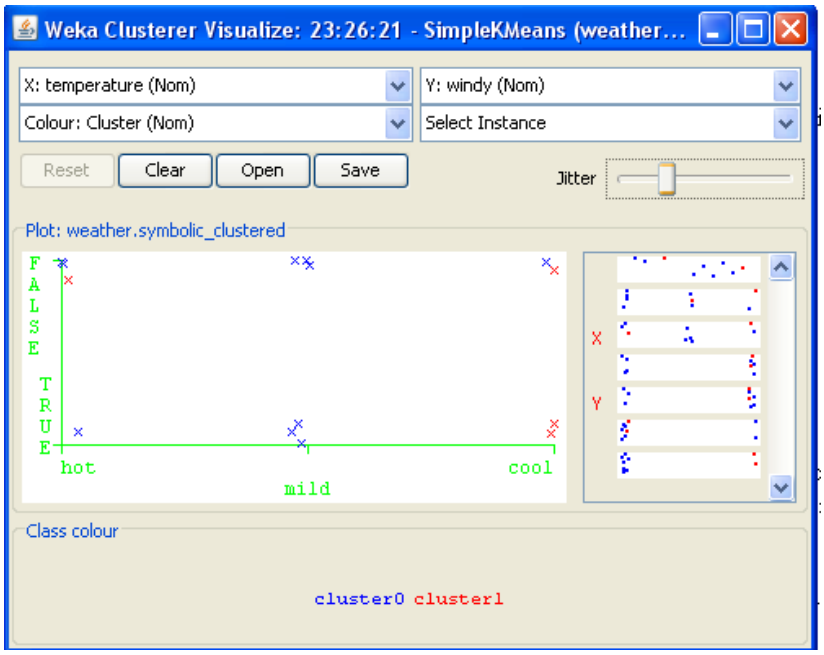
Project 2.1

Clustering

I clustered the samples with standard k-means with 500 iterations and in two classes. I was interested to see the clustering of windy and humidity: As you can see, when weather is cold and windy people mostly do not play golf and on the other hand, when weather is mild, people mostly play regardless of the wind.



=== Run information ===

Scheme: *weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10*

Relation: *weather.symbolic*

Instances: *14*

Attributes: *5*

outlook

temperature

humidity

windy

play

Test mode: evaluate on training data

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

kMeans

======

Number of iterations: 4

Within cluster sum of squared errors: 26.0

Missing values globally replaced with mean/mode

Cluster centroids:

Cluster#

Attribute Full Data 0 1

(14) (10) (4)

==========

outlook sunny sunny overcast

temperature mild mild cool

humidity high high normal

windy FALSE FALSE TRUE

play yes yes yes

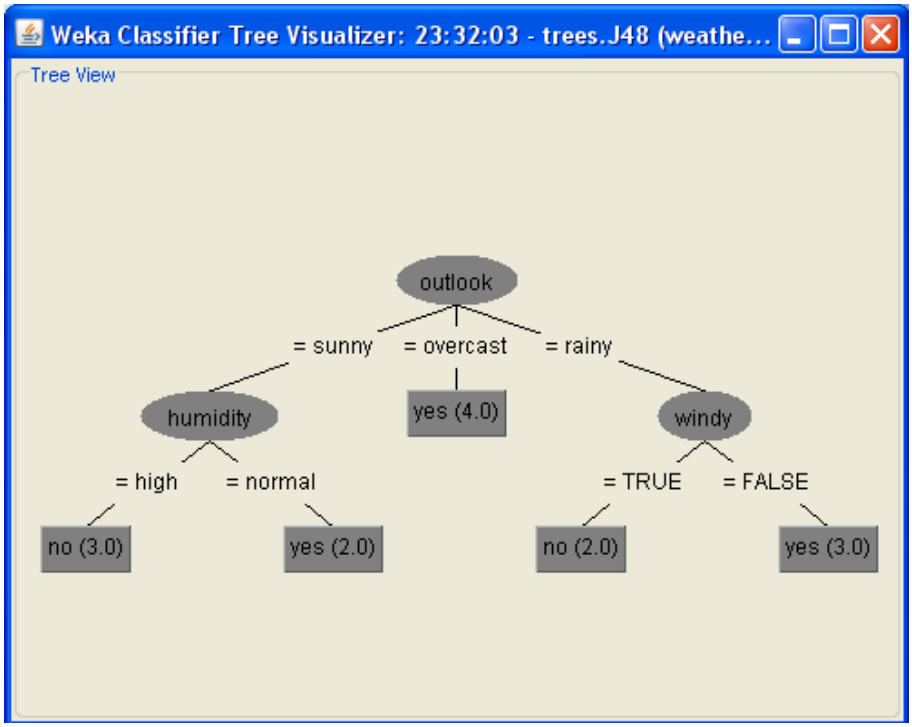
Clustered Instances

0 10 (71%)

1 4 (29%)

Classification with Decision Tree

I ran J48 with 10 fold cross-validation and obtained only 50% of accuracy. Below are the results and the tree, which is fairly reasonable.



=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: 10-fold cross-validation

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

J48 pruned tree

----

outlook = sunny

| humidity = high: no (3.0)

| humidity = normal: yes (2.0)

outlook = overcast: yes (4.0)

outlook = rainy

| windy = TRUE: no (2.0)

| windy = FALSE: yes (3.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*7*	*50*	*%*
Incorrectly Classified Instances	*7*	*50*	*%*
Kappa statistic	*-0.0426*		
Mean absolute error	*0.4167*		
Root mean squared error	*0.5984*		
Relative absolute error	*87.5*	*%*	
Root relative squared error	*121.2987*	*%*	
Total Number of Instances	*14*		

=== Detailed Accuracy By Class ===

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*ROC Area*	*Class*
	0.556	*0.6*	*0.625*	*0.556*	*0.588*	*0.633*	*yes*
	0.4	*0.444*	*0.333*	*0.4*	*0.364*	*0.633*	*no*
Weighted Avg.	*0.5*	*0.544*	*0.521*	*0.5*	*0.508*	*0.633*	

=== Confusion Matrix ===

a	*b*	*<-- classified as*

5 4 | a = yes

3 2 | b = no

Classification with Rule-Based Classifiers

***ConjuctiveRule**:*

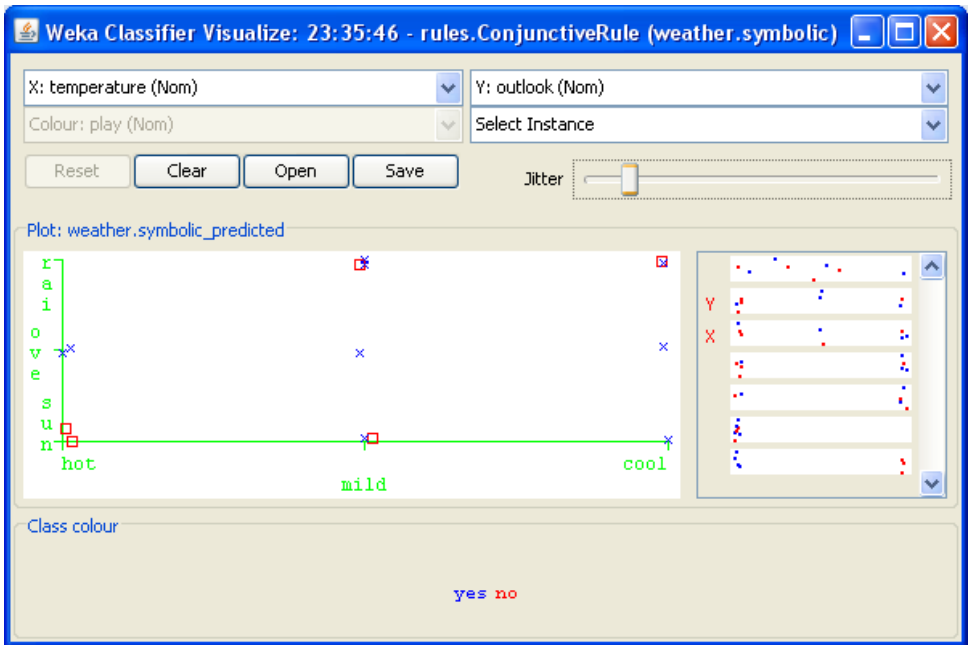
Achieves 64% of correct classification. Below is the visualization of classification error based on outlook, which indicates that classification errors happen mostly when the weather is rainy.

This class implements a single conjunctive rule learner that can predict for numeric and nominal class labels.

A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression. In this case, the consequent is the distribution of the available classes (or numeric value) in the dataset. If the test instance is not covered by this rule, then it's predicted using the default class distributions/value of the data not covered by the rule in the training data. This learner selects an antecedent by computing the Information Gain of each antecendent and prunes the generated rule using Reduced Error Prunning (REP).

For classification, the Information of one antecedent is the weighted average of the entropies of both the data covered and not covered by the rule.
For regression, the Information is the weighted average of the mean-squared errors of both the data covered and not covered by the rule.

In pruning, weighted average of accuracy rate of the pruning data is used for classification while the weighted average of the mean-squared errors of the pruning data is used for regression.



=== Run information ===

Scheme: weka.classifiers.rules.ConjunctiveRule -N 3 -M 2.0 -P -1 -S 1

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: 10-fold cross-validation

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

Single conjunctive rule learner:

=> play = yes

Class distributions:

Covered by the rule:

yes no

0.6 0.4

Not covered by the rule:

yes no

0 0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	9	64.2857 %
Incorrectly Classified Instances	5	35.7143 %
Kappa statistic	0	
Mean absolute error	0.4762	
Root mean squared error	0.5051	

Relative absolute error 100 %

Root relative squared error 102.3787 %

Total Number of Instances 14

=== Detailed Accuracy By Class ===

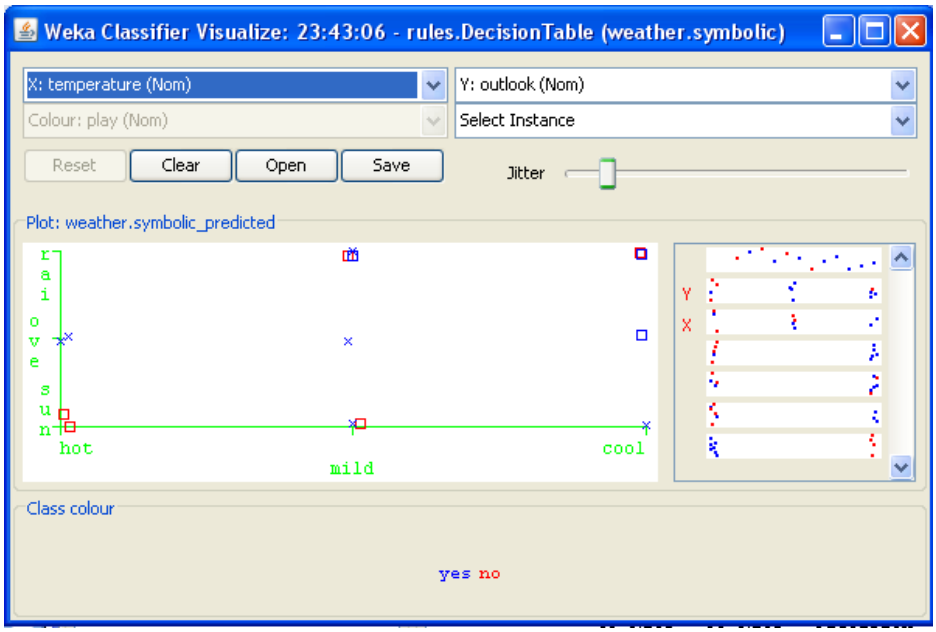
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.643	1	0.783	0.333	yes
	0	0	0	0	0	0.333	no
Weighted Avg.	0.643	0.643	0.413	0.643	0.503	0.333	

=== Confusion Matrix ===

a	b	<-- classified as
9	0	a = yes
5	0	b = no

Decision Table

Achieves a classification error of 42%. Again I visualized the classification error for the outlook vs temperature. This time we have errors even if the weather is cold.



=== Run information ===

Scheme: *weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"*

Relation: *weather.symbolic*

Instances: *14*

Attributes: *5*

outlook

temperature

humidity

windy

play

Test mode: *10-fold cross-validation*

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

Decision Table:

Number of training instances: 14

Number of Rules : 1

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 12

Merit of best subset found: 64.286

Evaluation (for feature selection): CV (leave one out)

Feature set: 5

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*6*	*42.8571 %*
Incorrectly Classified Instances	*8*	*57.1429 %*
Kappa statistic	*-0.3659*	
Mean absolute error	*0.5318*	

Root mean squared error 0.5583

Relative absolute error 111.6786 %

Root relative squared error 113.1584 %

Total Number of Instances 14

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.667	1	0.545	0.667	0.6	0.289	yes
	0	0.333	0	0	0	0.289	no
Weighted Avg.	0.429	0.762	0.351	0.429	0.386	0.289	

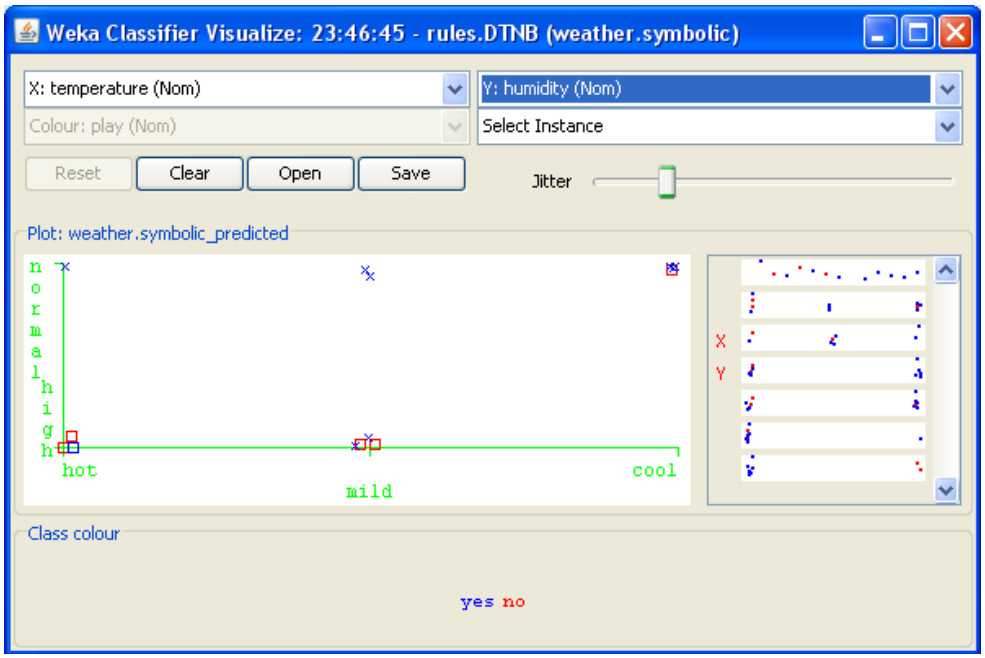
=== Confusion Matrix ===

a	b	<-- classified as
6	3	a = yes
5	0	b = no

DTNB

This classification achieves 57% of correct classification and if we visualize the classification error for temperature vs. humidity, we notice that we almost have no misclassification to No class when weather is mild or hot and the humidity is normal, which means that our classifier works well for this case which is consistent with our expectation from a decision table.

At each point in the search, the algorithm evaluates the merit of dividing the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes. A forward selection search is used, where at each step, selected attributes are modeled by naive Bayes and the remainder by the decision table, and all attributes are modelled by the decision table initially. At each step, the algorithm also considers dropping an attribute entirely from the model.



=== Run information ===

Scheme: *weka.classifiers.rules.DTNB -X 1*

Relation: *weather.symbolic*

Instances: *14*

Attributes: *5*

outlook

temperature

humidity

windy

play

Test mode: *10-fold cross-validation*

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

Decision Table:

Number of training instances: 14

Number of Rules : 12

Non matches covered by Majority class.

Evaluation (for feature selection): CV (leave one out)

Feature set: 1,2,4,5

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*8*	*57.1429 %*
Incorrectly Classified Instances	*6*	*42.8571 %*
Kappa statistic	*-0.1351*	
Mean absolute error	*0.5454*	
Root mean squared error	*0.5607*	
Relative absolute error	*114.5337 %*	
Root relative squared error	*113.6458 %*	
Total Number of Instances	*14*	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.889	1	0.615	0.889	0.727	0.078	yes
	0	0.111	0	0	0	0.078	no
Weighted Avg.	0.571	0.683	0.396	0.571	0.468	0.078	

=== Confusion Matrix ===

a	b	<-- classified as
8	1	a = yes
5	0	b = no

Nearest Neighbor Approaches

IB1 gives 50% of correct classification. KB1 with 3 neighbors gives about 64% correct classification which is improved because we are using more neighbors. If we use 7 neighbors instead, we get the same accuracy. If we use K-Star we get perfect classification with 20 for global blend. NNge works the same as well and achieves perfect classification. Generally approaches which use enough number of neighbors are more successful according to our experiments.

=== Run information ===

Scheme: weka.classifiers.lazy.IB1

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: 10-fold cross-validation

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

IB1 classifier

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*7*	*50 %*
Incorrectly Classified Instances	*7*	*50 %*
Kappa statistic	*0.0392*	
Mean absolute error	*0.5*	
Root mean squared error	*0.7071*	
Relative absolute error	*105 %*	
Root relative squared error	*143.3236 %*	
Total Number of Instances	*14*	

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.444 0.4 0.667 0.444 0.533 0.522 yes

0.6 0.556 0.375 0.6 0.462 0.522 no

Weighted Avg. 0.5 0.456 0.563 0.5 0.508 0.522

=== Confusion Matrix ===

a b <-- classified as

4 5 | a = yes

2 3 | b = no

=== Run information ===

Scheme: weka.classifiers.lazy.IBk -K 3 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: 10-fold cross-validation

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

IB1 instance-based classifier

using 3 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	*9*	*64.2857 %*
Incorrectly Classified Instances	*5*	*35.7143 %*
Kappa statistic	*0.1026*	
Mean absolute error	*0.4414*	
Root mean squared error	*0.4747*	
Relative absolute error	*92.699 %*	
Root relative squared error	*96.2242 %*	
Total Number of Instances	*14*	

=== Detailed Accuracy By Class ===

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*ROC Area*	*Class*
	0.889	*0.8*	*0.667*	*0.889*	*0.762*	*0.689*	*yes*
	0.2	*0.111*	*0.5*	*0.2*	*0.286*	*0.644*	*no*
Weighted Avg.	*0.643*	*0.554*	*0.607*	*0.643*	*0.592*	*0.673*	

=== Confusion Matrix ===

a b <-- classified as

8 1 | a = yes

4 1 | b = no

=== Run information ===

Scheme: weka.classifiers.rules.NNge -G 5 -I 5

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode: evaluate on training data

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

NNGE classifier

Rules generated :

class no IF : outlook in {rainy} ^ temperature in {mild,cool} ^ humidity in {high,normal} ^ windy in {TRUE} (2)

class yes IF : outlook in {overcast,rainy} ^ temperature in {hot,mild,cool} ^ humidity in {high,normal} ^ windy in {FALSE} (5)

class yes IF : outlook in {overcast} ^ temperature in {mild,cool} ^ humidity in {high,normal} ^ windy in {TRUE} (2)

class yes IF : outlook in {sunny} ^ temperature in {mild,cool} ^ humidity in {normal} ^ windy in {TRUE,FALSE} (2)

class no IF : outlook in {sunny} ^ temperature in {hot,mild} ^ humidity in {high} ^ windy in {TRUE,FALSE} (3)

Stat :

class yes : 3 exemplar(s) including 3 Hyperrectangle(s) and 0 Single(s).

class no : 2 exemplar(s) including 2 Hyperrectangle(s) and 0 Single(s).

Total : 5 exemplars(s) including 5 Hyperrectangle(s) and 0 Single(s).

Feature weights : [0.24674981977443894 0.029222565658954577 0.15183550136234153 0.04812703040826924]

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances 14 100 %

Incorrectly Classified Instances 0 0 %

Kappa statistic 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 14

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	yes
	1	0	1	1	1	1	no
Weighted Avg.	1	0	1	1	1	1	

=== Confusion Matrix ===

a	b	<-- classified as
9	0	a = yes
0	5	b = no