Name: Kimanh Pham (38936)
Phuong Lieu (38626)
Mahdi Sayyadi (508820)

DATA MINING ASSIGNMENT 2

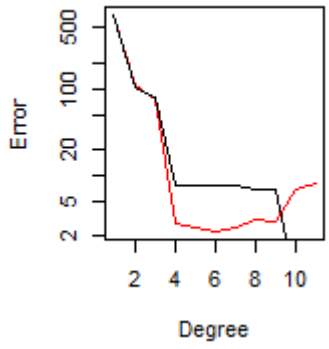
Q1.a:

There two possibilities to choose the model: first on the basis of training error and the second on the basis of validation error. Choosing the best model based on training error can be biased, but the validation error can give us the best result for this case. At degree n=6 can describe the best climate in Turku. The choice is ambiguous because with n=4,5,6,7,8,9 they showed similar pattern on the plot and when you get and compare the error. At the degrees 4,5,7 and 8 aren't much far away from the result from degree "6". However, at degree n=6, the validation error is smallest (2.317495) and reasonable training error (7.538568). We cannot choose the 10 or 11 because it might lead to overfitting.

Q1.b:

In order to evaluate if a polynomial model is reliable, we have to measure error of the model after a certain amount of time. Typically, we can evaluate the squared error to see how fit of the model does, based on the training error and variance between training error and validation error. The plot illustrates the relationship between the training error and the validation error. Additionally, we need untouched set of data. Test set can show us the reliability of the model and in order to assess the current model we require another data set to see whether the output model can give us a trustworthy result.

However, there is high probability of prediction weather due to external atmospheric factors. Therefore, it's difficult to evaluate a polynomial model of weather.



Q2.

The linear boundary does not separate well two classes. There are quite an amount of people with diabetes located in the same class with healthy people. Therefore, it is not feasible to build a good

classifier based on only 2 variables plasma.glucose and BMI. Moreover, there are several cases are not classified clearly because it is located in the linear line.

Q3.

The following classification performance measures are counted manually for linear model.

Contingency Table	Actually negative (observation)	Actually positive (observation)
Predicted negative (expectation)	30 (TN)	10 (FN)
Predicted positive (expectation)	18 (FP)	57 (TP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)=(30+57)/(30+10+18+57)=0.756
So, the accuracy rate is 75.6%, classification error is 24.4%

Q4.a:

Contingency Table with LDA model:

```
observation
prediction  0  1
         0 85 30
         1  6 32
```

Error rate `[1] 0.2352941`

- ⇨ Accuracy rate: 76.5%

Q4.b:

Prior probability of class `1:` `[1] 0.3360656`

- ⇨ 33.6%

Posterior class:

```
observation
prediction  0  1
         0  0  0
         1  91 62
```

- ⇨ Probability with LDA model: 40.5% (>33.6%)

Therefore, we can assume that LDA model gives more accurate prediction than the consistently guessing on the a priori maximum-likelihood class.

Q5:

	Threshold =0.25		Threshold =0.5		Threshold =0.75	
Prediction/Observation	0	1	0	1	0	1
0	70	13	85	30	88	43
1	21	49	6	32	3	19
Accuracy rate	77.7%		76.4%		69.9%	

In this case, false negative is worse than false negative because if the person is predicted as healthy but actually they have diabetes, the delay of research of healthy and treatment might happen which will

lead for worse health status. Therefore, threshold 0.25 is a better option in this case which gave the least amount of false negative. Moreover, it also gave the highest accuracy rate.

Q6.a:

Probabilities of unknown data

(0: non-diabetes and 1: diabetes)

Observation	0	1
370	0.9158641	0.084135863
389	0.3497488	0.650251210
44	0.7227943	0.277205738
211	0.6179306	0.382069383
622	0.9974387	0.002561332

Due to the balance value 0.5, we can classify the 370, 44, 211, 622 are non-diabetes and the 389 is diabetes.

Q6.b

The model gave a really high confidence with two cases 370 and 622 with highest probability of prediction accuracy (91.5% and 99.7%). Other cases in unknown dataset are also fairly high. Thus, we can trust this model.

Q7:

	LD1
pregnancies	0.305241113
plasma.glucose	0.856615805
blood.pressure	-0.223669036
triceps.skin.thickness	0.002242377
insulin	-0.070237055
bmi	0.493777925
diabetes.pedigree	0.199788578
age	0.162322335

Plasma.glucose has the most discriminatory power with respect to the presence of diabetes with coefficient 0.8566 and the least discriminatory power is triceps.skin.thickness with coefficient 0.002242377. High values in plasma glucose lead to very high possibility in the presence of diabetes and change in triceps skin thickness does not affect or affect little to the probability of diabetes presence.

Q8.

	Threshold =0.25 Hidden=3		Threshold =0.5, hidden =4		Threshold =0.7 Hidden=6	
Prediction/Observation	0	1	0	1	0	1
0	53	15	73	25	84	29
1	38	47	18	37	7	23
Accuracy rate	65.3%		71.9%		69.9%	

The neural network can perform well on this exact case, where output value is binary (contains diabetes and no diabetes) based on many different variables input which can be recommended to use for this case. Then classification can be used to determine the accuracy rate which is used to compare with the

one in LDA model. According to two tables above, they have slightly differences in accuracy rate and true positive value.