CAP6673 – Data Mining and Machine Learning

Data Mining and Machine Learning Assignment 1

Written by:

Christopher Foley
Z15092976

Academic Year: Spring 2017

Table of Contents

Assignment 1A: format data files

Create Data Set 1A (Prediction)

The original data sets were reformatted to add comments and conform to the arff structure as demonstrated in class the resulting files were as follows:

- TEST.arff (excerpt)

```
% CCCS - TEST  data set for Data mining classes
% Author: Christopher Foley
% email: cfoley3@fau.edu
% 09 attributes



@relation TEST


@attribute NUMUORS real
@attribute NUMUANDS real
@attribute TOTOTORS real
@attribute TOTOPANDS real
@attribute VG real
@attribute NLOGIC real
@attribute LOC real
@attribute ELOC real
@attribute FAULTS real
```

```
@data
6,12,127,45,10,0,641,55,0
5,5,41,12,1,0,407,17,0
23,28,95,66,4,2,241,20,0
```

- FIT.arff (excerpt):

```
% CCCS - FIT  data set for Data mining classes
% Author: Christopher Foley
% email: cfoley3@fau.edu
% 09 attributes


@relation FIT

@attribute NUMUORS real
@attribute NUMUANDS real
@attribute TOTOTORS real
@attribute TOTOPANDS real
@attribute VG real
@attribute NLOGIC real
@attribute LOC real
@attribute ELOC real
@attribute FAULTS real
```

```
@data
22,85,203,174,9,0,362,40,0
21,87,186,165,5,0,379,32,0
```

Create Data Set 2 (Prediction):

The data sets used in data set 1 were additionally formatted in CSV and a MACRO was used to add a 10^{th} data field (FaultProne). As per directions modules containing 2 or more FAULTS were classified as fp, all others nfp. The resulting excerpted data files were as follows:

- TEST2.arff

```
% CCCS - TEST  data set for Data mining classes
% used for prediction
% Author: Christopher Foley
% email: cfoley3@fau.edu
% 10 attributes



@relation TEST


@attribute NUMUORS real
@attribute NUMUANDS real
@attribute TOTOTORS real
@attribute TOTOPANDS real
@attribute VG real
@attribute NLOGIC real
@attribute LOC real
@attribute ELOC real
```

```
@attribute FAULTS real
@attribute FAULTPRONE {nfp,fp}

@data
6,12,127,45,10,0,641,55,0,nfp
5,5,41,12,1,0,407,17,0,nfp
```

- FIT2.arff

```
% CCCS - FIT  data set for Data mining classes
% classification assignment
% Author: Christopher Foley
% email: cfoley3@fau.edu
% 10 attributes


@relation FIT

@attribute NUMUORS real
@attribute NUMUANDS real
@attribute TOTOTORS real
@attribute TOTOPANDS real
@attribute VG real
@attribute NLOGIC real
@attribute LOC real
@attribute ELOC real
```

```
@attribute FAULTS real
@attribute FAULTPRONE {nfp, fp}

@data
22,85,203,174,9,0,362,40,0,nfp
21,87,186,165,5,0,379,32,0,nfp
30,107,405,306,25,0,756,99,0,nfp
```

Assignment 1B: Prediction

Data

The data provided appears to be a subset of the data created by a large command, control and communications system analyzed earlier [i] Using the author provided references the following product metrics were studied:

Attribute	Description
NUMORS	Number of Unique operators
NUMANDS	Number of Unique operands
TOTOTORS	Total number of operators
TOTOPANDS	Total number of operands
VG	McCabe Complexity Complex
NLOGIC	Number of logical ooperators
LOC	Number of lines of code
ELOC	Executable lines of code

A FIT data set containing 184 instances was used with the WEKA tool to train the model and create initial data. Two basic models were built, a linear regression model and a Decision Stump model.

After creating models with the FIT data, TEST data was then used and the results compared.

Analysis

Review of the data models generated will show that the total number of operands (TOTOPANDS) does not appear to have any affect on the measure of software quality as measured by the number of faults. The greatest indicator appears to be, as expected, the Number of Logical Operators (NLOGIC). This is expected as a logical operator indicates a decision or branch in the software which may be executed incorrectly or in unexpected ways. This conclusion will be demonstrated in the following sections which present the models and their data.

Linear Models

The unformatted FIT models are presented in Appendix A FIT Models and the unformatted test data in Appendix B – Test Models

Linear Models – Greedy

A greedy model will create a model and remove a rule from the rule tree. If the resulting model results in a better model, the rule is removed and the next rule is reviewed. When all rules have been examined, duplicates are removed and the resulting model used. The FIT model produced the folllowing model:

```
FAULTS =

      -0.0517 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0026 * TOTOTORS +
      -0      * TOTOPANDS +
      -0.0372 * VG +
       0.2118 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.309
```

Of particular note is that the TOTOPANDS attribute is given zero weight indicating that a reasonable model may be built ignoring the total number of operands. Applying the FIT model to

the test data we get:

```
FAULTS =
      -0.0482 * NUMUORS +
       0.0336 * NUMUANDS +
      -0.0021 * TOTOTORS +
      -0.0337 * VG +
       0.2088 * NLOGIC +
       0.0019 * LOC +
      -0.3255
```

Again of note the TOTOPANDS attribute is not present indicating that it does not affect the calculations however the TEST data also excludes the ELOC attribute. NLOGIC appears to have a much higher weight than others. Comparing the coefficients we see the following:

Attribute	FIT coefficient	TEST coefficient
NUMORS	-0.0517	-0.0482
NUMANDS	-0.0341	0.0336
TOTOTORS	-0.0026	-0.0021
TOTOPANDS	-0	n/a
VG	-0.0372	-0.0337
NLOGIC	0.2118	0.2088
LOC	0.0018	0.0019
ELOC	0.005	n/a
vertical shiift	-0.0309	-0.3225
Correlation Coefficient	0.7961	0.8314

Note: a value of n/a indicates that the attribute was not used in the final model.

A comparison of the data indicates that the TOTOPANDS (total number of operands) apparently did not affect the outcome and it's appearance in the FIT model is probably due to a value on the interval (-0.00005,0], since values are rounded to 4 decimal places. In both models a correlation > 0.70 indicates a strong positive correlation indicating that the number of faults could be predicted from both models.

Linear Models – M5

A M5 modes combines a decision tree with the possibility of linear regression at each node. The FIT model produced the following model:

```
FAULTS =
      -0.0516 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0027 * TOTOTORS +
      -0.0372 * VG +
       0.2119 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.3091
```

Of particular note in this model is that the TOTOPANDS attribute is not used, indicating that again the model may be built ignoring the total number of operands. Applying the FIT model to the test data we get:

```
FAULTS =
      -0.0516 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0027 * TOTOTORS +
      -0.0372 * VG +
       0.2119 * NLOGIC +
```

```
      0.0018 * LOC +
      0.005  * ELOC +
     -0.3091
```

Again of note the TOTOPANDS attribute is not present indicating that it does not affect the calculations. A comparison of the coefficients shows that both models have generated the same coefficients, however the correlation coefficient of the FIT data is 0.7935 and the TEST data is 0.839 both showing strong positive correlations. For completeness we compare the identical coefficients in a table similar to that shown earlier in "Linear Models – Greedy" on pagae 8:

Attribute	FIT coefficient	TEST coefficient
NUMORS	-0.0516	-0.0516
NUMANDS	-0.0341	-0.0341
TOTOTORS	-0.0027	-0.0027
TOTOPANDS	n/a	n/a
VG	-0.0372	-0.0372
NLOGIC	0.2119	0.2119
LOC	0.0018	0.0018
ELOC	0.005	n/a
vertical shiift	-0.3091	-0.3091
Correlation Coefficient	0.7935	0.829

Note: a value of n/a indicates that the attribute was not used in the final model.

A comparison of the data indicates that the TOTOPANDS (total number of operands) apparently did not affect the outcome and it's appearance in the FIT model is probably due to a value on the interval (-0.00005,0], since values are rounded to 4 decimal places. In both models a correlation > 0.70 indicates a strong positive correlation indicating that the number of faults could be predicted

from both models.

Linear Models – No attribute selection

With no attributes selected, the FIT model produced the folllowing model:

```
FAULTS =

      -0.0517 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0026 * TOTOTORS +
      -0      * TOTOPANDS +
      -0.0372 * VG +
       0.2118 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.309
```

Again, of particular note is that the TOTOPANDS attribute is given -0 weight indicating that its value is on the interval (-0.00005, 0], indicating that a reasonable model may be built ignoring the total number of operands. Applying the FIT model to the test data we get:

```
FAULTS =
      -0.0517 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0026 * TOTOTORS +
      -0      * TOTOPANDS +
      -0.0372 * VG +
```

```
 0.2118 * NLOGIC +
 0.0018 * LOC +
 0.005  * ELOC +
-0.309
```

Again of note the TOTOPANDS attribute is not present indicating that it does not affect the calculations. We also observe that both the FIT and TEST have the same coefficients with correlations of 0.7969 and 0.829 respectively, showing however the TEST data also excludes the ELOC attribute. Comparing the coefficients we see the following:

Attribute	FIT coefficient	TEST coefficient
NUMORS	-0.0517	-0.0517
NUMANDS	-0.0341	-0.0341
TOTOTORS	-0.0026	-0.0021
TOTOPANDS	-0	-0
VG	-0.0372	-0.0372
NLOGIC	0.2118	0.2118
LOC	0.0018	0.0018
ELOC	0.005	0.005
vertical shiift	-0.0309	-0.0309
Correlation Coefficient	0.7969	0.829

Note: a value of n/a indicates that the attribute was not used in the final model.

A comparison of the data indicates that the TOTOPANDS (total number of operands) apparently did not affect the outcome and it's appearance in the FIT model is probably due to a value on the interval (-0.00005,0], since values are rounded to 4 decimal places. In both models a correlation >

0.70 indicates a strong positive correlation indicating that the number of faults could be predicted from both models.

Linear Models – Decision Stump

A Decision stump model will attempt to merge and prune the decision tree to one node. The FIT data produced the folllowing model:

```
NLOGIC <= 14.0 : 1.3806818181818181

NLOGIC > 14.0 : 15.333333333333334

NLOGIC is missing : 2.271276595744681
```

Of particular note is that the NLOGIC attribute appears to be the key to the number of faults in the modules. This is expected due to the fact that logical operators indicate decisions which are usually the cause of faults in a program. Applying the FIT model to the test data we get similar results:

```
NLOGIC <= 14.0 : 1.3806818181818181

NLOGIC > 14.0 : 15.333333333333334

NLOGIC is missing : 2.271276595744681
```

A comparison of the WEKA analysis indicated that the larger data set was less reliable with a correlation coefficient of 0.5941. The smaller test data had a correlation of 0.7111 indicating a positive correlation.

Data Summary

The tabular data may be summarized as follows:

Attribute	Linear Regression					
	Greedy		M5		None	
	FIT	TEST	FIT	TEST	FIT	TEST
NUMORS	-0.0517	-0.0482	-0.0516	-0.0516	-0.0517	-0.0517
NUMANDS	-0.0341	0.0336	-0.0341	-0.0341	-0.0341	-0.0341
TOTOTORS	-0.0026	-0.0021	-0.0027	-0.0027	-0.0026	-0.0021
TOTOPANDS	-0	n/a	n/a	n/a	-0	-0
VG	-0.0372	-0.0337	-0.0372	-0.0372	-0.0372	-0.0372
NLOGIC	0.2118	0.2088	0.2119	0.2119	0.2118	0.2118
LOC	0.0018	0.0019	0.0018	0.0018	0.0018	0.0018
ELOC	0.005	n/a	0.005	n/a	0.005	0.005
Vertical Shift	-0.0309	-0.3225	-0.3091	-0.3091	-0.0309	-0.0309
Correlation	0.7961	0.8314	0.7935	0.829	0.7969	0.829

,

All Linear algorithms appeared to converge to the same values with reasonably high degrees of correlation. As expected the TOTOPANDS have no affect on the outcome and as shown in the Decision Stump analysis, the number of logical operands (NLOGIC) appear to be the most significant indicator of quality.

Appendix A – WEKA Output

FIT data – Linear Regression - Greedy

```
=== Run information ===


Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R
1.0E-8 -num-decimal-places 4
Relation:     FIT
Instances:    188
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 =
```

```
      -0.0482 * NUMUORS +
       0.0336 * NUMUANDS +
      -0.0021 * TOTOTORS +
      -0.0337 * VG +
       0.2088 * NLOGIC +
       0.0019 * LOC +
      -0.3255

Time taken to build model: 0.06 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7961
Mean absolute error                      1.6939
Root mean squared error                  2.8425
Relative absolute error                 58.6027 %
Root relative squared error             60.9977 %
Total Number of Instances              188
```

FIT data – Linear Regression – M5

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R
1.0E-8 -num-decimal-places 4
```

```
Relation:     FIT
Instances:    188
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 =

      -0.0516 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0027 * TOTOTORS +
      -0.0372 * VG +
       0.2119 * NLOGIC +
```

```
      0.0018 * LOC +
      0.005  * ELOC +
     -0.3091

Time taken to build model: 0.22 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7935
Mean absolute error                      1.7017
Root mean squared error                  2.8612
Relative absolute error                 58.8734 %
Root relative squared error             61.3972 %
Total Number of Instances              188
```

FIT data – Linear Regression – None

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R
1.0E-8 -num-decimal-places 4
Relation:     FIT
Instances:    188
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 =

      -0.0517 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0026 * TOTOTORS +
      -0      * TOTOPANDS +
      -0.0372 * VG +
       0.2118 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.309
```

```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7969
Mean absolute error                      1.6902
Root mean squared error                  2.8362
Relative absolute error                 58.4755 %
Root relative squared error             60.8616 %
Total Number of Instances              188
```

FIT Data – Decision Stump

```
=== Run information ===

Scheme:       weka.classifiers.trees.DecisionStump
Relation:     FIT
Instances:    188
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 : 1.3806818181818181
NLOGIC > 14.0 : 15.333333333333334
NLOGIC is missing : 2.271276595744681


Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5941
Mean absolute error                      2.2173
Root mean squared error                  3.7905
Relative absolute error                 76.7091 %
```

```
Root relative squared error             81.3406 %
Total Number of Instances              188
```

TEST Data Linear Regression - Greedy

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 2 -R
1.0E-8 -num-decimal-places 4
Relation:     FIT
Instances:    188
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 =

      -0.0482 * NUMUORS +
       0.0336 * NUMUANDS +
      -0.0021 * TOTOTORS +
      -0.0337 * VG +
       0.2088 * NLOGIC +
       0.0019 * LOC +
      -0.3255

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correlation coefficient                  0.8314
Mean absolute error                      1.8383
Root mean squared error                  3.6895
Relative absolute error                 58.6625 %
Root relative squared error             62.968  %
Total Number of Instances               94
```

TEST Data Linear Regression – M5

```
=== Run information ===


Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R
1.0E-8 -num-decimal-places 4
Relation:     FIT
Instances:    188
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 =
```

```
      -0.0516 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0027 * TOTOTORS +
      -0.0372 * VG +
       0.2119 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.3091

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                  0.829
Mean absolute error                      1.8376
Root mean squared error                  3.7324
Relative absolute error                 58.6423 %
Root relative squared error             63.7    %
Total Number of Instances               94
```

TEST Data Linear Regression - None

```
=== Run information ===


Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R
1.0E-8 -num-decimal-places 4
Relation:     FIT
Instances:    188
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 =
```

```
      -0.0517 * NUMUORS +
       0.0341 * NUMUANDS +
      -0.0026 * TOTOTORS +
      -0      * TOTOPANDS +
      -0.0372 * VG +
       0.2118 * NLOGIC +
       0.0018 * LOC +
       0.005  * ELOC +
      -0.309

Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correlation coefficient                  0.829
Mean absolute error                      1.8377
Root mean squared error                  3.7317
Relative absolute error                 58.6426 %
Root relative squared error             63.6881 %
Total Number of Instances               94
```

TEST Data Decision Stump

```
=== Run information ===


Scheme:       weka.classifiers.trees.DecisionStump
Relation:     FIT
Instances:    188
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 : 1.3806818181818181
NLOGIC > 14.0 : 15.333333333333334
```

```
NLOGIC is missing : 2.271276595744681


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

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

=== Summary ===

Correlation coefficient                  0.7111
Mean absolute error                      2.2952
Root mean squared error                  4.1928
Relative absolute error                 73.2439 %
Root relative squared error             71.5571 %
Total Number of Instances               94
```

i T.M. Khoshgoftaar and E.B. Allen, *Modeling Software Quality: The* Software *Measurement Analysis and Reliability Toolkit*, copy provided by author T.M. Khoshgoftaar as part of class notes CAP6674: Machine Learning and Data Mining, Florida Atlantic University, Spring 2017.