Assignment 4

Submission Date: Thursday, October 4, 2012

Learning Objectives:

1. Gain insight into the concept of confidence intervals
2. Learn how to use Weka's Remove Folds filter and the Experimenter

(1) We used the SpeakerIDforLect data set for this assignment.

The 5-fold info is as follows.
Fold 1
Success rate: 19.8094%
Number of instances: 3988

Fold 2
Success rate: 22.1721%
Number of instances: 3987

Fold 3
Success rate: 23.3258%
Number of instances: 3987

Fold 4
Success rate: 27.4893%
Number of instances: 3987

Fold 5
Success rate: 23.702%
Number of instances: 3987

We compute the confidence interval on fold 1, with Z = 1.65, we have
Confidence interval: [0.1878878, 0.2087121]

With doubled data set, the confidence interval will be smaller, because with more data, we are more certain about the point estimate.

(2-3) For a multiple classification problem like this one, an instance could be incorrectly classified into any class other than the actual class label. For each actual class value in this question, we focus on the two cells in the confusion matrix with the most errors, because that is the main source of errors.

For each value of the class label, we list the features that possibly caused misclassification, because the average values of those features are very different between the two cells with most errors and other cells in the confusion matrix.

For class value irf03, feature 13, 14

For class value irf04, feature 1

For class value irf06, feature 16, 18, 19, 17, 23

For class value irf07, feature 23, 17, 20, 19, 18, 16

For class value irm02, feature 19, 20, 21, 23

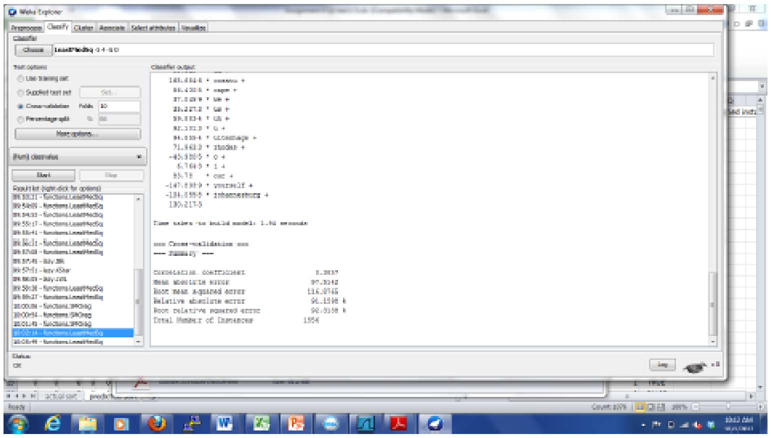
For class value irm05, feature 19

For class value irm06, no apparent indicator since most of the instances were classified correctly

For class value irm07, feature 20

(4) Here is one example solution:

Using the normal data set, replacement preprocessing and the LinearMedSq classifier the correlation coefficient = .3857.



Classifier: LeastMedSq -S 4 -G 0

Cross-validation Folds 10

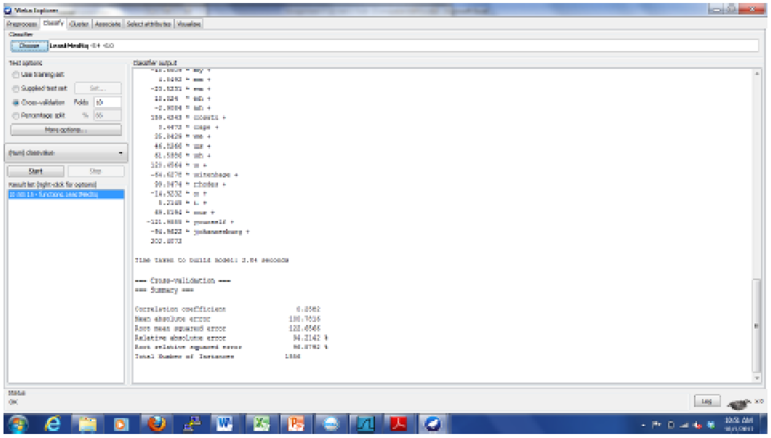
(Num) classvalue

Classifier output

```
      185.4848 * cosmos +
       86.4305 * cage +
       37.0899 * ue +
       20.2272 * ud +
       59.0034 * ub +
       82.1513 * u +
       94.0384 * citeshape +
       71.9422 * rhodes +
      -40.9205 * o +
        6.7649 * i +
       55.75   * our +
     -147.8939 * yourself +
     -134.0595 * johannesburg +
      130.2173

Time taken to build model: 1.96 seconds

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

Correlation coefficient                  0.3857
Mean absolute error                     97.3142
Root mean squared error                116.0746
Relative absolute error                 94.1598 %
Root relative squared error             92.3158 %
Total Number of Instances             1556
```

Status
OK

Simple data set



Classifier: LeastMedSq -S 4 -G 0

Cross-validation Folds 10

(Num) classvalue

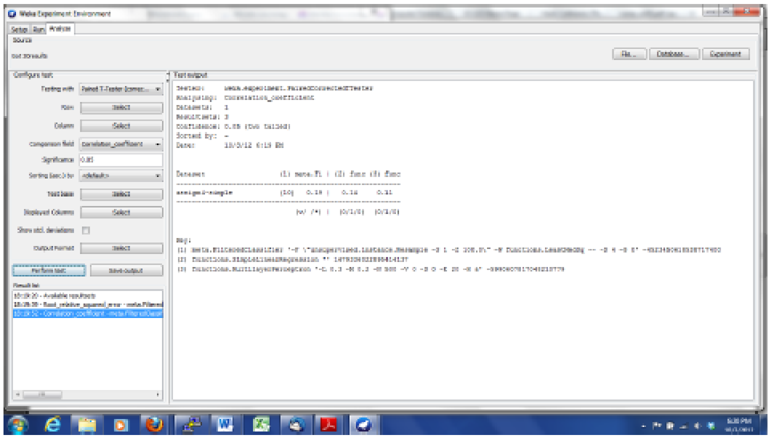
Classifier output

```
      -13.0808 * say +
        1.8482 * mm +
      -23.8232 * mw +
       13.024  * mb +
       -2.9094 * mb +
      158.4843 * coasts +
        3.4473 * cage +
       35.8429 * ww +
       46.3266 * ms +
       61.5856 * wb +
      123.4564 * m +
      -64.6278 * citeshape +
       39.3474 * rhodes +
      -14.5232 * n +
        8.2168 * i +
       49.0184 * our +
     -121.9888 * yourself +
      -92.9622 * johannesburg +
      202.0873

Time taken to build model: 2.84 seconds

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

Correlation coefficient                  0.2562
Mean absolute error                    100.7616
Root mean squared error                122.6546
Relative absolute error                 94.2142 %
Root relative squared error             96.8782 %
Total Number of Instances             1556
```

Status
OK

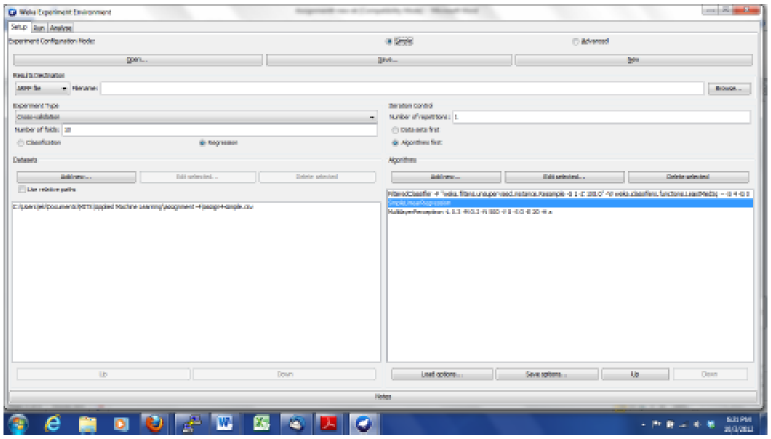
(1) Is the algorithm you picked significantly different from SimpleLinearRegression on the simple version of the dataset? The LinearMedSq algorithm does not perform significantly different from the SimpleLinearRegression algorithm. This is because the performance values fall within the confidence interval of the baseline. (2) And for the algorithm you picked, does its performance differ significantly between the two versions of the dataset? The performance on the simple data set performs more poorly.

These pages contain the comparison to two other algorithms. These results do not indicate a significant improvement over the Multilayer preceptor or the SimpleLinearRegression algorithm.



Testing with: Paired T-Tester (correc...
Row: Select
Column: Select
Comparison field: Correlation_coefficient
Significance: 0.05
Sorting (asc.) by: <default>
Test base: Select
Displayed Columns: Select
Show std. deviations: ☐
Output Format: Select

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Correlation_coefficient
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/3/12 6:13 PM


Dataset                   (1) meta.Fi | (2) func (3) func
------------------------------------------------------------
assign2-simple            (10)   0.19 |   0.16     0.21
------------------------------------------------------------
                              (v/ /*) |  (0/1/0)  (0/1/0)


Key:
(1) meta.FilteredClassifier '-F \"unsupervised.instance.Resample -S 1 -Z 100.0\" -W functions.LeastMedSq -- -S 4 -G 0' -4523450618538717400
(2) functions.SimpleLinearRegression '' 1679336022895414137
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
```



Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination: ARFF file Filename:

Experiment Type: Cross-validation

Number of folds: 10

○ Classification ● Regression

Iteration Control: Number of repetitions: 1

○ Data sets first

● Algorithms first

Datasets: □ Use relative paths

C:\Users\jel\Documents\MIT\Applied Machine Learning\assignment 4\design4-simple.csv

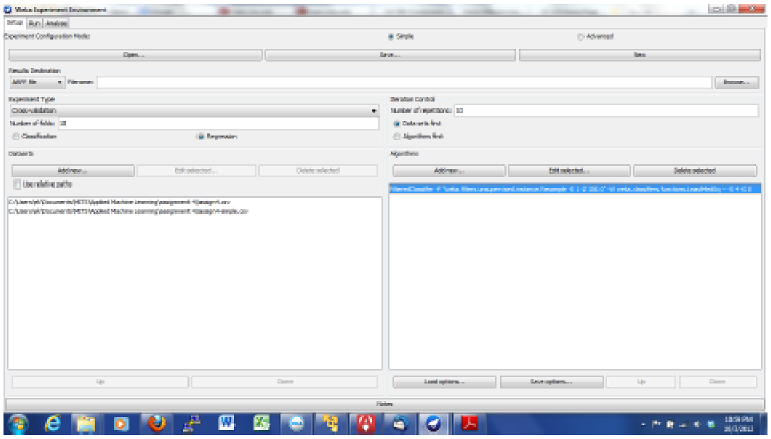
Algorithms:

FilteredClassifier -F "weka.filters.unsupervised.instance.Resample -S 1 -Z 100.0" -W weka.classifiers.functions.LeastMedSq -- -S 4 -G 0

SimpleLinearRegression

MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

The results from the comparison between the two data sets does not show significant improvement between the two data sets since the confidence intervals overlap by .01.



Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination: ARFF file Filename:

Experiment Type: Cross-validation

Number of folds: 10

○ Classification ● Regression

Iteration Control: Number of repetitions: 10

● Data sets first

○ Algorithms first

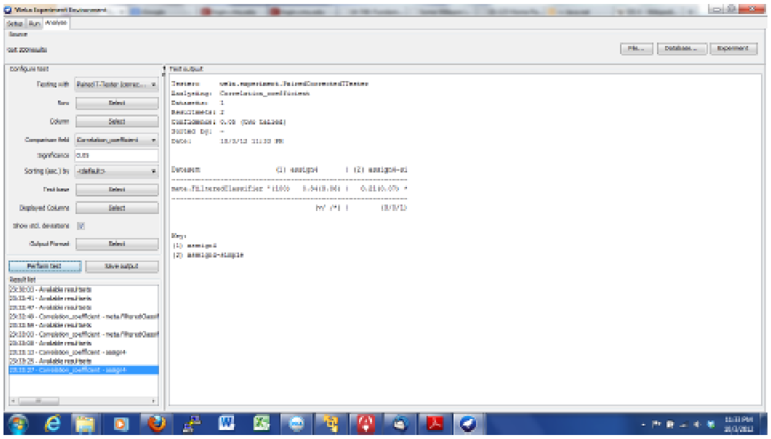
Datasets: □ Use relative paths

C:\Users\jel\Documents\MIT\Applied Machine Learning\assignment 4\design4.csv

C:\Users\jel\Documents\MIT\Applied Machine Learning\assignment 4\design4-simple.csv

Algorithms:

FilteredClassifier -F "weka.filters.unsupervised.instance.Resample -S 1 -Z 100.0" -W weka.classifiers.functions.LeastMedSq -- -S 4 -G 0



Testing with: Paired T-Tester (correc...

Row: Select

Column: Select

Comparison field: Correlation_coefficient

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations: ☑

Output Format: Select

```
[illegible] - Available resultsets
[illegible] - Available resultsets
[illegible] - Available resultsets
[illegible] - Correlation_coefficient - meta.FilteredClassif
[illegible] - Available resultsets
[illegible] - Correlation_coefficient - meta.FilteredClassif
[illegible] - Available resultsets
[illegible] - Correlation_coefficient - assign4
[illegible] - Available resultsets
[illegible] - Correlation_coefficient - assign4
```

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Correlation_coefficient
Datasets:   1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       10/3/12 11:30 PM


Dataset                   (1) assign4      | (2) assign4-si
------------------------------------------------------------
meta.FilteredClassifier '(100)   0.54(0.06) |   0.21(0.07) *
------------------------------------------------------------
                                   (v/ /*) |        (0/0/1)


Key:
(1) assign4
(2) assign4-simple
```