NNML Formative Assessment, 2019

Name - Harish Iyer
Student No. - 18018054

One paragraph explaining how you solve task 1.

For this activity confusion matrix is produced manually for Gaussian Process as well as Simple Regression. As given 1 to denote a good enhancer, and 0 poor enhancer, naturally 1 is associated to Positive and 0 is associated to Negative.
For Gaussian Process (GP), first given prediction data is loaded in excel ERGP column.
In an adjacent column ERActuals (Actual values), actual outcome is loaded.
By comparing these two columns (using excel-attached), it is clear whether or not Predictions match the actual given observations for each prediction.
Confusion matrix for Gaussian Process is created as below

TN = Filter matching 0 value (ERGP Predicted column), these are True Negatives.
TP = Filter matching 1 value, these are True Positives
FN = Filter non-matching 0 predictions, these are predicted negative, but should be actually positive (as per Actual values). Hence these are false negative.
FP = Filter non-matching 1 predictions (ERGP Column), these are positive values which should rather be negative. Hence these are false positive.

TN	FP
FN	TP

56	3
4	8

Similarly, we calculate for Simple Regression (SR)

TN	FP
FN	TP

53	6
7	5

A copy of the source code you have written for task 2.

```
disp('Load Actual Results')
load ER_labels.txt;
ERActual=ER_labels;

disp('Load GP Predictions')
load ER_GP_predictions.txt;
```

```
ERGP=ER_GP_predictions;
disp('Confusion Matrix for GP Data')
[C_GP,Rate_GP]=confmat(ERGP,ERActual)
TN=C_GP(1,1);
FN=C_GP(2,1);
FP=C_GP(1,2);
TP=C_GP(2,2);
disp('Calculations for GP')
GP_Recall= TP/(TP+FN)
GP_Precision=TP/(TP+FP)
GP_F_Score=2*GP_Recall*GP_Precision/(GP_Recall+GP_Precision)
GP_FP_Rate=FP/(FP+TN)

disp('Load SR Predictions')
load ER_SR_predictions.txt;
ERSR=ER_SR_predictions;
disp('Confusion Matrix for SR Data')
[C_SR,Rate_SR]=confmat(ERSR,ERActual)
TN=C_SR(1,1);
FN=C_SR(2,1);
FP=C_SR(1,2);
TP=C_SR(2,2);
disp('Calculations for SR')
SR_Recall= TP/(TP+FN)
SR_Precision=TP/(TP+FP)
SR_F_Score=2*SR_Recall*SR_Precision/(SR_Recall+SR_Precision)
SR_FP_Rate=FP/(FP+TN)
```

Details of the command line instructions you used for task 2.

In the above script first actual data is loaded and assigned to ERActual which is a table of actual predictions. Similarly predicted GP data is laoded.
Then by calling confmat function, Confusion (C_GP) and Rate (RateGP) matrices are generated.
Each of the column in the C_GP matrix are loaded element wise to corresponding TN,FN,FP and TP.
Then calculations are carried out based on given/known formulas.

Then SR data is loaded and calculations are carried out for SR in the same way as GP.

The final results obtained in task 2.

Output of the Matlab script is as below. It matches with the manual calculations.

```
K>> Load_Data_and_Calc_Rates
Load Actual Results
Load GP Predictions
Confusion Matrix for GP Data
C_GP =
```

```
   56     3
    4     8

Rate_GP =

  90.1408   64.0000

Calculations for GP
GP_Recall =

   0.6667
GP_Precision =

   0.7273
GP_F_Score =

   0.6957
GP_FP_Rate =

   0.0508

Load SR Predictions
Confusion Matrix for SR Data
C_SR =

   53     6
    7     5

Rate_SR =

  81.6901   58.0000

Calculations for SR
SR_Recall =

   0.4167

SR_Precision =

   0.4545

SR_F_Score =

   0.4348

SR_FP_Rate =
```

```
   0.1017

K>>
```

An answer to the question presented in task 3.

It can be observed from the above calculations that Precision for Gaussian process is 0.72 whereas for Simple Regression is 0.45. So Gaussian process has given better precision. Similarly, False Positive Rate for Gaussian process is 0.05 and that of Simple regression is 0.1. This means Simple regression is giving almost 50% more False Positive rate as compared to Gaussian process. For a larger amount of data, in case we use Simple Regression, we are likely to get a lot of inaccurate false positive, which corresponds to good enhancers predicted incorrectly. Hence, it may be more appropriate to use Gaussian Process for these calculations.