equal to

$$Var(X) = \frac{1}{\theta^2}$$

so for our example we have:

$$R^* = E[E_y[(Y - f^*(X))^2|X]]$$

$$R^* = E[\frac{1}{(c+X)^2}] = \int_0^{10} \frac{1}{(c+X)^2}$$

$$R^* = \left[\frac{-1}{c+X}\right]_0^{10}$$

$$R^* = \frac{-1}{c+10} + \frac{1}{c}$$

If we want we can try to discuss the result with different value for c and compare them

2 Exercise 2

We consider a supervised regression problem, Y = R. We have seen that when the loss used is the square loss l2(y, z) = (y z), then the Bayes predictor is the conditional expectation

2.1 Question 1

We will use the same distribution from the first exercise with c = 0

First, let's calculate a Bayes estimator with the squared loss function We use our result from the first exercise

$$f^*(x) = E[Y|X = x] = \frac{1}{\theta}$$

When we use the squared error we get the Esp() of the exponential
Now if we switch to an absolute loss function we have

$$f^*(x) = E[|y - z||X = x] = \frac{Ln(2)}{\theta}$$

We can conclude that it's possible that the Bayes estimator can change if the loss function is changed

2.2 Question 2

Let's devellop the function g(z)

First let's call h the density function for a better comprehension

$$g(z) = \int_{-\infty}^{\infty} |y - z|h(y)dy$$

$$g(z) = \int_{-\infty}^{z} (z - y)h(y)dy + \int_{z}^{\infty} (y - z)h(y)dy$$

$$g(z) = z[1 - F(y)] + \int_{-\infty}^{z} -y * h(y)dy - z[F(y)] + \int_{z}^{\infty} yh(y)dy$$

Next we need to check the value where the derivate of g is equal to 0

$$g(z) = z[1 - F(y)] + \int_{-\infty}^{z} -y * h(y)dy - z[F(y)] + \int_{z}^{\infty} yh(y)dy$$

$$g'(z) = 1 - F(y) + zh(y) - zh(y) - F(y) + zh(y) - zh(y)$$

$$g'(z) = 1 - 2F(y)$$

$$F(y) = \frac{1}{2}$$

This result is telling us that we also have the median

3 Exercise 3

We want to show that in the linear model, fixed design we have

$$E[R_x(\widehat{\theta})] = \frac{n-d}{n}\sigma^2$$

In this expression, both y and ˆ are random variables, that are not independent, as theta is the OLS estimator. The expectation is over the distribution of both variables.

3.1 step 1

Show that :

$$E[R_n(\widehat{\theta})] = E_\epsilon[\frac{1}{n}\|I_n - X(X^tX)^{-1}X^t)\epsilon\|^2]$$

$$E[R_n(\widehat{\theta})] = E[\frac{1}{n}\|y - X\widehat{\theta}\|^2]$$

$$E[R_n(\widehat{\theta})] = E[\frac{1}{n}\|X\theta + \epsilon - X\widehat{\theta}\|^2]$$

$$E[R_n(\widehat{\theta})] = E[\frac{1}{n}\|X\theta + \epsilon - X(X^tX)^{-1}X^ty\|^2]$$

$$E[R_n(\widehat{\theta})] = E[\frac{1}{n}\|X\theta + \epsilon - X(X^tX)^{-1}X^t(X\theta + \epsilon)\|^2]$$

$$E[R_n(\widehat{\theta})] = E_\Theta[\frac{1}{n}\|X\theta - X(X^tX)^{-1}X^tX\theta\|^2] + E_\epsilon[\frac{1}{n}\|\epsilon - X(X^tX)^{-1}X^t\epsilon\|^2]$$

$$E[R_n(\widehat{\theta})] = E_\Theta[\frac{1}{n}\|0\|^2] + E_\epsilon[\frac{1}{n}\|(I_n - X(X^tX)^{-1}X^t)\epsilon\|^2]$$

$$E[R_n(\widehat{\theta})] = E_\epsilon[\frac{1}{n}\|(I_n - X(X^tX)^{-1}X^t)\epsilon\|^2]$$

3.2 step 2

Let A Rn,n. Show that

$$\sum_{i=1}^{n} A_{ij}^2 = tr(A^t A)$$

To show this demonstration we can use the definition of the scalar and the trace function

$$tr(A^t A) =< A, A >= \sum_{i=1}^{n} A_{ij}^2$$

3.3 step 3

Show that

$$E[\frac{1}{n}\|A\epsilon\|^2] = \frac{\sigma^2}{n} tr(A^t A)$$

$$E[\frac{1}{n}\|A\epsilon\|^2] = E[\frac{1}{n}\sum_{i=1}^{n}(A_{ij}\epsilon)^2]$$

$$E[\frac{1}{n}\|A\epsilon\|^2] = \frac{1}{n} tr(A^t A) E[\epsilon^2]$$

But

$$E[\epsilon^2] = E[\epsilon^2 - E[\epsilon]^2] = Var = \sigma^2$$

Since in the subject it's said that E(epsilon) is actually equal to 0 So in the end we have

$$E[\frac{1}{n}\|A\epsilon\|^2] = \frac{\sigma^2}{n} tr(A^t A)$$

3.4 step 4

We note

$$A = I_n - X(X^t X)^{-1} X^t$$

Show that

$$A^t A = A$$

$$A^t A = (I_n - X(X^t X)^{-1} X^t)^t (I_n - X(X^t X)^{-1} X^t)$$

$$A^t A = I_n - 2I_n X(X^t X)^{-1} X^t + X(X^t X)^{-1} X^t X(X^t X)^{-1} X^t$$

We can reduce the last term to this:

$$X^t X(X^t X)^{-1} = I_d$$

$$X(X^t X)^{-1} X^t X(X^t X)^{-1} X^t = X(X^t X)^{-1} X^t$$

In the end we can simplify

$$A^tA = I_n - 2I_nX(X^tX)^{-1}X^t + X(X^tX)^{-1}X^tX(X^tX)^{-1}X^t$$

$$A^tA = I_n - 2X(X^tX)^{-1}X^t + X(X^tX)^{-1}X^t$$

$$A^tA = I_n - X(X^tX)^{-1}X^t$$

$$A^tA = A$$

3.5 step 5

We can now show that our proposition in the start is true

$$E[R_x(\widehat{\theta})] = E[[E[R_n(\widehat{\theta})]]$$

we have

$$[E[R_n(\widehat{\theta})] = \frac{\sigma^2}{n}tr(I_n - X(X^tX)^{-1}X^t)$$

we now use trace proprieties tr(a+b) = tr(a) + tr(b):

$$[E[R_n(\widehat{\theta})] = \frac{\sigma^2}{n}tr(I_n) - tr(X(X^tX)^{-1}X^t)$$

we now use trace proprieties tr(abc) = tr(cab)):

$$[E[R_n(\widehat{\theta})] = \frac{\sigma^2}{n}tr(I_n) - tr(X^tX(X^tX)^{-1})$$

$$[E[R_n(\widehat{\theta})] = \frac{\sigma^2}{n}tr(I_n) - tr(I_d)$$

$$[E[R_n(\widehat{\theta})] = \frac{\sigma^2}{n}(n-d)$$

if we take back our first equation:

$$E[R_x(\widehat{\theta})] = E[[E[R_n(\widehat{\theta})]] = \frac{\sigma^2}{n}(n-d)$$

3.6 step 6

The expected value is:

$$\frac{\|y - \widehat{\theta}\|}{n-d}\sigma^2 = \frac{n-d}{n-d}\sigma^2 = \sigma^2$$

3.7 step 7

With our setting we can approach a estimation of sigma Look at the python file and the image

4 Exercise 4

For the regression part and the classification part, we have decided to chose the scikit-learn library. In fact, it contains many regression and classification models that will be useful for these 2 parts. There is 2 important properties that we have to retrieve from the dataset. There is 1000 entries and 20 features. This needs a regression model that can work with a small set of samples and 20 features. Moreover, the r2 score is influenced by the random initialization of the models parameters, The Lasso regression seems to be the best regression model in this case.

5 Exercise 5

Like the previous section, the dataset has 1000 entries and 20 features. For this reason, we need a classification model that works with a small set of samples and 20 features. We can also note that the accuracy is highly influenced by the random initialization of models parameters, the linear SVC seems to be the best model in this case.

Both result for the question 4 and 5 can be found in the Jupiter file