Kevin Tudor
9/9/2022

Assignment 1B:
Modeling Assignment: Prediction

Due date: Sept. 28, '22

No late submissions are accepted.

Students must work individually to complete this assignment.

Submit 4 days in advance and you will get 10% bonus or extra points

This part of the project will build models to predict the number of faults based on the other attributes of the instances.

Each model is to be first built and evaluated using 10-fold cross validation on the fit data set, and then validated using the test data set.

Use the data sets prepared for prediction in the previous assignment of the project.

Build the following prediction models:

1. Linear Regression

- For **linear regression**, compare the model selection methods: greedy, M5, no selection.

2. Decision Stump

Compare the models, how many and which independent variables were selected?

Use the statistical indicators provided by Weka to perform the comparisons.

Your report should include all the results based on 10-fold cross-validation and on the test data set. You should also compare the results of all the methods.

Table of Contents

1. Normalized Table Data

Cross-Validation (10-Fold)	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Total Number of Instances
No Selection	0.8358	1.6177	2.6141	53.9848 %	55.0596 %	192
M5	0.8266	1.6459	2.6873	54.9255 %	56.6001 %	192
Greedy	0.8261	1.6569	2.6824	55.2924 %	56.4984 %	192
Decision Stump	0.6746	2.1841	3.5237	72.8887 %	74.2175 %	192

FAULTS	No Selection	M5	Greedy	Decision Stump Classifications:
NUMUORS *	-0.054 +	-0.054 +	-0.0436 +	
NUMUANDS *	0.0366 +	0.0366 +	0.0351 +	**NLOGIC <= 14.0**
TOTOTORS *	0.0031 +	0.0031 +		1.3854748603351956
TOTOPANDS *	-0.0062 +	-0.0062 +	-0.0024 +	
VG *	-0.0441 +	-0.0441 +	-0.0369 +	**NLOGIC > 14.0:**
NLOGIC *	0.2162 +	0.2162 +	0.2214 +	15.615384615384615
LOC *	0.0013 +	0.0013 +	0.0017 +	
ELOC *	0.0051 +	0.0051 +		**NLOGIC is missing:**
FAULTS	-0.3001	-0.3001	-0.3583	2.3489583333333335

User – Supplied Test Set	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Total Number of Instances
No Selection	0.8687	1.9577	3.9570	56.532 %	59.719 %	96
M5	0.8687	1.9577	3.9570	56.532 %	59.719 %	96
Greedy	0.8710	1.9616	3.9258	56.6442 %	59.2485 %	96
Decision Stump	0.7578	2.4427	4.5254	70.5379 %	68.2982 %	96

FAULTS	No Selection	M5	Greedy	Decision Stump Classifications:
NUMUORS *	-0.054 +	-0.054 +	-0.0436 +	
NUMUANDS *	0.0366 +	0.0366 +	0.0351 +	**NLOGIC <= 14.0**
TOTOTORS *	0.0031 +	0.0031 +		1.3854748603351956
TOTOPANDS *	-0.0062 +	-0.0062 +	-0.0024 +	
VG *	-0.0441 +	-0.0441 +	-0.0369 +	**NLOGIC > 14.0:**
NLOGIC *	0.2162 +	0.2162 +	0.2214 +	15.615384615384615
LOC *	0.0013 +	0.0013 +	0.0017 +	
ELOC *	0.0051 +	0.0051 +		**NLOGIC is missing:**
FAULTS	-0.3001	-0.3001	-0.3583	2.3489583333333335

2. Conclusion

1. Compare the models, how many and which independent variables were selected?

When using the **Cross-Validation (10-Fold)** test mode:

- 192 instances
- The most optimal **linear regression** model would be the "No Selection" method. This is because the Correlation Coefficient (-1 to +1) is highest at (0.8358) and the errors are the lowest overall.
- The less optimal **linear regression model** would be the "Greedy" method because the Correlation Coefficient (-1 to +1) is lower at (0.8261) and the overall errors are the higher.
- The least optimal model would be the "Decision Stump" method because the Correlation Coefficient (-1 to +1) is lowest at (0.6746) and the errors are by far the highest.

When using the **User-Supplied Test Set** mode:

- 96 instances
- The most optimal **linear regression** model would be the "Greedy" method. This is because even though the errors are higher for Mean Absolute Error and Relative Absolute Error the Correlation Coefficient (-1 to +1) is highest at (0.8710) meaning the strength between variables is the overall strongest.
- The less optimal **linear regression models** would be the "No Selection and M5" methods because the Correlation Coefficient (-1 to +1) is lower at (0.8687).
- The least optimal model would be the "Decision Stump" method because the Correlation Coefficient (-1 to +1) is lowest at (0.7578) and the errors are by far the highest.

In conclusion, when using Cross-Validation (10-Fold) mode the best model for **linear regression** would be the "No Selection" method due to its high correlation and lower overall errors. When using the User-Supplied Test Set mode the best model for **linear regression** would be the "Greedy" method due to its highest overall correlation. The best overall model may be more difficult to decide as the Cross-Validation (10-Fold) model using the "No Selection" method has a lower correlation than the User-Supplied Test set model using the "Greedy" method by 0.0352. Although the correlation is lower, the errors for the Cross-Validation (10-Fold) model using the "No Selection" method are far lower than using the User-Supplied Test set model with the "Greedy" method making it a better candidate for the "best model".

"No Selection" and "M5" use all nine attributes while "Greedy" only uses seven (not utilizing variables from TOTOTORS and ELOC columns).

3. Raw WEKA data

Fit Prediction (Linear Regression) Cross-Validation (10-Fold) greedy:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
     -0.0436 * NUMUORS +
      0.0351 * NUMUANDS +
     -0.0024 * TOTOPANDS +
     -0.0369 * VG +
      0.2214 * NLOGIC +
      0.0017 * LOC +
     -0.3583

Time taken to build model: 0.01 seconds

=== Cross-validation ===

=== Summary ===
Correlation coefficient                  0.8261
Mean absolute error                      1.6569
Root mean squared error                  2.6824
Relative absolute error                 55.2924 %
Root relative squared error             56.4984 %
Total Number of Instances              192
```

Fit Prediction (Linear Regression) Supplied test set (Test_pred) greedy:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
    -0.0436 * NUMUORS +
     0.0351 * NUMUANDS +
    -0.0024 * TOTOPANDS +
    -0.0369 * VG +
     0.2214 * NLOGIC +
     0.0017 * LOC +
    -0.3583

Time taken to build model: 0 seconds

=== Evaluation on test set ===
Time taken to test model on supplied test set: 0 seconds

=== Summary ===
Correlation coefficient                  0.871
Mean absolute error                      1.9616
Root mean squared error                  3.9258
Relative absolute error                 56.6442 %
Root relative squared error             59.2485 %
Total Number of Instances               96
```

Fit Prediction (Linear Regression) Cross-Validation (10-Fold) M5:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
     -0.054  * NUMUORS +
      0.0366 * NUMUANDS +
      0.0031 * TOTOTORS +
     -0.0062 * TOTOPANDS +
     -0.0441 * VG +
      0.2162 * NLOGIC +
      0.0013 * LOC +
      0.0051 * ELOC +
     -0.3001

Time taken to build model: 0.04 seconds

=== Cross-validation ===

=== Summary ===
Correlation coefficient                  0.8266
Mean absolute error                      1.6459
Root mean squared error                  2.6873
Relative absolute error                 54.9255 %
Root relative squared error             56.6001 %
Total Number of Instances              192
```

Fit Prediction (Linear Regression) Supplied test set (Test_pred) M5:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
     -0.054  * NUMUORS +
      0.0366 * NUMUANDS +
      0.0031 * TOTOTORS +
     -0.0062 * TOTOPANDS +
     -0.0441 * VG +
      0.2162 * NLOGIC +
      0.0013 * LOC +
      0.0051 * ELOC +
     -0.3001

Time taken to build model: 0 seconds

=== Evaluation on test set ===
Time taken to test model on supplied test set: 0 seconds

=== Summary ===
Correlation coefficient                  0.8687
Mean absolute error                      1.9577
Root mean squared error                  3.957
Relative absolute error                 56.532  %
Root relative squared error             59.719  %
Total Number of Instances               96
```

Fit Prediction (Linear Regression) Cross-Validation (10-Fold) No Selection:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
     -0.054  * NUMUORS +
      0.0366 * NUMUANDS +
      0.0031 * TOTOTORS +
     -0.0062 * TOTOPANDS +
     -0.0441 * VG +
      0.2162 * NLOGIC +
      0.0013 * LOC +
      0.0051 * ELOC +
     -0.3001

Time taken to build model: 0 seconds

=== Cross-validation ===

=== Summary ===
Correlation coefficient                  0.8358
Mean absolute error                      1.6177
Root mean squared error                  2.6141
Relative absolute error                 53.9848 %
Root relative squared error             55.0596 %
Total Number of Instances              192
```

Fit Prediction (Linear Regression) Supplied test set (Test_pred) No Selection:

```
=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4
Relation:     TudorFitRegression
Instances:    192
Attributes:   9
              NUMUORS
              NUMUANDS
              TOTOTORS
              TOTOPANDS
              VG
              NLOGIC
              LOC
              ELOC
              FAULTS

Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

Linear Regression Model

FAULTS =
     -0.054  * NUMUORS +
      0.0366 * NUMUANDS +
      0.0031 * TOTOTORS +
     -0.0062 * TOTOPANDS +
     -0.0441 * VG +
      0.2162 * NLOGIC +
      0.0013 * LOC +
      0.0051 * ELOC +
     -0.3001

Time taken to build model: 0 seconds

=== Evaluation on test set ===
Time taken to test model on supplied test set: 0 seconds

=== Summary ===
Correlation coefficient                  0.8687
Mean absolute error                      1.9577
Root mean squared error                  3.957
Relative absolute error                 56.532  %
Root relative squared error             59.719  %
Total Number of Instances               96
```

Fit Prediction (Decision Stump) Cross-Validation (10-Fold) Default:

=== Run information ===

Scheme: weka.classifiers.trees.DecisionStump
Relation: TudorFitRegression
Instances: 192
Attributes: 9
NUMUORS
NUMUANDS
TOTOTORS
TOTOPANDS
VG
NLOGIC
LOC
ELOC
FAULTS

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===
Decision Stump

Classifications
NLOGIC <= 14.0 : 1.3854748603351956
NLOGIC > 14.0 : 15.615384615384615
NLOGIC is missing : 2.3489583333333335

Time taken to build model: 0 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.6746
Mean absolute error	2.1841
Root mean squared error	3.5237
Relative absolute error	72.8887 %
Root relative squared error	74.2175 %
Total Number of Instances	192

Fit Prediction (Decision Stump) Supplied test set (Test_pred) Default:

=== Run information ===

Scheme: weka.classifiers.trees.DecisionStump
Relation: TudorFitRegression
Instances: 192
Attributes: 9
NUMUORS
NUMUANDS
TOTOTORS
TOTOPANDS
VG
NLOGIC
LOC
ELOC
FAULTS

Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===
Decision Stump

Classifications
NLOGIC <= 14.0 : 1.3854748603351956
NLOGIC > 14.0 : 15.615384615384615
NLOGIC is missing : 2.3489583333333335

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correlation coefficient	0.7578
Mean absolute error	2.4427
Root mean squared error	4.5254
Relative absolute error	70.5379 %
Root relative squared error	68.2982 %
Total Number of Instances	96