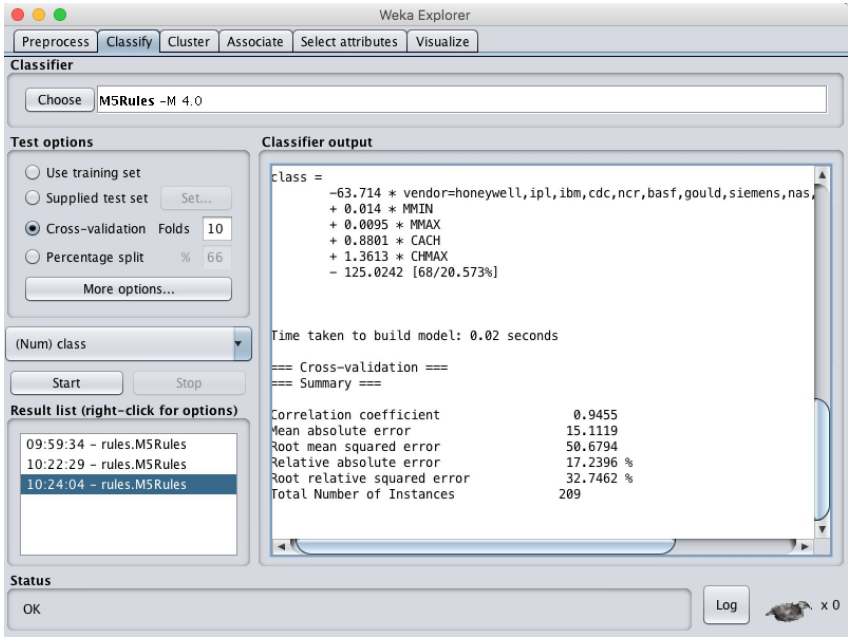
Fundamentals of Data Mining

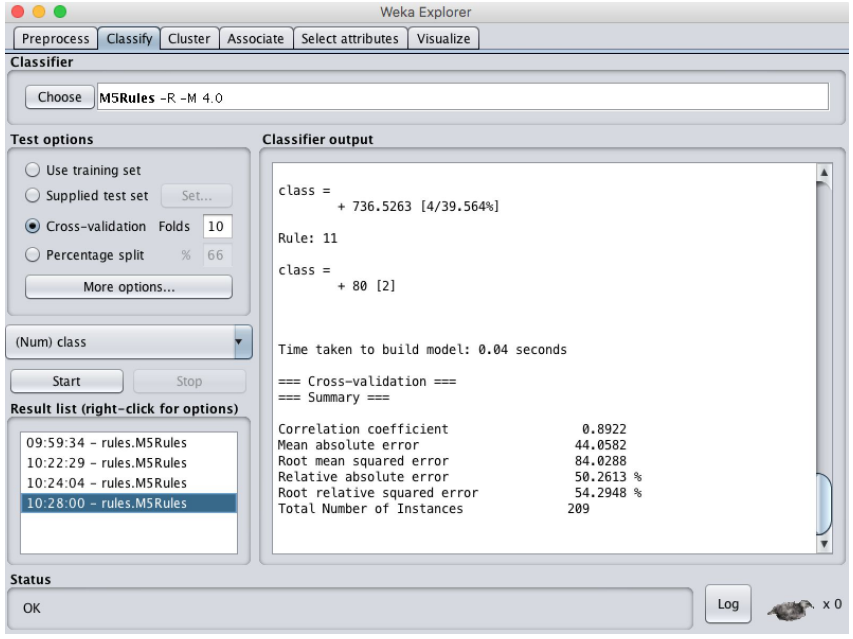
John M. Warlop, Assignment #5, Due 11/13/2017

1	Describe the similarities and difference between Decision and Regression Tree learning. Decision trees work best for nominal values whereas a regression tree works better on attributes and classes that are real valued. In addition, the way the split points(branches of tree) are derived differs. The Regression tree uses sum of squared errors to determine split point and entropy is used in a decision tree. The main similarity is that they are both tree structures.
2	Use the Regression tree learning scheme (weka.classifiers.M5') to analyze the CPU.arff. Evaluate the difference between a model tree and a regression tree (right click on the option provides "build Regression Tree" option). Experiment with the available parameters to understand their significance and discuss how they influence the model? Using M5 on the cpu.arff file I found out the following. With default values, with a regression tree, the correlation coefficient was 0.89(11 rules) and the correlation w/o a regression tree was .94(2 rules). Both of these values show a strong correction either with or without a regression tree. When I chose no regression tree and unpruned, I got a .98 correlation coefficient and 20 rules. Whereas with a regression tree I got 0.77 correlation coefficient and 20 rules. When prunning is turned on, there are much fewer rules. This makes since because prunning of tree will mean less nodes and thus less rules.

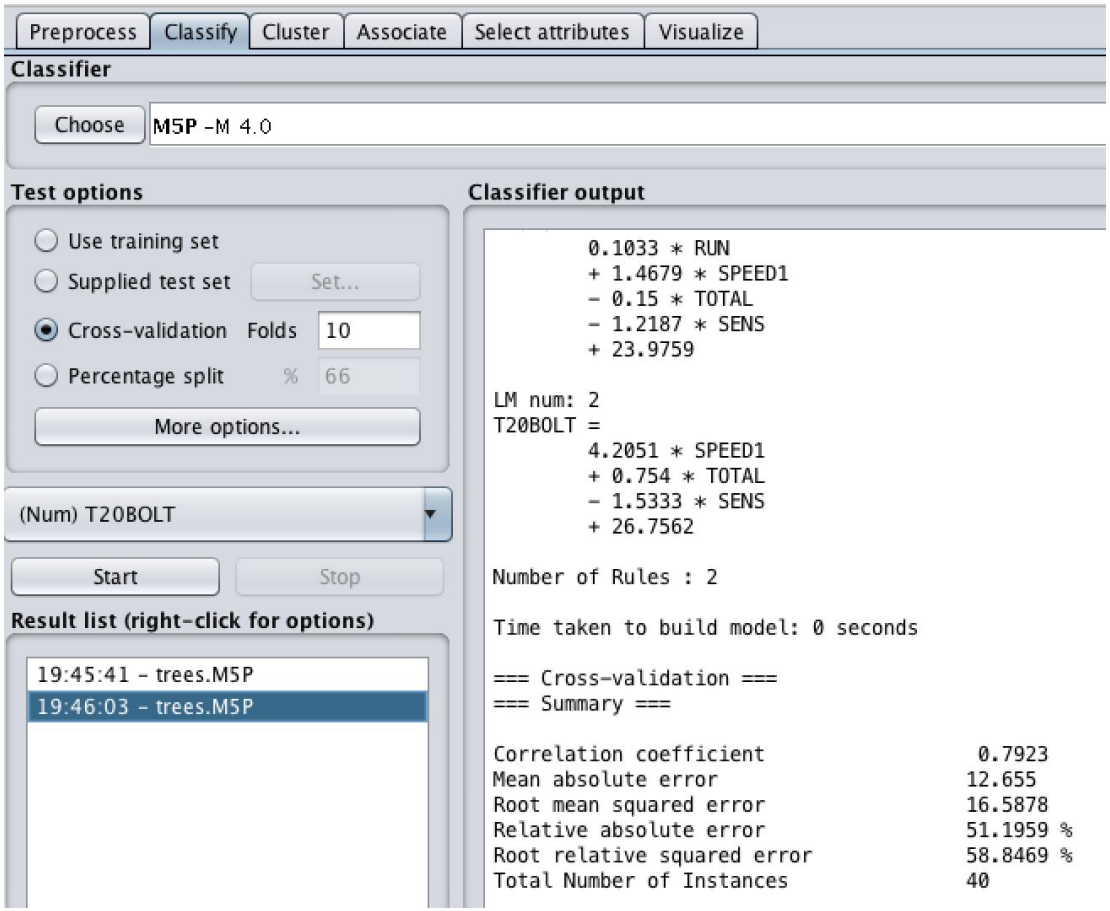
No Regression Tree



With Regression Tree

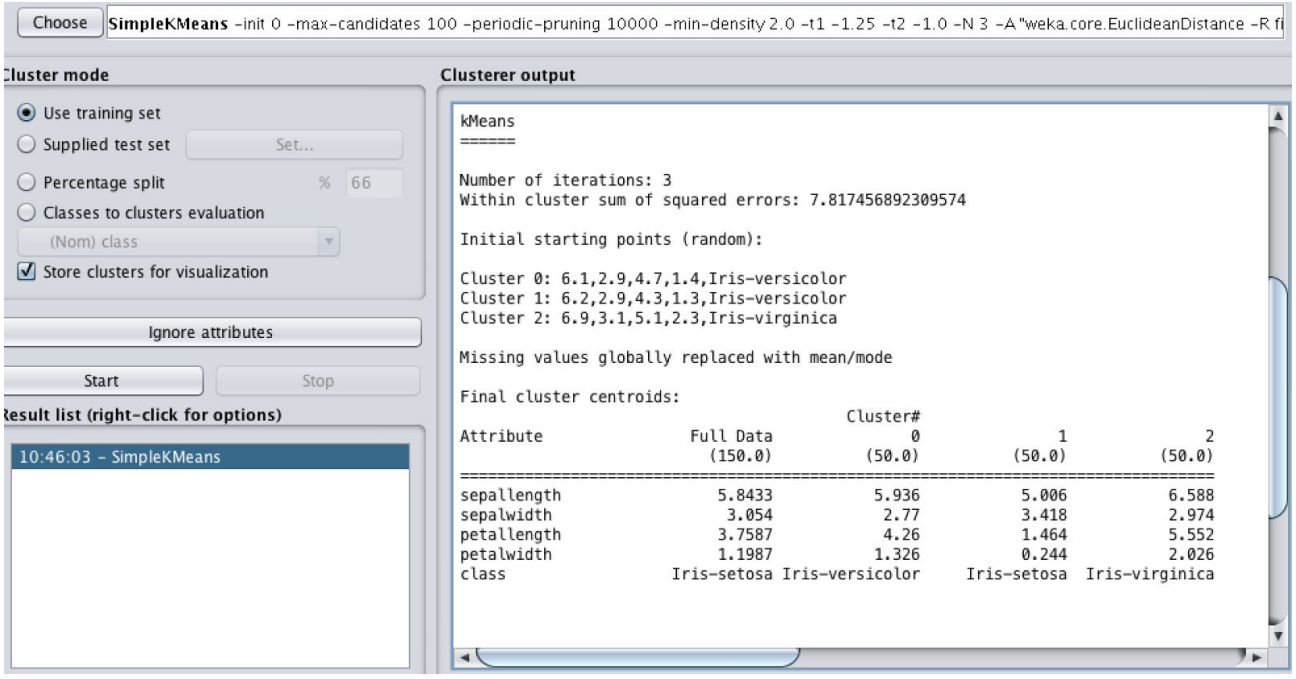


3 Use M5P Model tree learning scheme(weka.classifiers.M5') to analyze the bolts data(bolts.arff without the TIME attribute):



Analyze the data. What adjustments have the greatest effect on the time to count 20 bolts?

As I look at the two linear models, the feature with the largest coefficient is SPEED1. In the first linear model, SPEED1 has a coefficient of 1.47 and in the second it has a coefficient of 4.2. I would think the feature with the largest coeffient will lead to the biggest change in the time to count 20 bolts.

	How does this model differ from the Decision Tree induced tree? When I did the decision tree model, I also concluded that SPEED1 would be the biggest determing factor for the time to count 20 bolts.
4	Use a k-means clustering technique to analyze the iris data set. What did you set the ***k*** value to be? Try several different values. What was the random seed value? Experiment with different random seed values. How did changing of these values influence the produced model? Use different distance functions. Did they produce significantly different clustering models? I chose a k value of 3 because I knew that there are 3 clusters: Iris-versacolor, Iris-setosa and Iris-virginia. The sum of squared errors goes down as you add more clusters. When I changed to just one cluster, the cluster was Iris-vericolor. When I chose 6 clusters, there was one Iris-vertosa, two Iris-versacolors and three Iris-verginica's. This makes sense because given more clusters, the distances will decrease. I did not see any change in performance when I changed the seed value. I tried the Manhattan distance and I did not see much change in k-means performance 

Appendix

2

M5 & Regression Tree - CPU.ARFF

HOME

```
%
% !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
%
% Attributes 2 and 8 deleted.
%
% As used by Kilpatrick, D. & Cameron-Jones, M. (1998). Numeric prediction
% using instance-based learning with encoding length selection. In Progress
% in Connectionist-Based Information Systems. Singapore: Springer-Verlag.
%
% !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
%
% 1. Title: Relative CPU Performance Data
%
% 2. Source Information
%    -- Creators: Phillip Ein-Dor and Jacob Feldmesser
%      -- Ein-Dor: Faculty of Management; Tel Aviv University; Ramat-Aviv;
%         Tel Aviv, 69978; Israel
%    -- Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779
%    -- Date: October, 1987
%
% 3. Past Usage:
%     1. Ein-Dor and Feldmesser (CACM 4/87, pp 308-317)
%        -- Results:
%           -- linear regression prediction of relative cpu performance
```

```
%           -- Recorded 34% average deviation from actual values
%     2. Kibler,D. & Aha,D. (1988).  Instance-Based Prediction of
%        Real-Valued Attributes.  In Proceedings of the CSCSI (Canadian
%        AI) Conference.
%        -- Results:
%           -- instance-based prediction of relative cpu performance
%           -- similar results; no transformations required
%     - Predicted attribute: cpu relative performance (numeric)
%
% 4. Relevant Information:
%    -- The estimated relative performance values were estimated by the authors
%       using a linear regression method.  See their article (pp 308-313) for
%       more details on how the relative performance values were set.
%
% 5. Number of Instances: 209
%
% 6. Number of Attributes: 10 (6 predictive attributes, 2 non-predictive,
%                              1 goal field, and the linear regression's guess)
%
% 7. Attribute Information:
%    1. vendor name: 30
%       (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec,
%        dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson,
%        microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry,
%        sratus, wang)
%    2. Model Name: many unique symbols
%    3. MYCT: machine cycle time in nanoseconds (integer)
%    4. MMIN: minimum main memory in kilobytes (integer)
%    5. MMAX: maximum main memory in kilobytes (integer)
%    6. CACH: cache memory in kilobytes (integer)
%    7. CHMIN: minimum channels in units (integer)
%    8. CHMAX: maximum channels in units (integer)
%    9. PRP: published relative performance (integer)
```

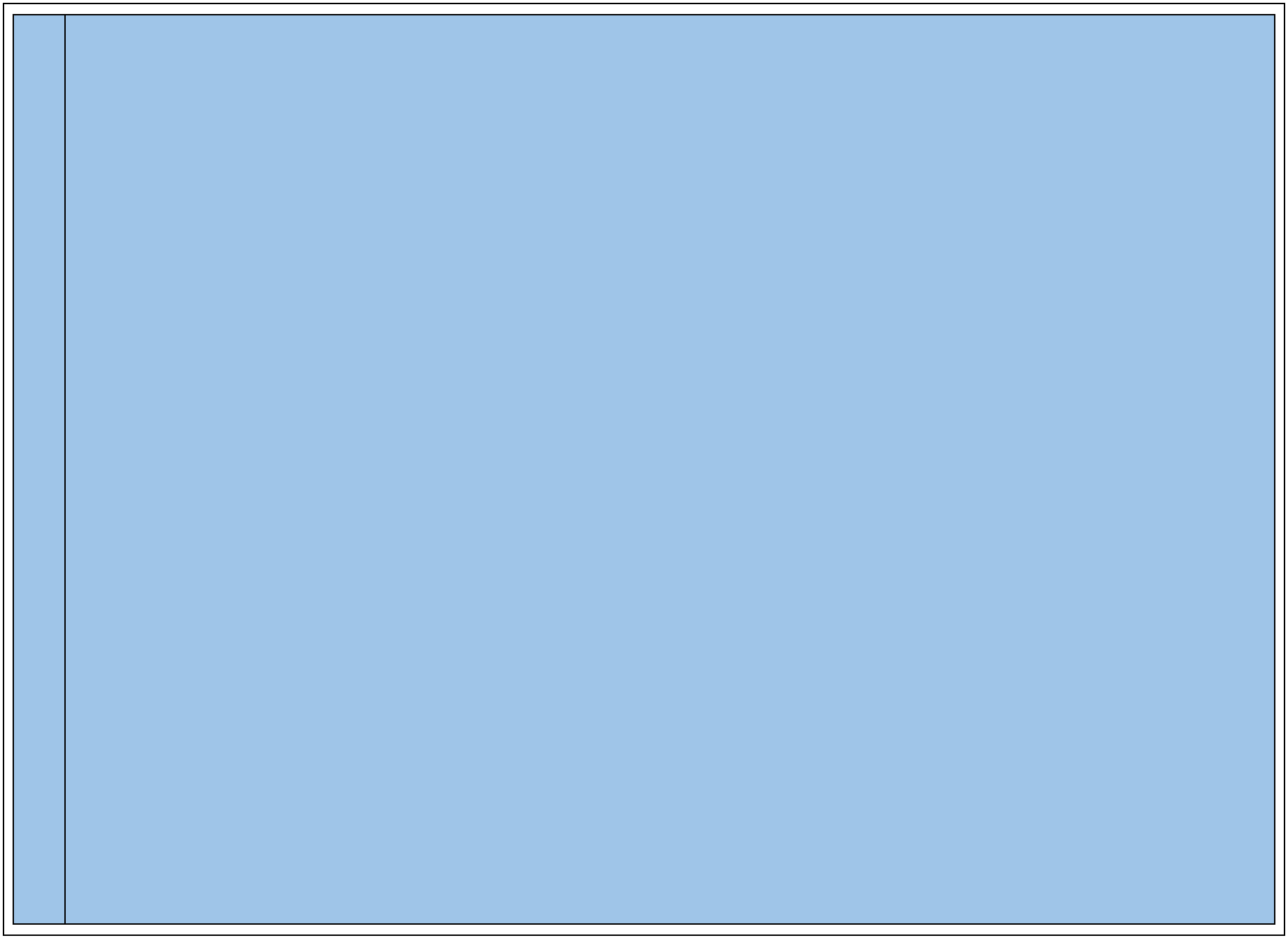
```
%   10. ERP: estimated relative performance from the original article (integer)
%
% 8. Missing Attribute Values: None
%
% 9. Class Distribution: the class value (PRP) is continuously valued.
%    PRP Value Range:   Number of Instances in Range:
%    0-20               31
%    21-100             121
%    101-200            27
%    201-300            13
%    301-400            7
%    401-500            4
%    501-600            2
%    above 600          4
%
% Summary Statistics:
%            Min  Max   Mean    SD      PRP Correlation
%    MCYT:   17   1500  203.8   260.3   -0.3071
%    MMIN:   64   32000 2868.0  3878.7   0.7949
%    MMAX:   64   64000 11796.1 11726.6  0.8630
%    CACH:   0    256   25.2    40.6     0.6626
%    CHMIN:  0    52    4.7     6.8      0.6089
%    CHMAX:  0    176   18.2    26.0     0.6052
%    PRP:    6    1150  105.6   160.8    1.0000
%    ERP:   15    1238  99.3    154.8    0.9665
%
```

M5 Rules & No Regression Tree - CPU.ARFF

HOME

```
=== Run information ===

Scheme:       weka.classifiers.rules.M5Rules -M 4.0
Relation:     cpu
Instances:    209
Attributes:   8
              vendor
              MYCT
              MMIN
              MMAX
              CACH
              CHMIN
              CHMAX
              class
Test mode:    10-fold cross-validation

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

M5 pruned model rules
(using smoothed linear models) :
Number of Rules : 2

Rule: 1
IF
      MMAX <= 14000
THEN

class =
      -2.0542 * vendor=honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl
      + 5.4303 * vendor=adviser,sperry,amdahl
      - 5.7791 * vendor=amdahl
```

```
      + 0.0064 * MYCT
      + 0.0016 * MMIN
      + 0.0034 * MMAX
      + 0.5524 * CACH
      + 1.1411 * CHMIN
      + 0.0945 * CHMAX
      + 4.1463 [141/2.365%]

Rule: 2

class =
      -63.714 * vendor=honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl
      + 0.014 * MMIN
      + 0.0095 * MMAX
      + 0.8801 * CACH
      + 1.3613 * CHMAX
      - 125.0242 [68/20.573%]



Time taken to build model: 0.02 seconds

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

Correlation coefficient                  0.9455
Mean absolute error                     15.1119
Root mean squared error                 50.6794
Relative absolute error                 17.2396 %
Root relative squared error             32.7462 %
Total Number of Instances              209
```



M5 Rules with CPU.ARFF & With Regression Tree

```
=== Run information ===

Scheme:       weka.classifiers.rules.M5Rules -R -M 4.0
Relation:     cpu
Instances:    209
Attributes:   8
              vendor
              MYCT
              MMIN
              MMAX
              CACH
              CHMIN
              CHMAX
              class
Test mode:    10-fold cross-validation

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

M5 pruned regression rules
(using smoothed linear models) :
Number of Rules : 11

Rule: 1
IF
      MMAX <= 14000
      CACH <= 8.5
      MMAX <= 6100
THEN
```

```
class =
     + 33.8382 [75/3.056%]

Rule: 2
IF
     MMAX <= 22485
     CACH <= 27
     MMAX <= 10000
THEN

class =
     + 61.2024 [38/2.917%]

Rule: 3
IF
     MMAX <= 22485
     CACH > 40
THEN

class =
     + 121.2831 [18/17.632%]

Rule: 4
IF
     MMAX <= 22485
     MMIN <= 3310
     MMAX <= 14000
THEN

class =
     + 97.5739 [16/3.53%]

Rule: 5
```

```
IF
      MMAX <= 22485
      CHMIN <= 7
THEN

class =
      + 131.4625 [24/5.422%]

Rule: 6
IF
      CACH > 56
      CHMIN > 10
      MMAX <= 48000
THEN

class =
      + 406.6518 [10/41.106%]

Rule: 7
IF
      CACH <= 56
      MMAX > 22485
      vendor=amdahl <= 0.5
THEN

class =
      + 240.3034 [8/8.062%]

Rule: 8
IF
      MMAX > 22485
      MMAX <= 48000
THEN
```

```
class =
      + 382.2619 [9/8.884%]

Rule: 9
IF
      MMAX <= 42485
      MMIN > 4620
THEN

class =
      + 328.345 [5/2.112%]

Rule: 10
IF
      vendor=basf,gould,siemens,nas,adviser,sperry,amdahl > 0.5
THEN

class =
      + 736.5263 [4/39.564%]

Rule: 11

class =
      + 80 [2]



Time taken to build model: 0.04 seconds

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

Correlation coefficient	0.8922
Mean absolute error	44.0582
Root mean squared error	84.0288
Relative absolute error	50.2613 %
Root relative squared error	54.2948 %
Total Number of Instances	209

3

M5P & Bolts.arff

HOME

```
=== Run information ===

Scheme:       weka.classifiers.trees.M5P -M 4.0
Relation:     bolts-weka.filters.unsupervised.attribute.Remove-R7
Instances:    40
Attributes:   7
              RUN
              SPEED1
              TOTAL
              SPEED2
              NUMBER2
              SENS
              T20BOLT
Test mode:    10-fold cross-validation

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

M5 pruned model tree:
(using smoothed linear models)

SPEED1 <= 5 : LM1 (24/19.755%)
SPEED1 >  5 : LM2 (16/66.007%)

LM num: 1
T20BOLT =
     0.1033 * RUN
     + 1.4679 * SPEED1
```

```
      - 0.15 * TOTAL
      - 1.2187 * SENS
      + 23.9759

LM num: 2
T20BOLT =
      4.2051 * SPEED1
      + 0.754 * TOTAL
      - 1.5333 * SENS
      + 26.7562

Number of Rules : 2

Time taken to build model: 0 seconds

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

Correlation coefficient                  0.7923
Mean absolute error                     12.655
Root mean squared error                 16.5878
Relative absolute error                 51.1959 %
Root relative squared error             58.8469 %
Total Number of Instances               40
```

4 HOME

K-Means & Iris.ARFF

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000
-min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500
-num-slots 1 -S 10
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574
```

```
Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0              1              2
                          (150.0)          (50.0)         (50.0)         (50.0)
=============================================================================
sepallength                5.8433           5.936          5.006          6.588
sepalwidth                  3.054            2.77          3.418          2.974
petallength                3.7587            4.26          1.464          5.552
petalwidth                 1.1987           1.326          0.244          2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```
