Fintech – Homework 1

1) Linear regression:

Question a

In this question, I convert the data to make it usable in the rest of the homework. I chose to use the pandas library to create data frames and to easily manipulate the data.

I also had to replace the text parts by binary values. I chose the following codes:

School:	GP -> 1	MS -> 0
Sex:	M -> 1	F -> 0
Famsize :	GT3 -> 1	LE3 -> 0
Activities, higher, internet, romantic	Yes -> 1	No -> 0

To normalize the data, I subtract to each value the mean and then divide it by the standard deviation of the training set. Finally, I split the data into two parts: the training set (800 lines) and the test set (200 lines). The lines are randomly split.

Question b

In this part, I implement a linear regression without bias. It is possible to obtain the optimal weights thanks to this formula:

$$Weights = (X^T X)^{-1} * X^T * Y$$

X = training set data
Y = G3 of the training set
Then, I calculate G3 predicted (Ypred): $Ypred = Xtest * Weights$

The RMSE is calculated with the following formula:

$$RMSE = \|Ypred - Ytest\| = \sqrt{\frac{1}{m}\sum(Ypred[i] - Ytest[i])^2}$$

m = length of Ypred
Ytest = G3 of test set

Question c:

In this part, I regularise the linear regression to get a better result by adding a term to the cost function.

$$J(w) = \text{MSEtrain} + \frac{\lambda}{2} w^T w = \frac{1}{m}\|Ytrain - Xtrain * w\|^2 + + \frac{\lambda}{2} w^T w$$

$$Weights = \left(X^T X + m * \frac{\lambda}{2} * I\right)^{-1} * X^T * Y$$

m = row number of the training set

The maximum likelihood criterion can be used to get the optimal weights. It is given by:

$$Wopti = arg\max_{W} \sum \log Pmodel(X(i), W)$$

Indeed, it is possible to show that maximising the likelihood is the same than minimising the cross-entropy function between (Pmodel,P(X(i))). In other words, this process tries to match P(X(i)) with Pmodel and give us the closest Ypred from Ytest.

If we apply this principle to the problem, we can consider that Pmodel is a Gaussian distribution with a mean vector W and a standard deviation I (identity):

$$Wopti = arg\max_{W} \sum \log N(X(i), W, I)$$

We can solve this optimisation problem with the gradient descent method or the Newton's method and get the optimal weights.

Question d:

To get a better result, I add a bias term (b) to the model:

$$\mathrm{J(w)} = \mathrm{MSEtrain} + \frac{\lambda}{2}\mathrm{w^T w} = \frac{1}{m}\|Ytrain - (Xtrain * w + b)\|^2 + + \frac{\lambda}{2}\mathrm{w^T w}$$

To simplify the calculations, I add a column of 1 to the Xtrain and the Xtest matrix. In a result, a new value will appear in the weight vector which will be the bias term (b). We can calculate the optimal Weights and bias with the following formula:

$$Weights = \left(X^T X + m * \frac{\lambda}{2} * I\right)^{-1} * X^T * Y$$

Question e:

In this part, I implement a Bayesian linear regression. Then, I use the means of the posterior model as weights for our models.

$$Weights = (X^T X + \alpha I)^{-1} * X^T * Y$$

We can notice that we almost get the same formula than for the d part.

Question f:

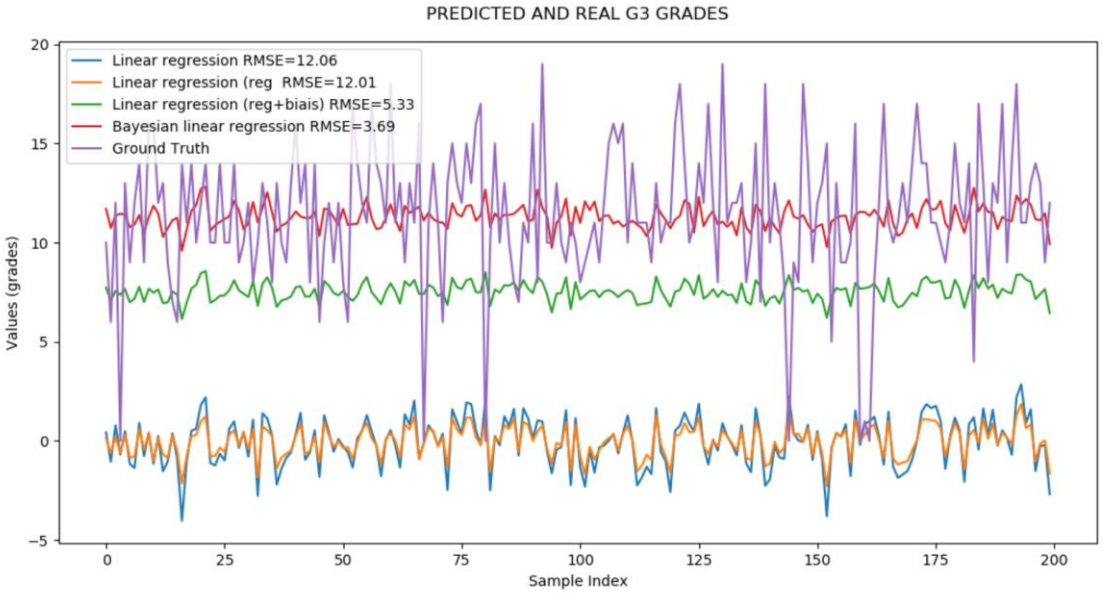
I obtained the following curves during one of my tests (remember that the training and test set are randomly combined).

It is noteworthy that the simple linear regression RMSE is very close from the regularized one. Nevertheless, it seems like the second one has less sharps picks inducted by the generalisation error. It is more likely to be closer to the reality.

Even if these curves have a good shape, they are still centred to zero far away from the real data. That is why I added a bias term to the other regressions (simple with bias and Bayesian). It is noteworthy that the RMSE is highly improved thanks to this term. It is located around 5,3 for the simple regression and 3,6 for the Bayesian linear regression.

We can se that the 3rd and 4th models come from the almost identical formula and are very similar. But if we had chosen a more complex prior distribution, the Bayesian regression would have

been more precise. This let me think that It is still possible to get a better RMSE with the Bayesian regression.



2) Hidden test set:

In this question, I use the e model to a new dataset: test_no_G3. In order to chose which alpha could be the optimal, I created a function optialpha. This one calculates the RMSE of the testset with different values of alpha and return the one with the lowest RMSE. Whit this method and with a training set composed of the 800 first data lines of “train.csv”, I found a value α=2,89.

Then, I calculate the predicted values of G3 and finally put then into the text file.