Fundamental of Data Mining
Assignment #2
Due(10/23/2017)
John M. Warlop

1. How does One-R pick the 'most important' attribute?

   The One-R classifier only uses one rule, thus the name One-R. The one rule that it ends up using is the rule that has the total smallest error. This is done by creating a frequency table for each feature/predictor.

2. Why Naïve Bayes Classification is called "naïve"?
   A synonym for naive is unsophisticated. Bayes takes an unsophisticted approach to producing a classifier. The unsophistication comes from the fact that this classifyer assumes there is no interdependencies between predictors/features. For example: if there were two features -- overcast & raining -- beyes assumes no relationship between overcast and rain.

3. We will use Weka package as the basis of the assignments for this class. If you haven't gone through the Weka Tutorial – please do so before moving forward with the assignment. Weka provides implementations of a wide range of learning procedures as well as the environment for executing systematic experiments and reporting model evaluation for the results without requiring any programming.

   These exercises serve two purposes:

   - Enable you to discover Weka capabilities and how to use them.
   - Empower you to see the learning procedures that we discuss in the lectures in action.

   Through the Weka Explorer interface, under preprocess tab load the Weather.Nominal data set.

Initially “preprocess” will have been selected as you launch the Explorer. This is the tab you select when you want Weka to find the data set that you want to use.

Weka processes data sets that are in its own ARFF format. Conveniently, the download will have set up a folder within the Weka-3.7 (or higher depending on your version of Weka) folder called “Data”. This contains a selection of data files in ARFF format. Additional files needed for the assignment can be found under Resources section in the Blackboard.

In the Explorer window, click on “Open file” and then use the browser to navigate to the ‘Data’ folder within the Weka-3.7 folder. Select the file called weather.nominal.arff. (This is in fact the file listed above).

This is an artificially created small data set, used in class for demonstration purposes. In this case, the normal usage is to learn to predict the ‘play’ attribute from four others providing information about the weather.

Most of the information it displays is self-explanatory: it is a data set containing 14 examples (instances) each of which has 5 attributes. The ‘play’ attribute has been suggested as the class attribute (i.e. the one that will be predicted from the others).

Most of the right hand of the window gives you information about the attributes. Initially, it will give you information about the first attribute (‘outlook’). This shows that it has 3 possible values tells you how many there are of each value. The bar chart in the lower right shows how the values of the suggested class variable are distributed across the possible values of the ‘outlook’.

If you click on ‘temperature’ in the panel on the left, the information about the ‘outlook’ attribute will be replaced by the corresponding information about the temperature attribute.

Choosing a method

Next select a machine learning procedure to apply to this training data set. The task is classification so click on the ‘Classify’ tab near the top of the Explorer window.

By default, a classifier called ZeroR has been selected. Click the ‘Start’ button and Weka will run the Zero-R model on the data set.

Now change the training classifier to a different classifier by clicking on the Choose button. A hierarchical pop up menu appears. Click to expand ‘Bayes’, which appears at the top of this menu, then select NaiveBayes.

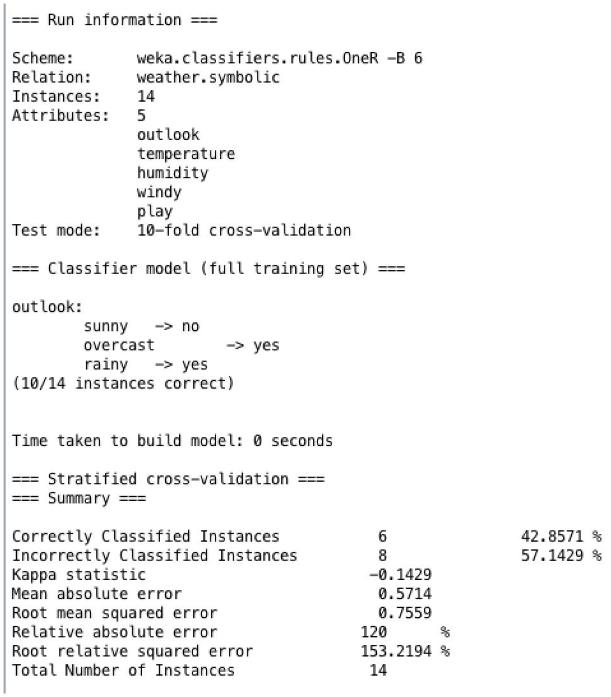
The panel headed ‘Test options’ allows the user to choose the experimental procedure. We covered this topic in more details later in the course. For the present exercise click on ‘Use training set’. (This will simply build a Naïve Bayes model using all the examples in the data set). We will use the cross-validation option for the rest of the assignments for the duration of the course.

Training the model

The small panel halfway down the left hand side indicates which attribute will be used as the classification attribute. It will currently be set to ‘play’. (Note that this is what actually determines the classification attribute – the ‘class’ attribute on the pre-process screen is simply to allow you to see how a variable appears to depend on the values of other attributes).

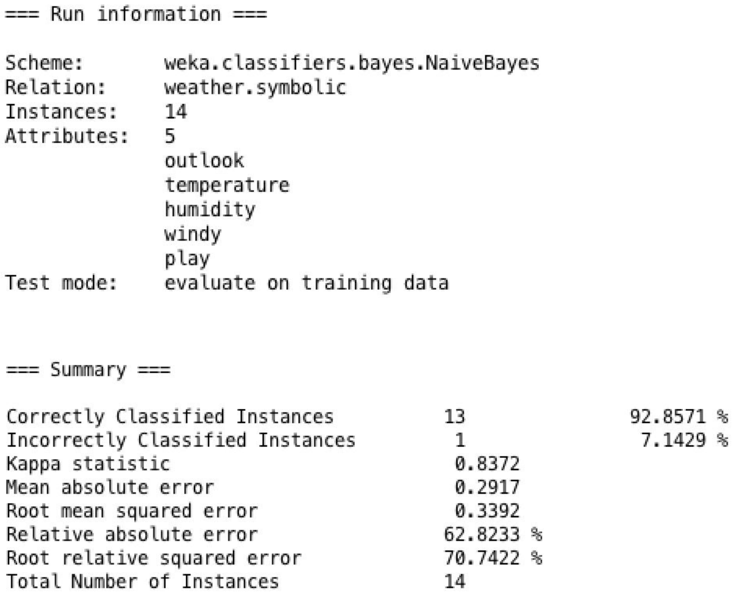
Click the start button and the Naïve Bayes model will run. The results will appear in the scrollable panel on the right of the Explorer window. Take a look at the model and evaluation of the model presented.

One-R



```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

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

outlook:
        sunny   -> no
        overcast        -> yes
        rainy   -> yes
(10/14 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.1429
Mean absolute error                      0.5714
Root mean squared error                  0.7559
Relative absolute error                120      %
Root relative squared error            153.2194 %
Total Number of Instances               14
```

Naive Beyes



```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data


=== Summary ===

Correctly Classified Instances          13               92.8571 %
Incorrectly Classified Instances         1                7.1429 %
Kappa statistic                          0.8372
Mean absolute error                      0.2917
Root mean squared error                  0.3392
Relative absolute error                 62.8233 %
Root relative squared error             70.7422 %
Total Number of Instances               14
```

What is the % of correctly classified instances when using One-R vs. Naïve Bayes?

	One-R	Naive Baynes
Correctly Classified	42.8571	92.8571

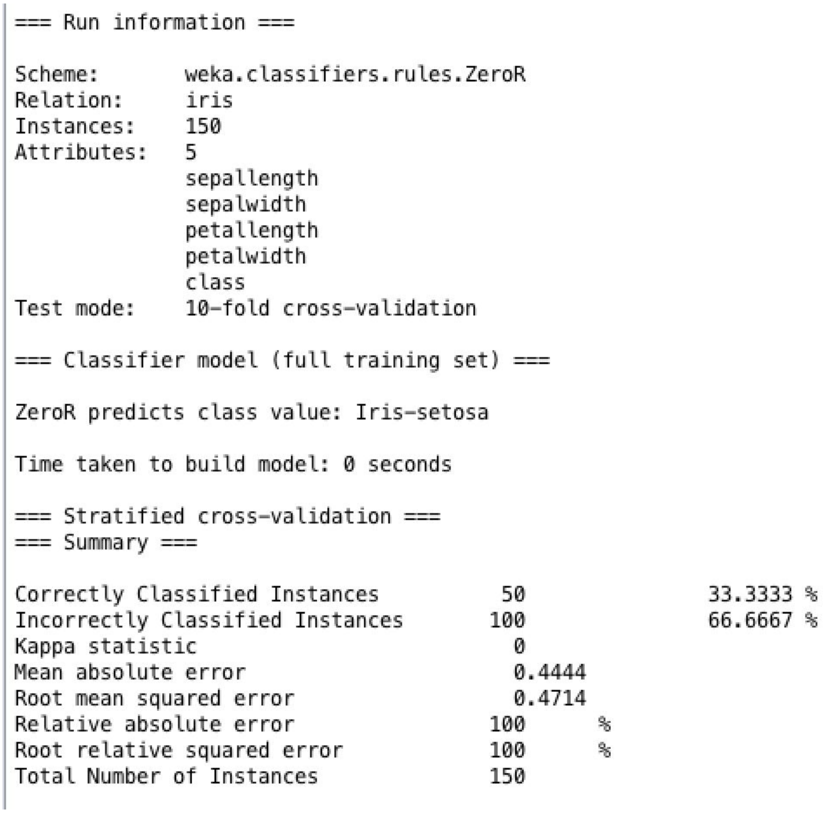
Why do you think that is? This is the percentage of instances that are classified correctly given a certain classifier.

The Naive Baynes classifier uses the conditional probabilities of each feature, whereas One-R does not. I believe this is why Naive Baynes is more accurate classifier.

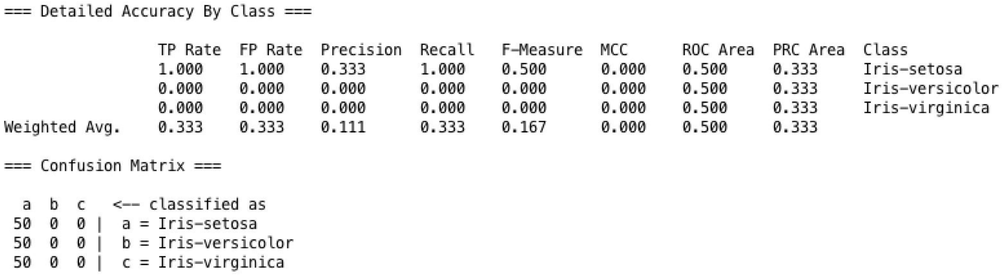
4. Use the following learning schemes to analyze the iris data (in iris.arff):

ZeroR	Weka.classifiers.ZeroR
OneR	Weka.classifiers.OneR
Naive Bayes	Weka.classifiers.NaiveBayes

############################ Zero-R ################################

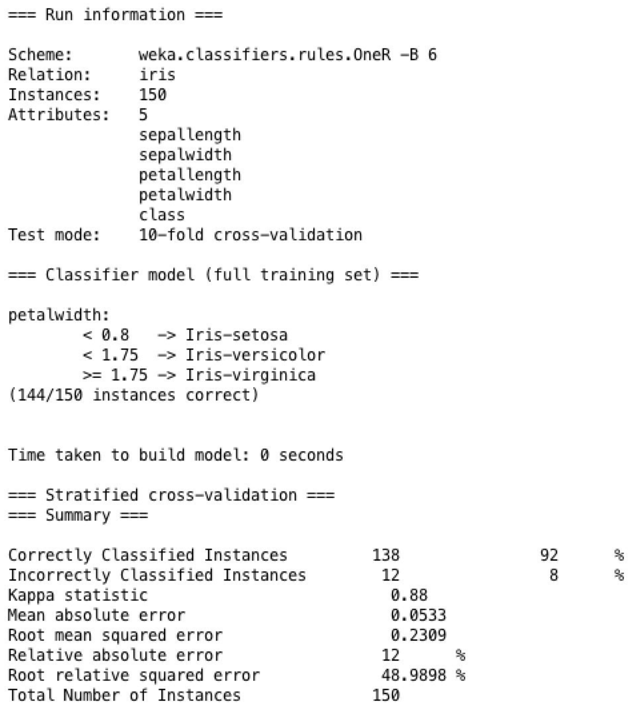


```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

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

ZeroR predicts class value: Iris-setosa

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          50               33.3333 %
Incorrectly Classified Instances       100               66.6667 %
Kappa statistic                          0
Mean absolute error                      0.4444
Root mean squared error                  0.4714
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              150
```

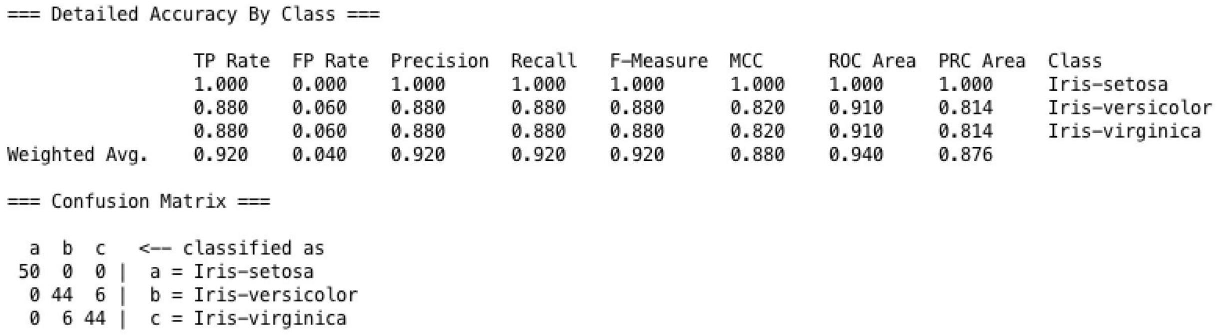


```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.333      1.000    0.500      0.000    0.500     0.333     Iris-setosa
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.333     Iris-versicolor
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.333     Iris-virginica
Weighted Avg.    0.333    0.333    0.111      0.333    0.167      0.000    0.500     0.333

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
 50  0  0 |  b = Iris-versicolor
 50  0  0 |  c = Iris-virginica
```

One-R

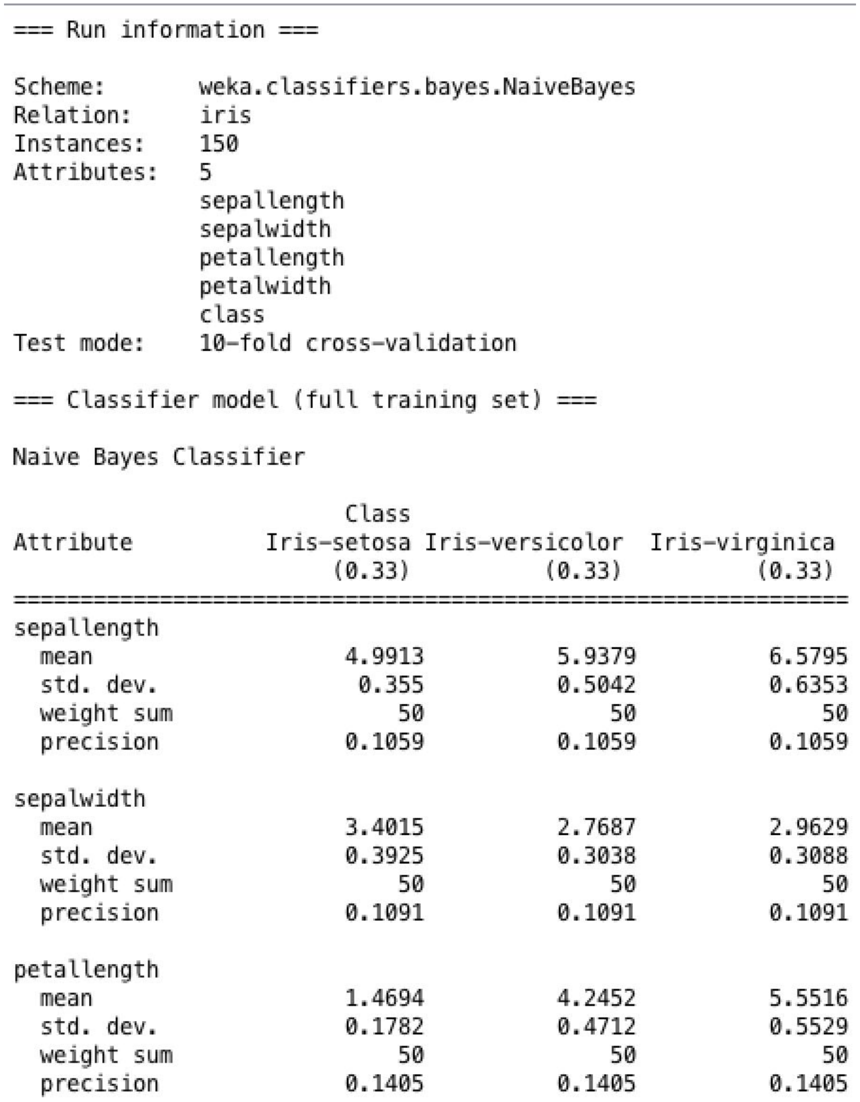


```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

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

petalwidth:
        < 0.8   -> Iris-setosa
        < 1.75  -> Iris-versicolor
        >= 1.75 -> Iris-virginica
(144/150 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         138               92      %
Incorrectly Classified Instances        12                8      %
Kappa statistic                          0.88
Mean absolute error                      0.0533
Root mean squared error                  0.2309
Relative absolute error                 12      %
Root relative squared error             48.9898 %
Total Number of Instances              150
```

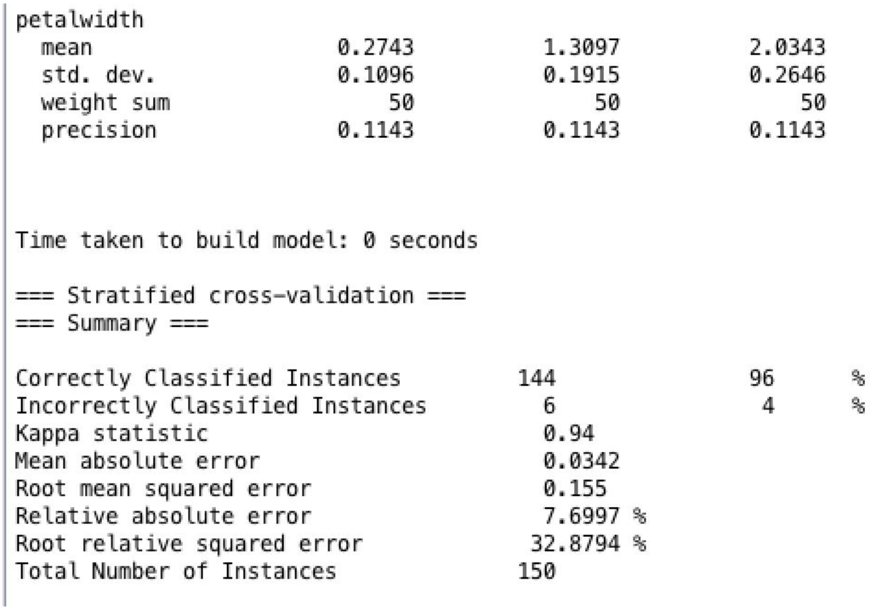


```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.880    0.060    0.880      0.880    0.880      0.820    0.910     0.814     Iris-versicolor
                 0.880    0.060    0.880      0.880    0.880      0.820    0.910     0.814     Iris-virginica
Weighted Avg.    0.920    0.040    0.920      0.920    0.920      0.880    0.940     0.876

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 44  6 |  b = Iris-versicolor
  0  6 44 |  c = Iris-virginica
```

######################   Naive Beyes ########################



```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

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

Naive Bayes Classifier

                         Class
Attribute          Iris-setosa Iris-versicolor  Iris-virginica
                        (0.33)          (0.33)          (0.33)
===============================================================
sepallength
  mean                   4.9913          5.9379          6.5795
  std. dev.               0.355          0.5042          0.6353
  weight sum                 50              50              50
  precision              0.1059          0.1059          0.1059

sepalwidth
  mean                   3.4015          2.7687          2.9629
  std. dev.              0.3925          0.3038          0.3088
  weight sum                 50              50              50
  precision              0.1091          0.1091          0.1091

petallength
  mean                   1.4694          4.2452          5.5516
  std. dev.              0.1782          0.4712          0.5529
  weight sum                 50              50              50
  precision              0.1405          0.1405          0.1405
```

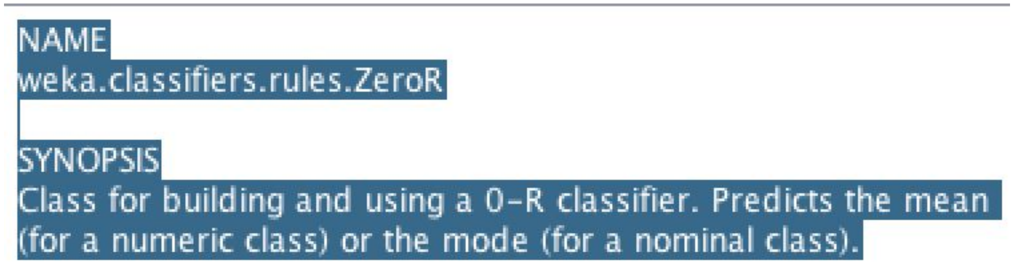


```
petalwidth
  mean                     0.2743         1.3097         2.0343
  std. dev.                0.1096         0.1915         0.2646
  weight sum                   50             50             50
  precision                0.1143         0.1143         0.1143



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0342
Root mean squared error                  0.155
Relative absolute error                  7.6997 %
Root relative squared error             32.8794 %
Total Number of Instances              150
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.960    0.040    0.923      0.960    0.941      0.911    0.992     0.983     Iris-versicolor
                 0.920    0.020    0.958      0.920    0.939      0.910    0.992     0.986     Iris-virginica
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.994     0.989

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

- Give a brief description of each of the methods. (hint: you can get synopsis of each method by right clicking on the chosen classifier and selecting “more” in the about section)

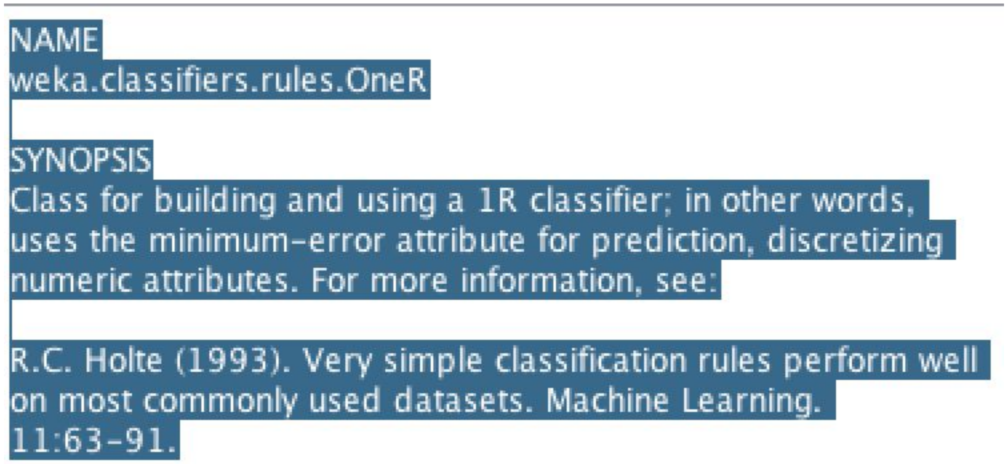
Zero-R



NAME
weka.classifiers.rules.ZeroR

SYNOPSIS
Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class).

One-R



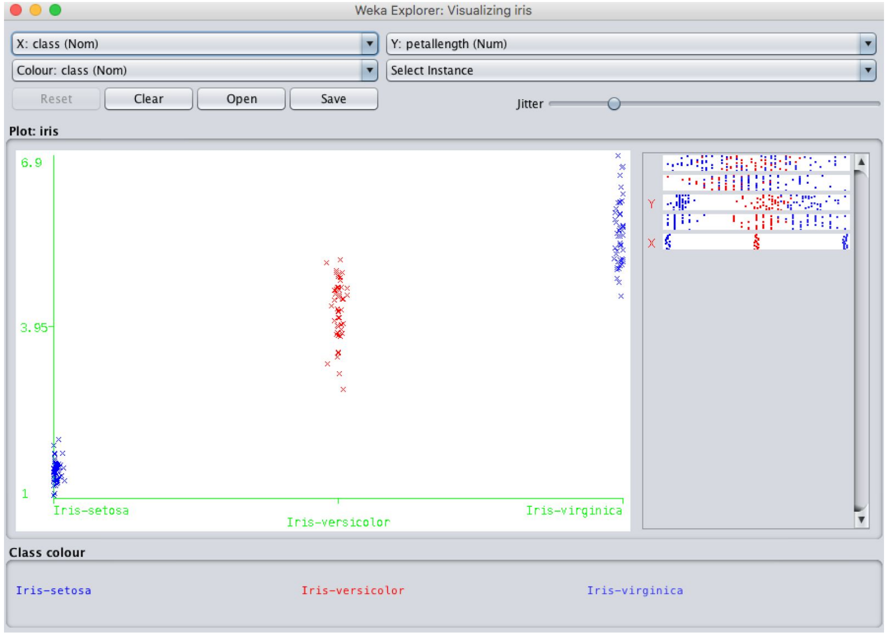
NAME
weka.classifiers.rules.OneR

SYNOPSIS
Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes. For more information, see:

R.C. Holte (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning. 11:63-91.

Naive-Bayes

- Do the decisions/models produced by the classifiers, in relation to the iris dataset, make sense to you? Why? How do they differ from each other? Yes, they do make sense. The zero-R has no rules and only chooses the class with highest mode(or 1st one if all equal), that is why zero-R was only 33% correct because there are three class values with equal number of occurances. One-R seen below used petalwidth as its one classifier. As we look at visual. We see that there was no confusion regarding classifying Iris-setosa, but there was a little overlap between Iris-versicolor and Iris-virginica(by the tune of 6% FP’s) The naive-bayes was the best classifier, but only a little over One-R.



Weka Explorer: Visualizing iris

X: class (Nom)

Y: petallength (Num)

Colour: class (Nom)

- How did each one of the methods perform? We will cover the evaluation techniques later in the class – for now you can choose common sense or one of the techniques that Weka presents with the model (either training data set, cross-validation or % split). Naive Bayes performed the best, while Zero-R was the poorest performer.

- Which method provided you with the most/least knowledge (insight into your data set/rules/patterns) and why?

	Naive Bayes	Zero-R
Most Knowledge	This classifier gives you the most insight into your data because it looks for interdependencies amongst features.	[redacted]
Least Knowledge	[redacted]	Zero-R makes no assumes there are no inderdendencies on any predictors. This method of classification does not lead to a deeper understanding of the data you are presented.

Appendix

################################### Zero-R ###################################

=== Run information ===

```
Scheme:       weka.classifiers.rules.ZeroR
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data
```

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

```
ZeroR predicts class value: Iris-setosa


Time taken to build model: 0 seconds
```

=== Evaluation on training set ===

```
Time taken to test model on training data: 0 seconds
```

=== Summary ===

```
Correctly Classified Instances          50               33.3333 %
Incorrectly Classified Instances       100               66.6667 %
Kappa statistic                          0
Mean absolute error                      0.4444
Root mean squared error                  0.4714
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              150
```

=== Detailed Accuracy By Class ===

```
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.333      1.000    0.500      0.000    0.500     0.333     Iris-setosa
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.333     Iris-versicolor
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.333     Iris-virginica
Weighted Avg.    0.333    0.333    0.111      0.333    0.167      0.000    0.500     0.333
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
 50  0  0 |  b = Iris-versicolor
 50  0  0 |  c = Iris-virginica
```

One-R

=== Run information ===

```
Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data
```

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

```
petalwidth:
       < 0.8    -> Iris-setosa
       < 1.75   -> Iris-versicolor
       >= 1.75  -> Iris-virginica
(144/150 instances correct)



Time taken to build model: 0 seconds
```

=== Evaluation on training set ===

```
Time taken to test model on training data: 0 seconds
```

=== Summary ===

```
Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0267
Root mean squared error                  0.1633
Relative absolute error                  6      %
Root relative squared error             34.641  %
Total Number of Instances              150
```

=== Detailed Accuracy By Class ===

```
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.980    0.050    0.907      0.980    0.942      0.913    0.965     0.896     Iris-versicolor
                 0.900    0.010    0.978      0.900    0.938      0.910    0.945     0.914     Iris-virginica
Weighted Avg.    0.960    0.020    0.962      0.960    0.960      0.941    0.970     0.937
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 49  1 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

Naive Beyes

=== Run information ===

```
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    evaluate on training data
```

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

```
Naive Bayes Classifier

                         Class
Attribute          Iris-setosa Iris-versicolor  Iris-virginica
                        (0.33)          (0.33)          (0.33)
===============================================================
sepallength
  mean                   4.9913          5.9379          6.5795
  std. dev.               0.355          0.5042          0.6353
  weight sum                 50              50              50
  precision              0.1059          0.1059          0.1059
```

```
sepalwidth
  mean                   3.4015          2.7687          2.9629
  std. dev.              0.3925          0.3038          0.3088
  weight sum                 50              50              50
  precision              0.1091          0.1091          0.1091

petallength
  mean                   1.4694          4.2452          5.5516
  std. dev.              0.1782          0.4712          0.5529
  weight sum                 50              50              50
  precision              0.1405          0.1405          0.1405

petalwidth
  mean                   0.2743          1.3097          2.0343
  std. dev.              0.1096          0.1915          0.2646
  weight sum                 50              50              50
  precision              0.1143          0.1143          0.1143



Time taken to build model: 0 seconds
```

=== Evaluation on training set ===

```
Time taken to test model on training data: 0.01 seconds
```

=== Summary ===

```
Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0324
Root mean squared error                  0.1495
Relative absolute error                  7.2883 %
Root relative squared error             31.7089 %
Total Number of Instances              150
```

=== Detailed Accuracy By Class ===

```
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.960    0.040    0.923      0.960    0.941      0.911    0.993     0.986     Iris-versicolor
                 0.920    0.020    0.958      0.920    0.939      0.910    0.993     0.987     Iris-virginica
Weighted Avg.    0.960    0.020    0.960      0.960    0.960      0.940    0.995     0.991
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

References

ZeroR: https://youtu.be/kUbYN4AcPmA OneR: https://youtu.be/phnkMGDZUNI Naive Bayes: https://youtu.be/XcwH9JGfZOU	