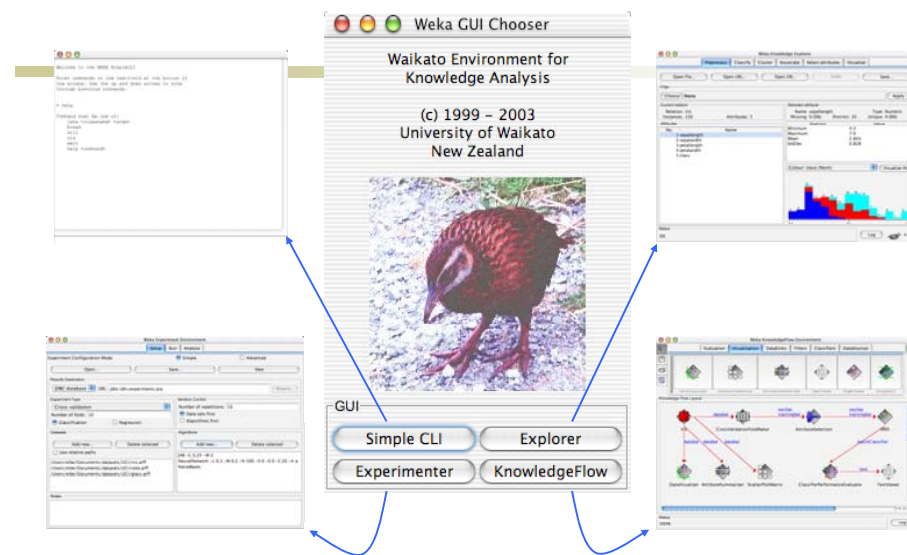


# Weka: Brief Introduction

- Features Covered in this Lecture
  - Preprocessing – Examining datasets and using filters.
  - Classification – selecting and running classifiers.
  - Visualization Tools – brief exposure



## Explorer: Preprocessing the data

- Data can be imported from a file in various formats: **ARFF**, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, **attribute selection**, transforming and combining attributes, ...

## WEKA only deals with “flat” files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
    atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

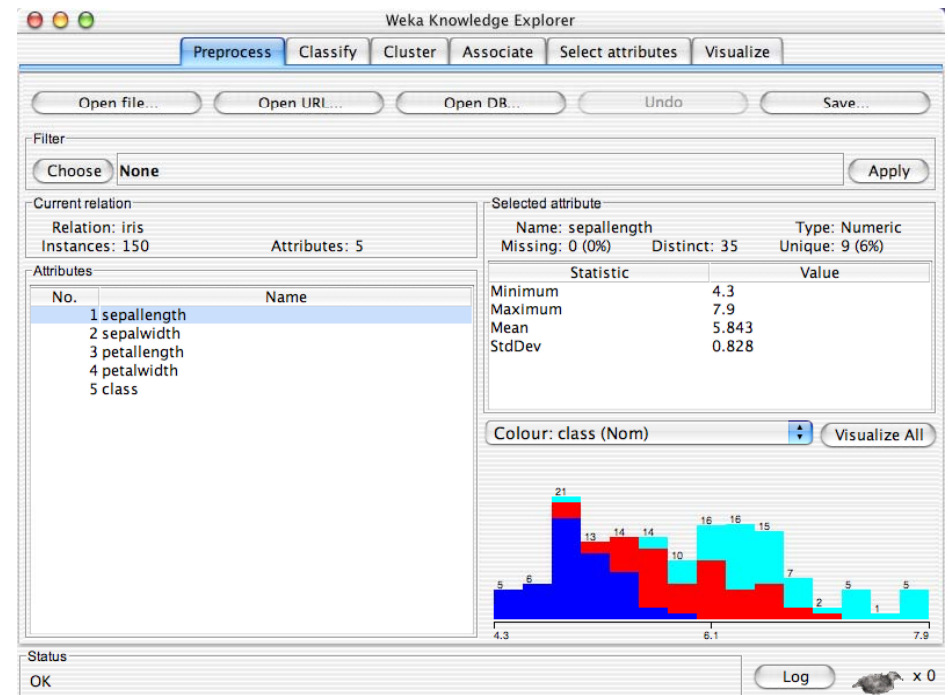
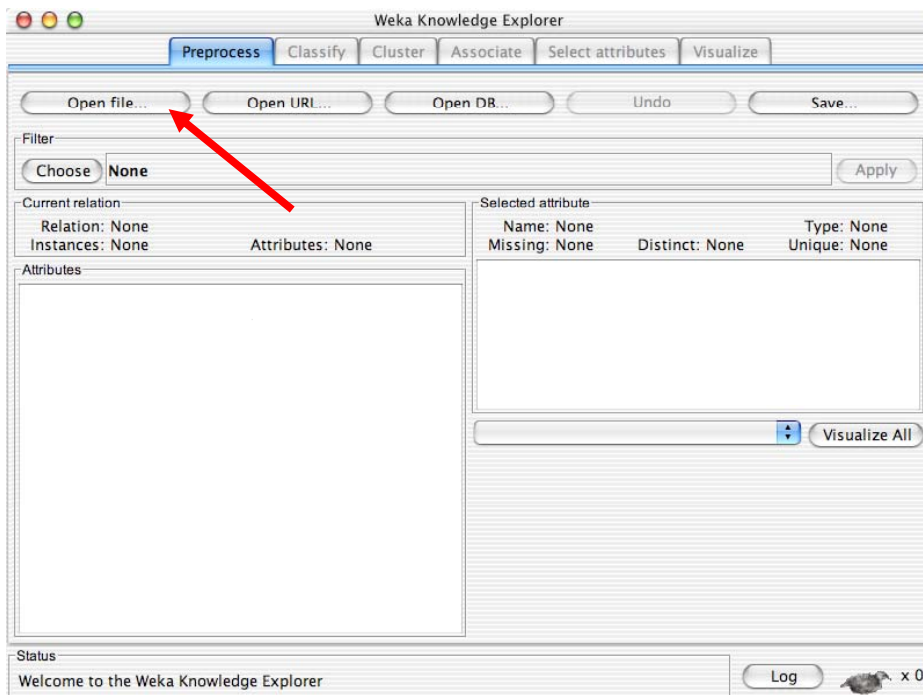
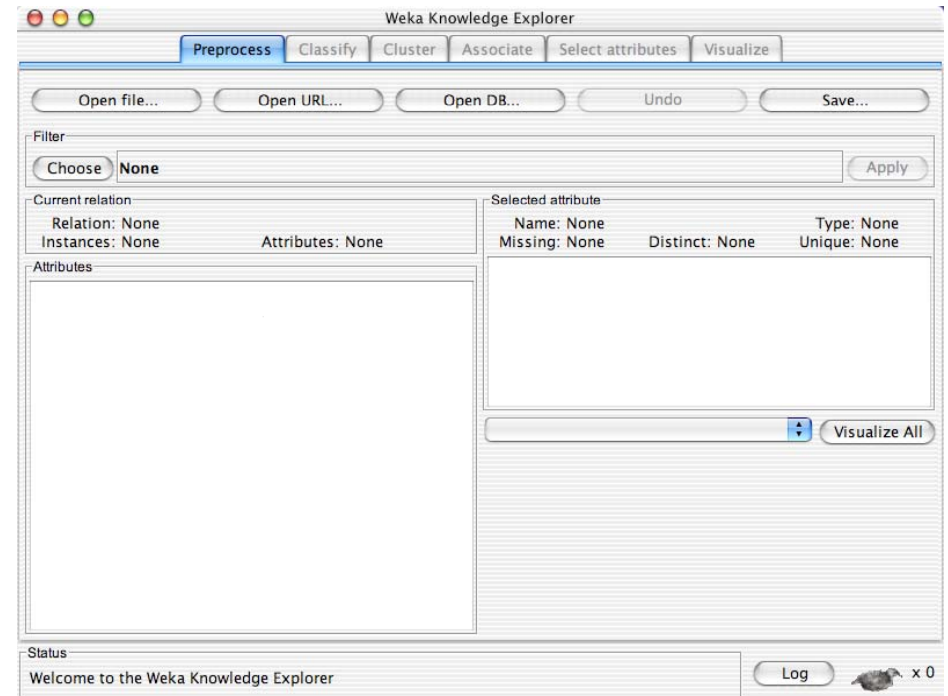
@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

Flat file in ARFF format

# WEKA only deals with “flat” files

@relation heart-disease-simplified  
 @attribute age numeric **numeric attribute**  
 @attribute sex { female, male} **nominal attribute**  
 @attribute chest\_pain\_type { typ\_angina, asympt, non\_anginal, atyp\_angina}  
 @attribute cholesterol numeric  
 @attribute exercise\_induced\_angina { no, yes}  
 @attribute class { present, not\_present}

@data  
 63,male,typ\_angina,233,no,not\_present  
 67,male,asympt,286,yes,present  
 67,male,asympt,229,yes,present  
 38,female,non\_anginal,?,no,not\_present  
 ...



Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose None Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute: Name: sepalength Type: Numeric Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Attributes:

No.	Name
1	sepalength
2	sepalwidth
3	petalength
4	petalwidth
5	class

Colour: class (Nom) Visualize All

Status: OK Log x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose None Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute: Name: class Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Attributes:

No.	Name
1	sepalength
2	sepalwidth
3	petalength
4	petalwidth
5	class

Colour: class (Nom) Visualize All

Status: OK Log x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose None Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute: Name: class Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Attributes:

No.	Name
1	sepalength
2	sepalwidth
3	petalength
4	petalwidth
5	class

Colour: class (Nom) Visualize All

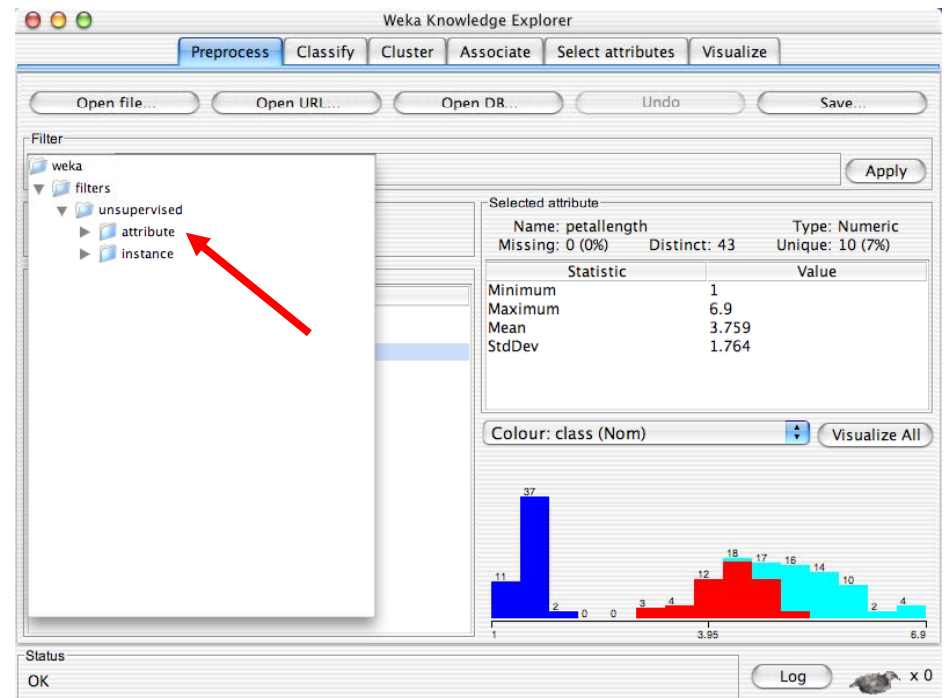
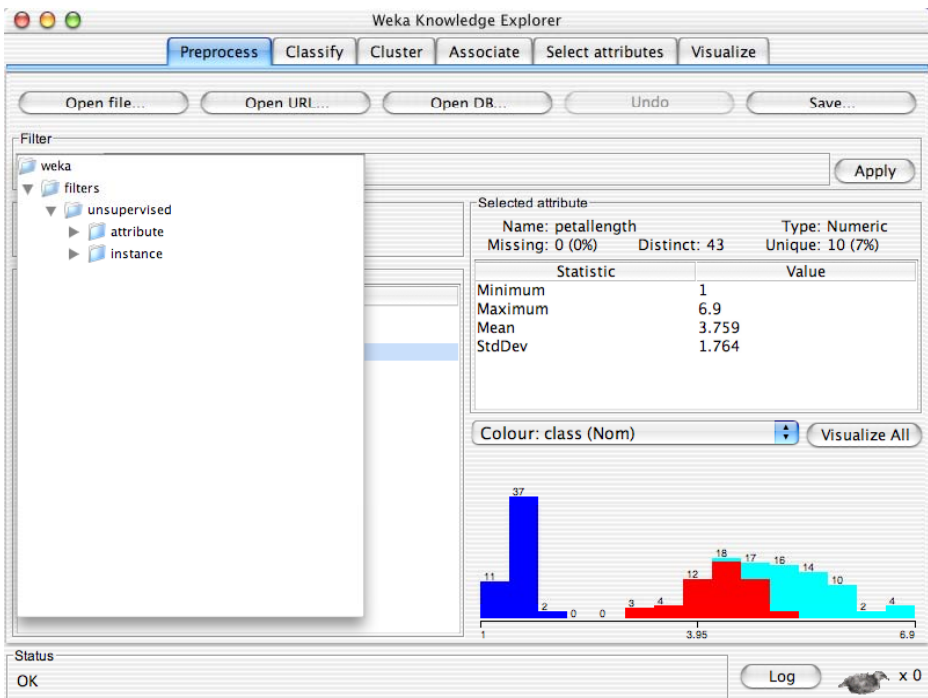
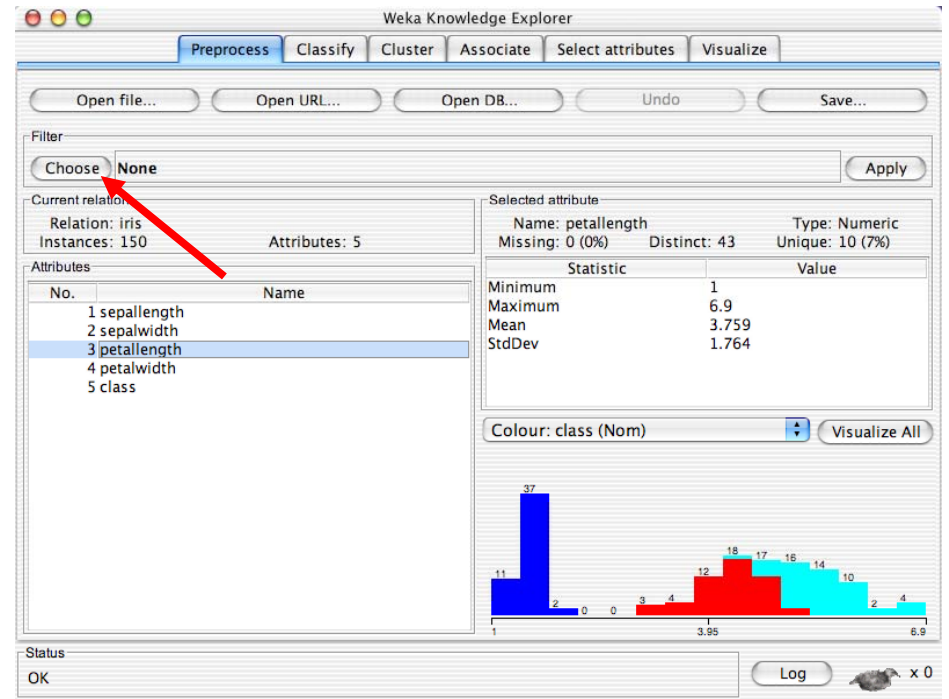
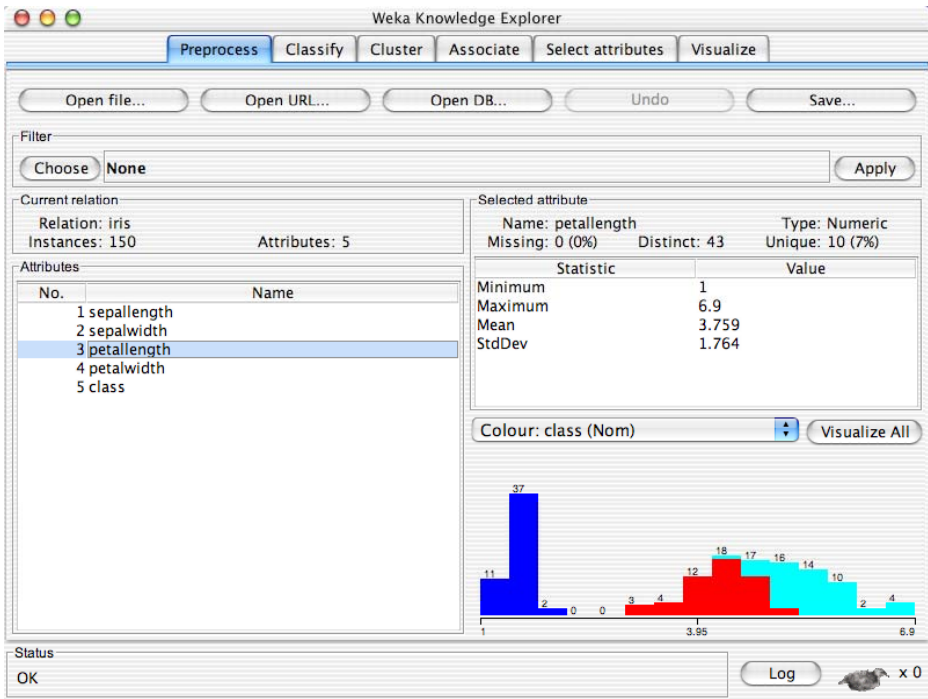
Status: OK Log x 0

Weka Knowledge Explorer

sepalength sepalwidth petalength

petalwidth class

Status: OK Log x 0



Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

- weka
  - filters
    - attribute
      - Add
      - AddCluster
      - AddExpression
      - AddNoise
      - Copy
      - Discretize
      - FirstOrder
      - MakeIndicator
      - MergeTwoValues
      - NominalToBinary
      - Normalize
      - NumericToBinary
      - NumericTransform
      - Obfuscate
      - PKIDiscretize
      - Remove
      - RemoveType

Apply

Selected attribute

Name: petalength      Type: Numeric  
Missing: 0 (0%)      Distinct: 43      Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)      Visualize All

Status: OK      Log      x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose Discretize -B 10 -R first-last      Apply

Current relation

Relation: iris      Attributes: 5  
Instances: 150

Selected attribute

Name: petalength      Type: Numeric  
Missing: 0 (0%)      Distinct: 43      Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)      Visualize All

Status: OK      Log      x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose Discretize -B 10 -R first-last      Apply

Current relation

Relation: iris      Attributes: 5  
Instances: 150

Selected attribute

Name: petalength      Type: Numeric  
Missing: 0 (0%)      Distinct: 43      Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)      Visualize All

Status: OK      Log      x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose Discretize -B 10 -R first-last      Apply

Current relation

Relation: iris      Attributes: 5  
Instances: 150

Selected attribute

Name: petalength      Type: Numeric  
Missing: 0 (0%)      Distinct: 43      Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom)      Visualize All

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.      More

attributeIndices: first-last

bins: 10

findNumBins: False

invertSelection: False

makeBinary: False

useEqualFrequency: False

Visualize All

Open...      Save...      OK      Cancel

Status: OK      Log      x 0

Weka Knowledge Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose **Discretize -B 10 -R first-last** weka.gui.GenericObjectEditor Apply

Current relation: Relation: iris Instances: 150 Attributes: : Numeric : 10 (7%)

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. More

Attributes:

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

attributeIndices: first-last

bins: 10

findNumBins: False

invertSelection: False

makeBinary: False

useEqualFrequency: False

Open... Save... OK Cancel

Status: OK Log x 0

Weka Knowledge Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose **Discretize -B 10 -R first-last** weka.gui.GenericObjectEditor Apply

Current relation: Relation: iris Instances: 150 Attributes: : Numeric : 10 (7%)

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. More

Attributes:

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

attributeIndices: first-last

bins: 10

findNumBins: False

invertSelection: False

makeBinary: False

useEqualFrequency: True

Open... Save... OK Cancel

Status: OK Log x 0

Weka Knowledge Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose **Discretize -B 10 -R first-last** weka.gui.GenericObjectEditor Apply

Current relation: Relation: iris Instances: 150 Attributes: : Numeric : 10 (7%)

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. More

Attributes:

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

attributeIndices: first-last

bins: 10

findNumBins: False

invertSelection: False

makeBinary: False

useEqualFrequency: True

Open... Save... OK Cancel

Status: OK Log x 0

Weka Knowledge Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Save...

Filter: Choose **Discretize -F -B 10 -R first-last** Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute:

Name: petal.length Type: Numeric

Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

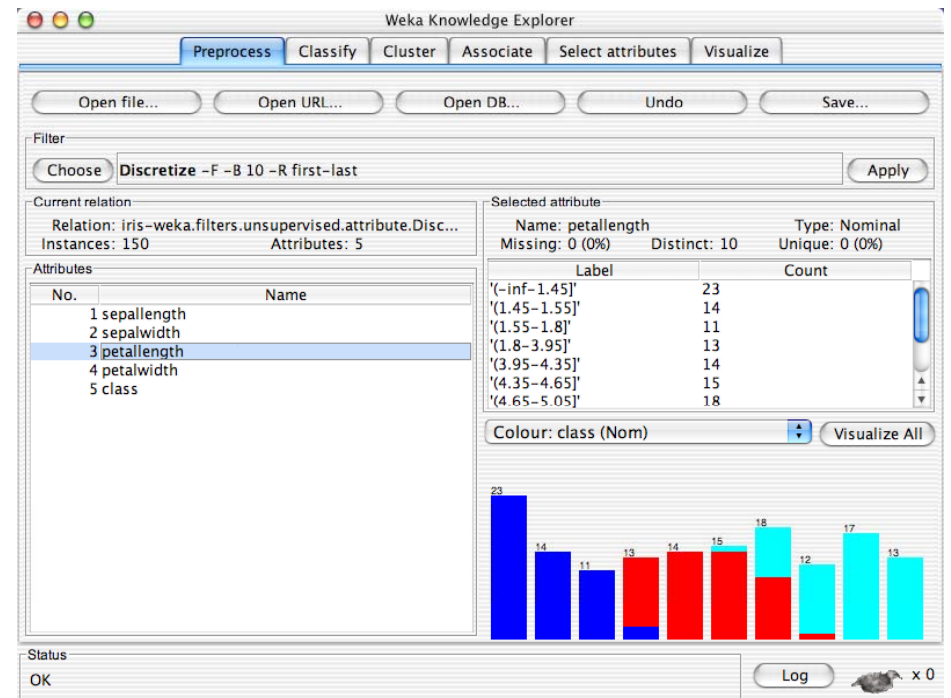
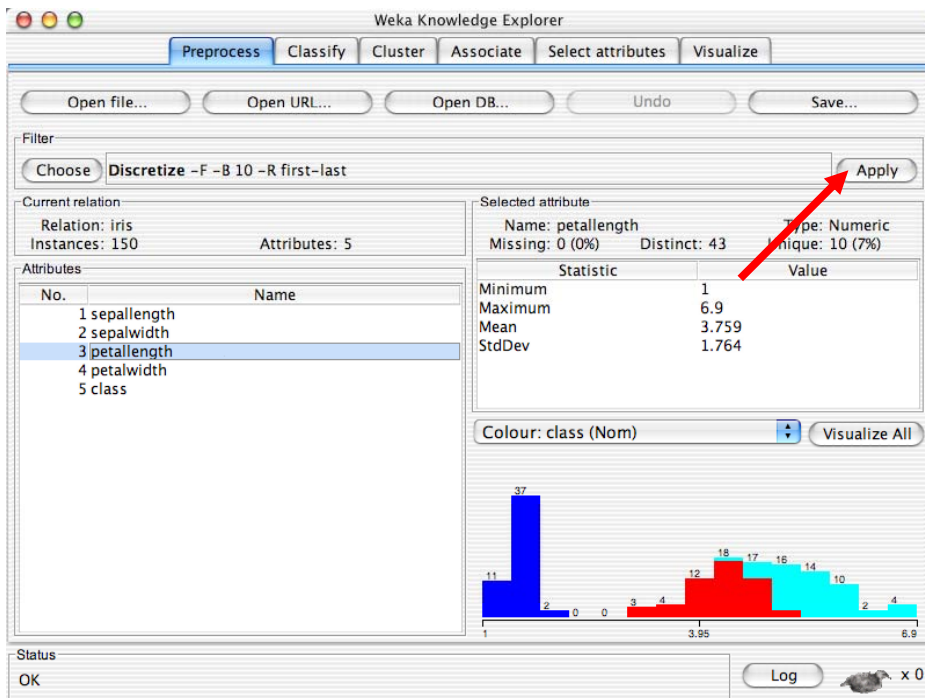
Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Attributes:

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

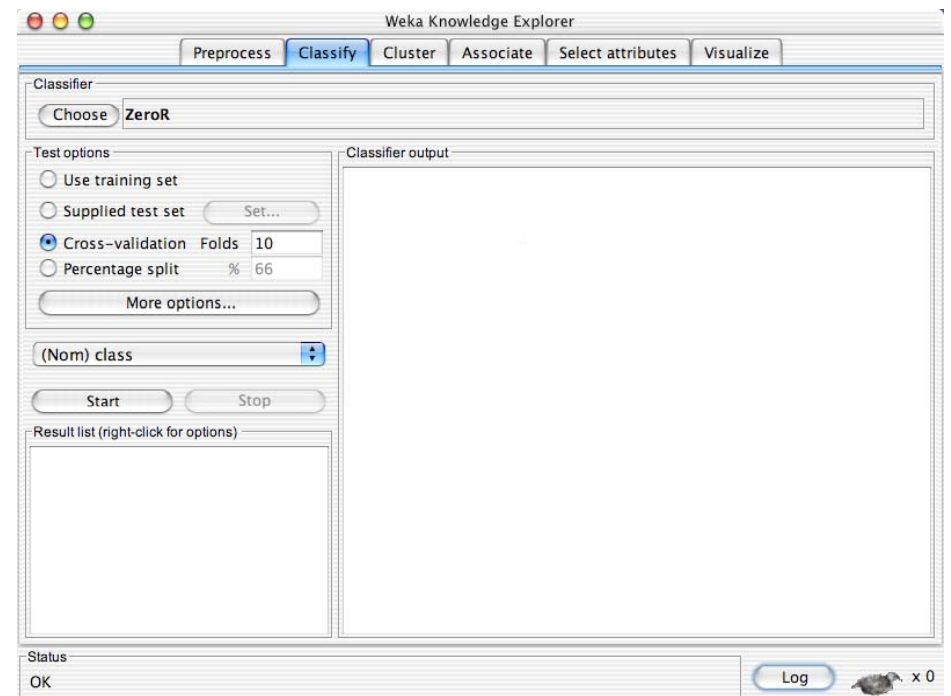
Colour: class (Nom) Visualize All

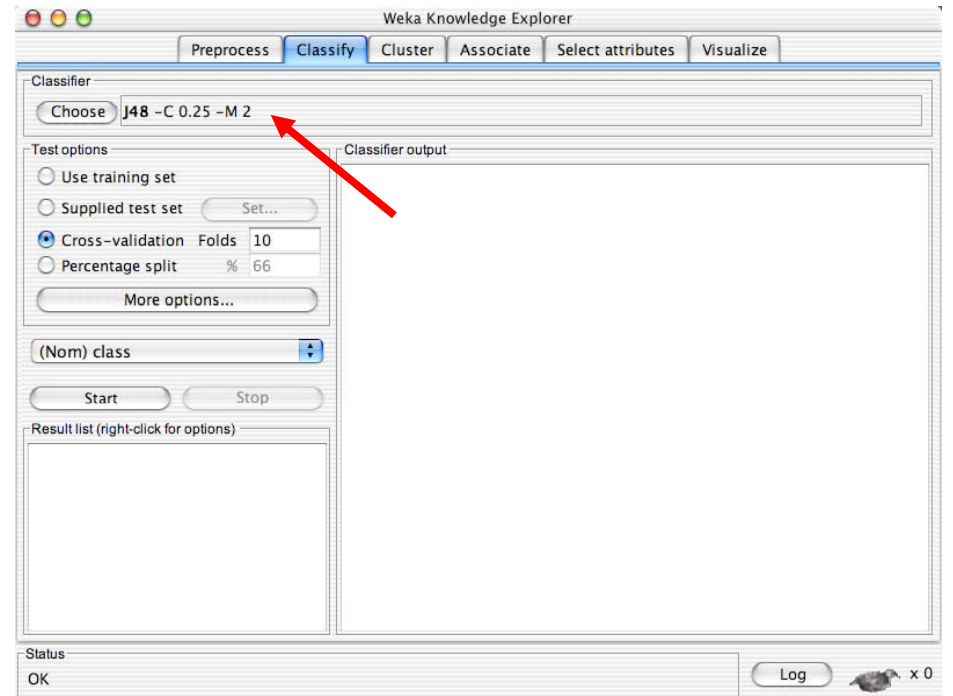
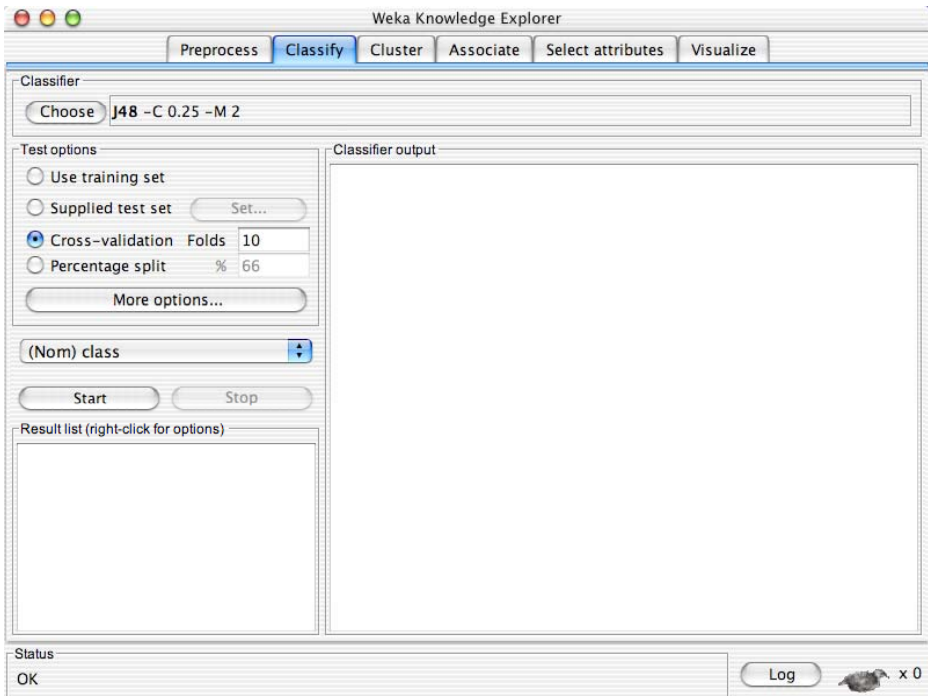
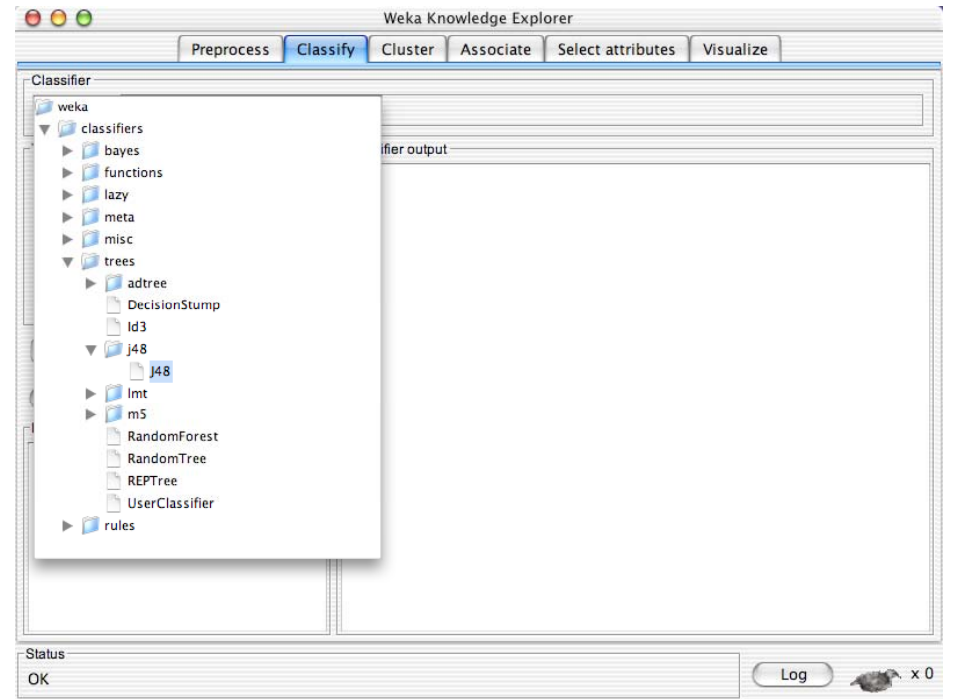
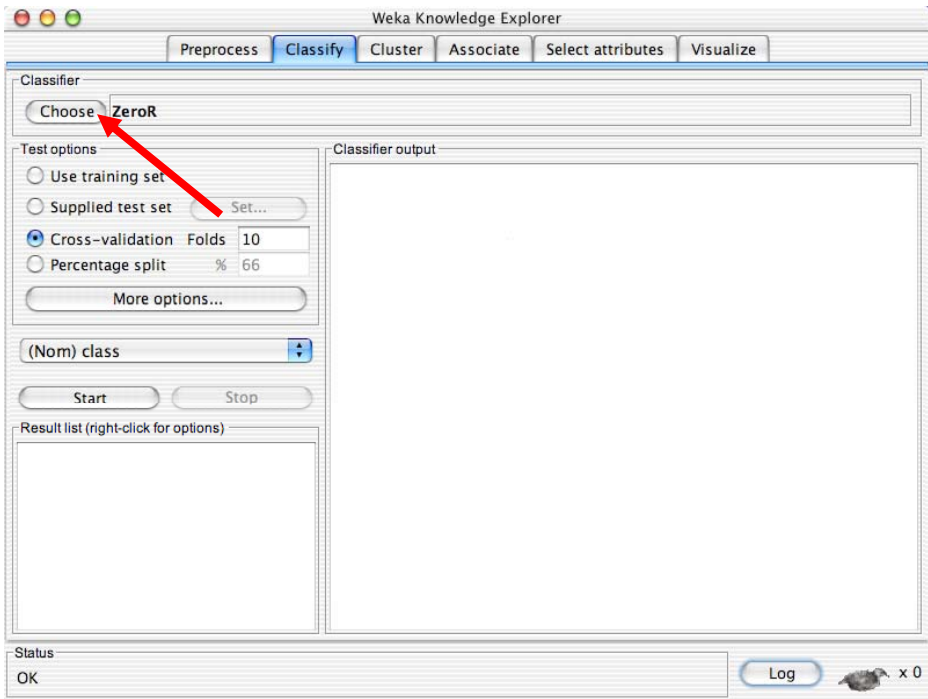
Status: OK Log x 0

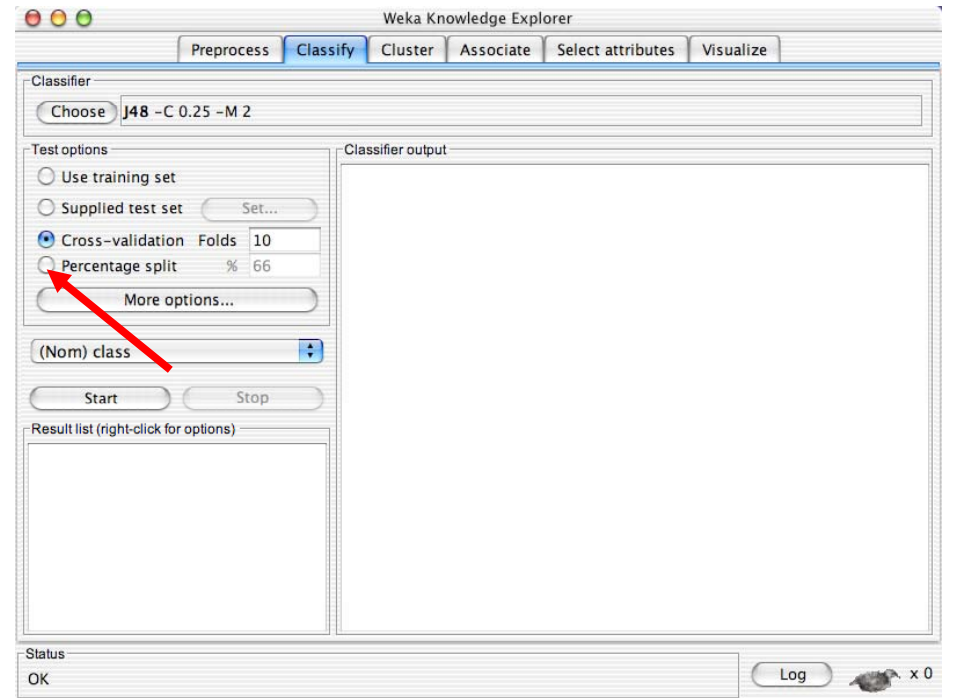
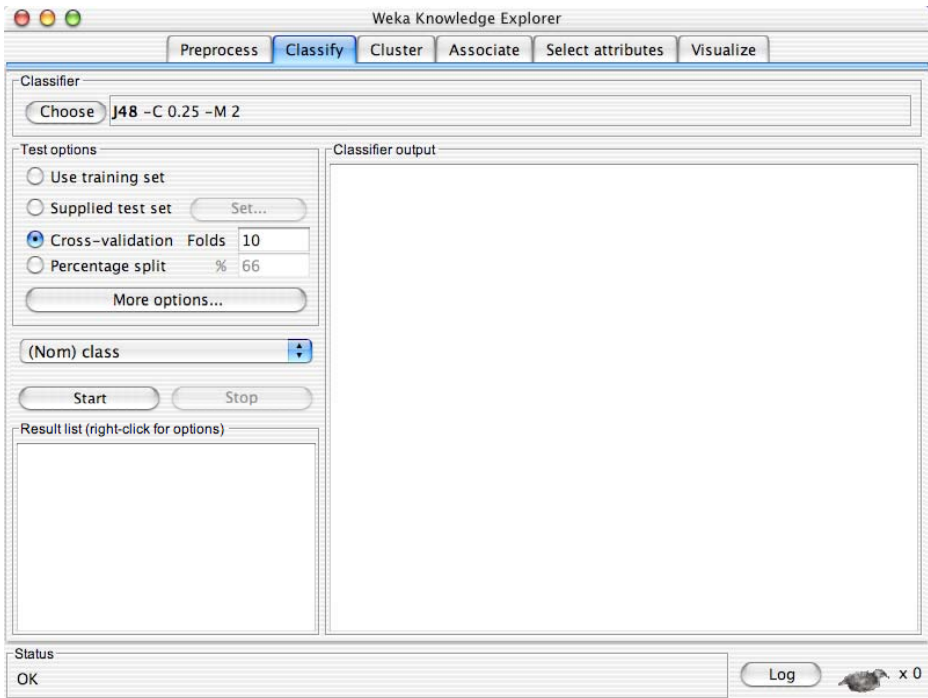
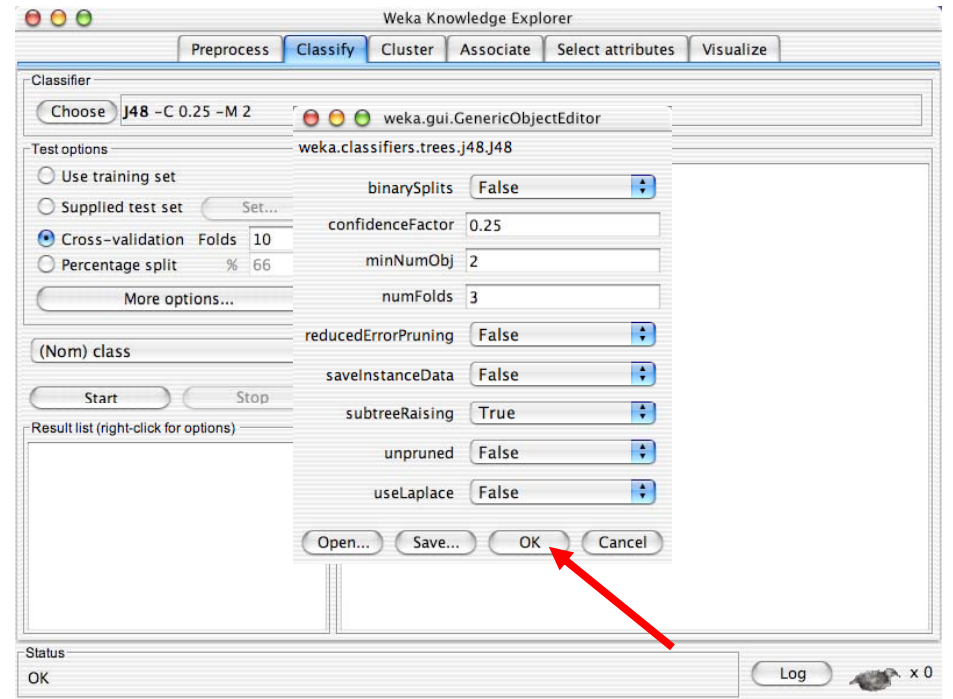
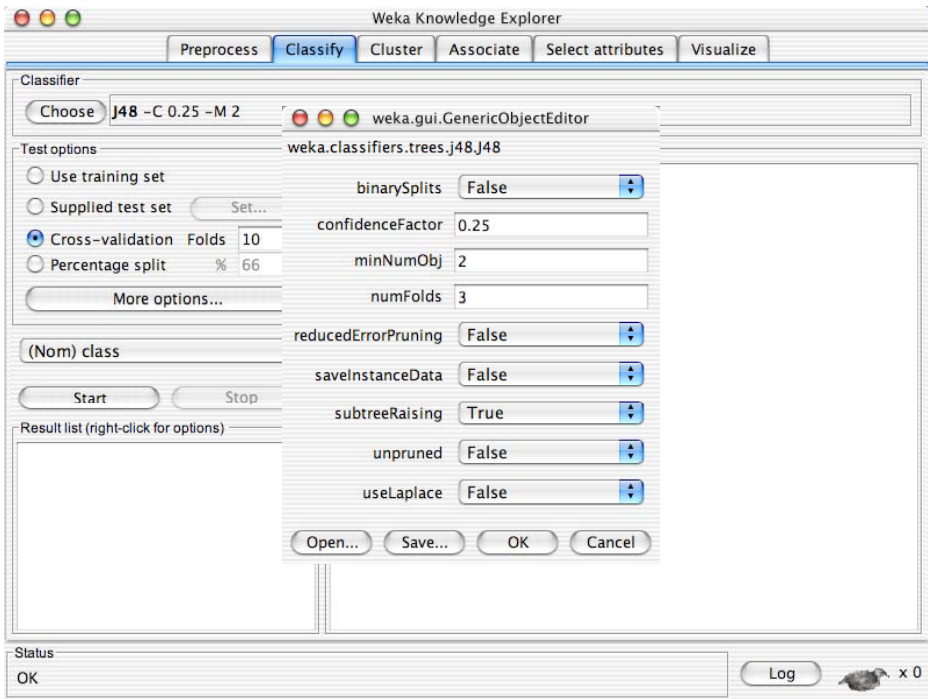


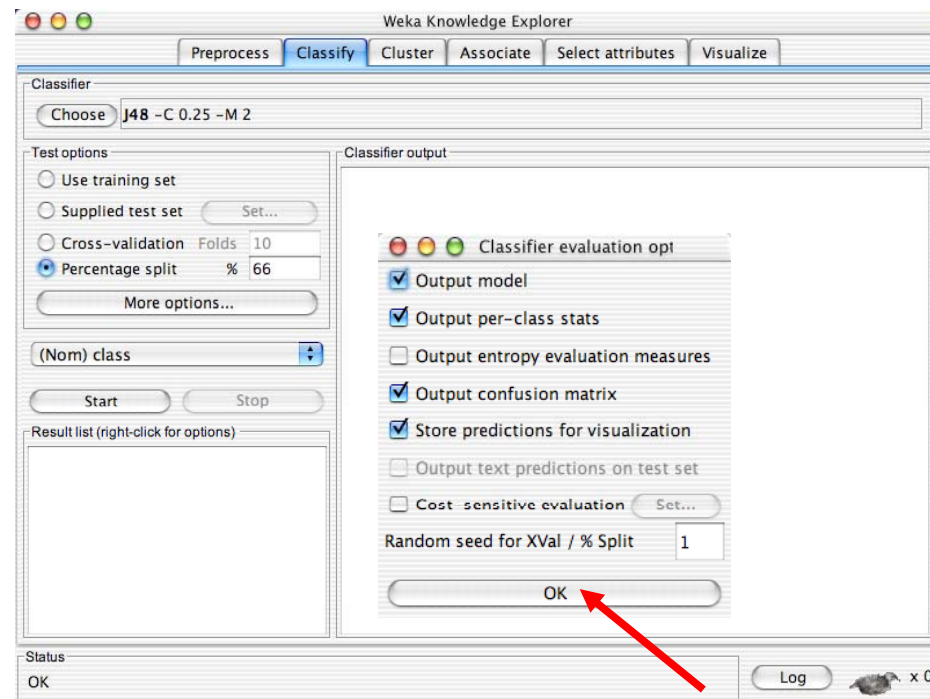
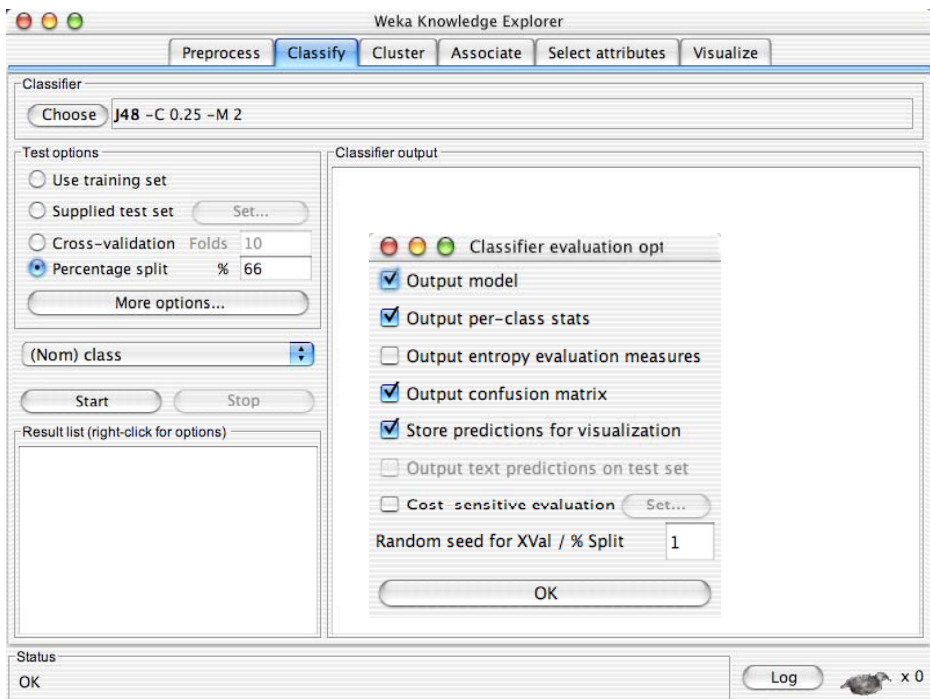
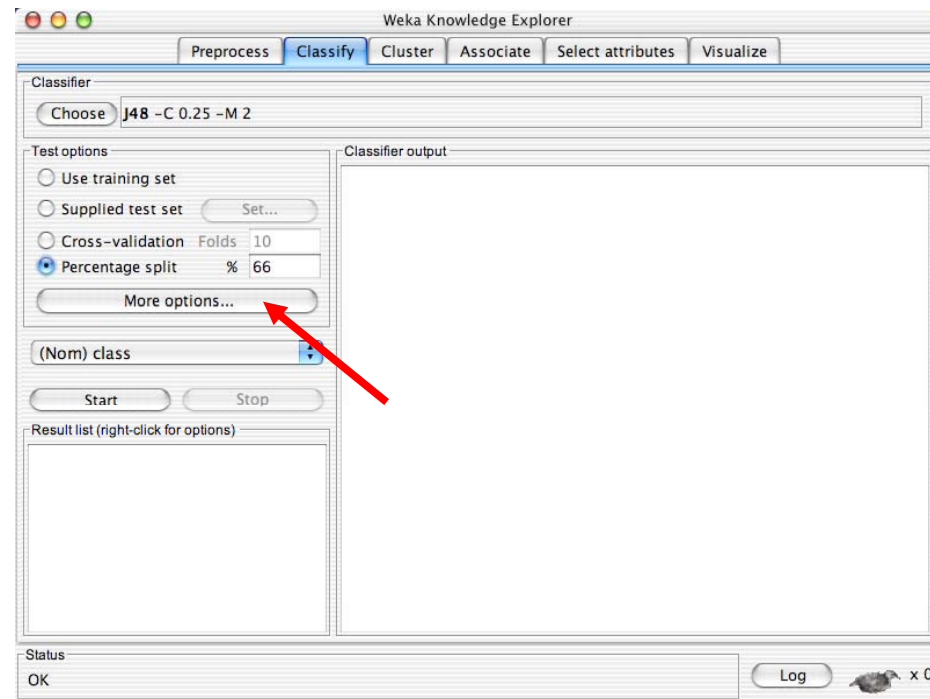
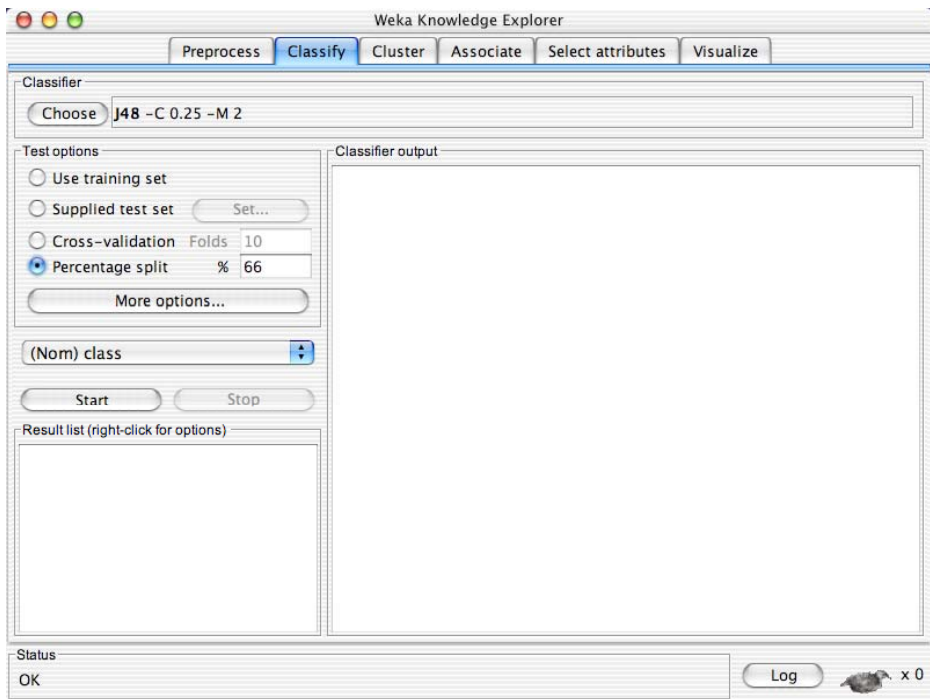
## Explorer: building “classifiers”

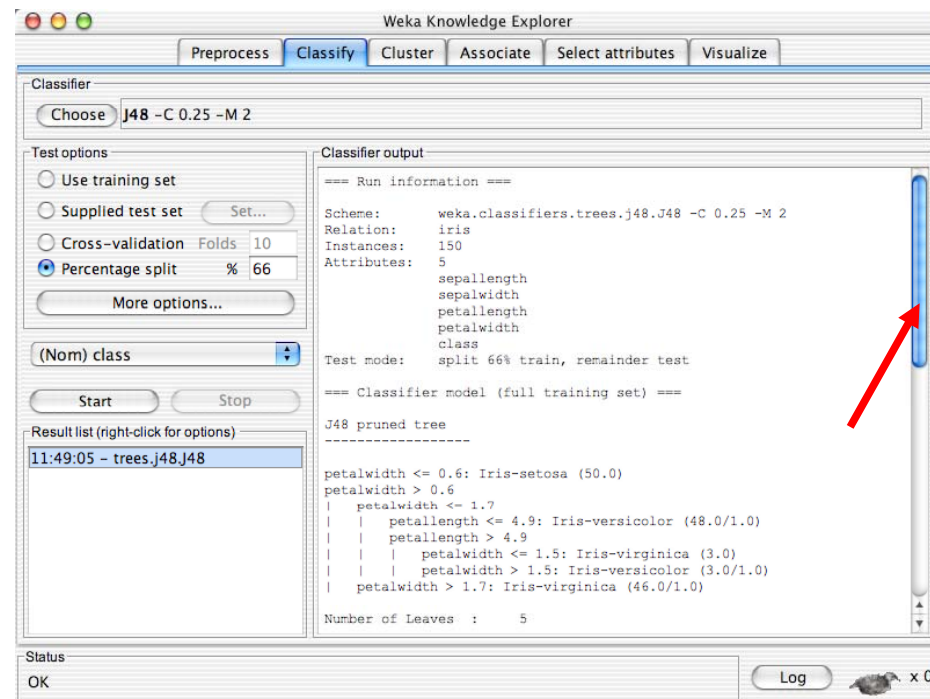
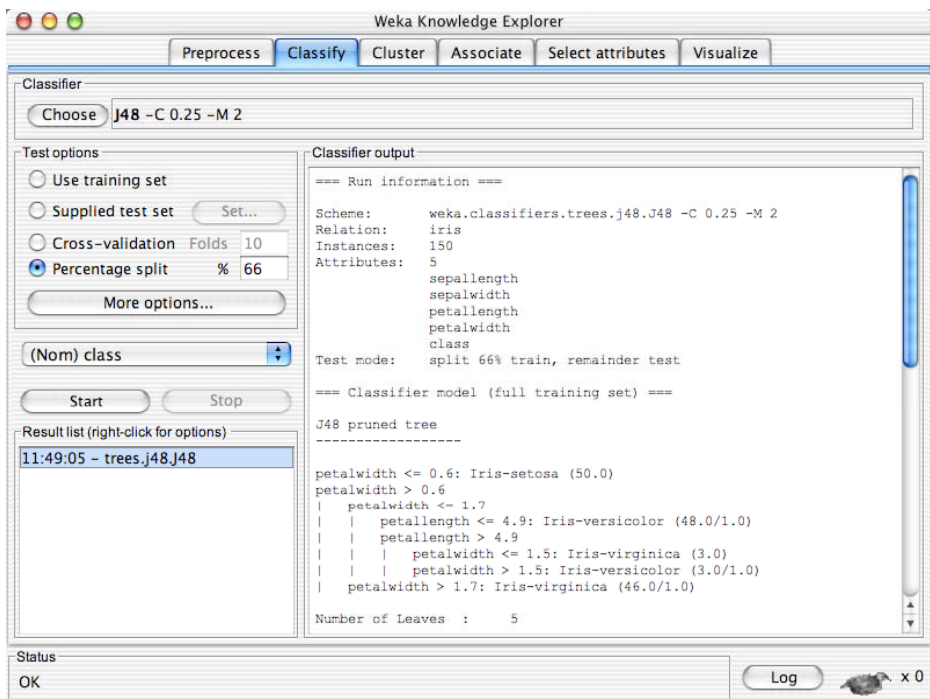
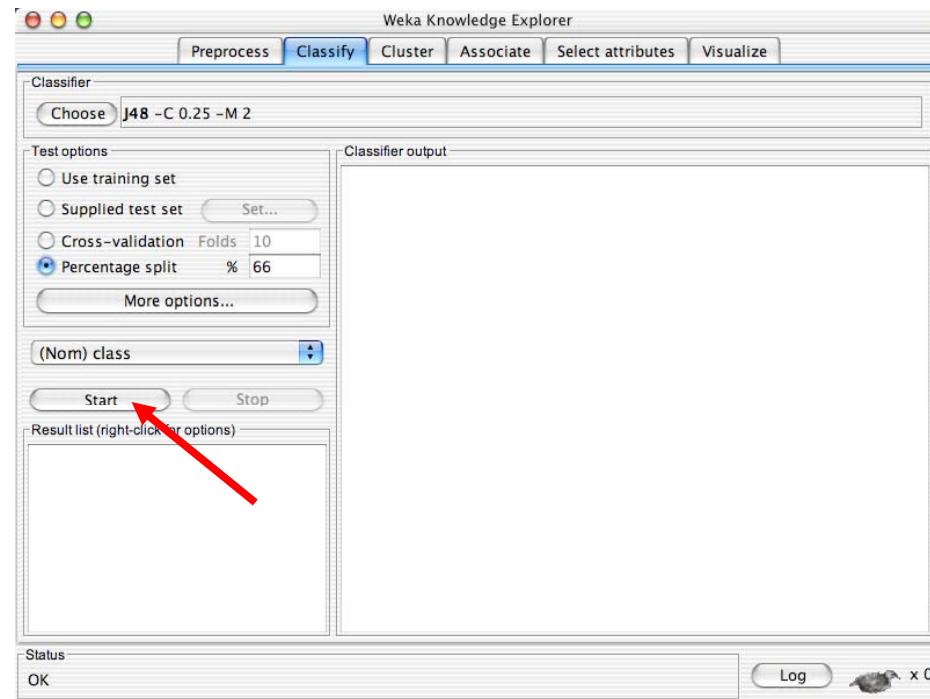
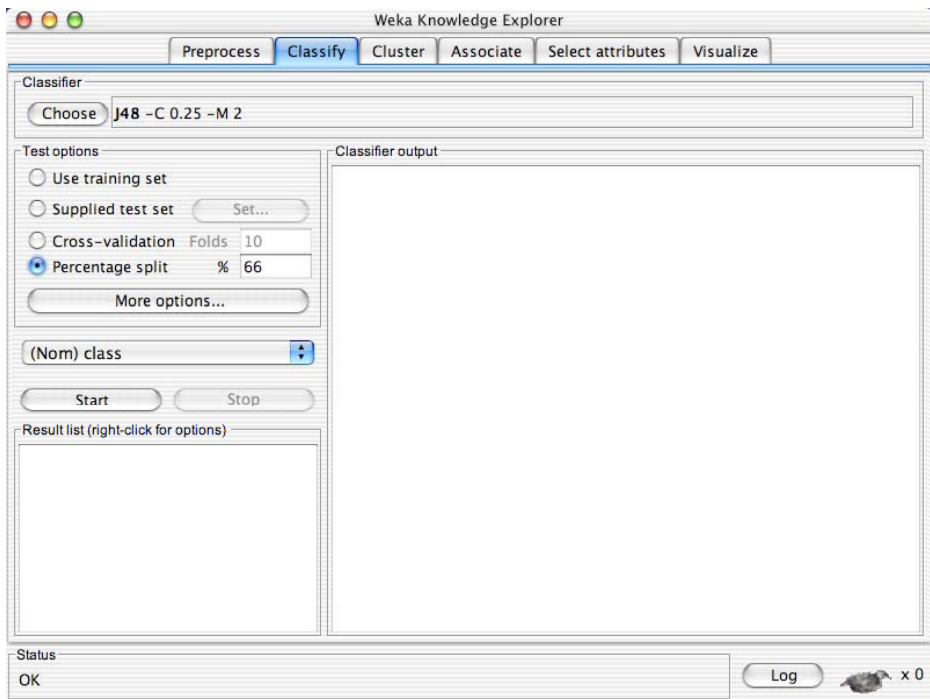
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, ...
- “Meta”-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, ...











Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options:
 

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:
 

```

    Time taken to build model: 0.24 seconds
    === Evaluation on test split ===
    === Summary ===
    Correctly Classified Instances      49      96.0784 %
    Incorrectly Classified Instances    2       3.9216 %
    Kappa statistic                    0.9408
    Mean absolute error                 0.0396
    Root mean squared error            0.1579
    Relative absolute error             8.8979 %
    Root relative squared error        33.4091 %
    Total Number of Instances          51

    === Detailed Accuracy By Class ===
    TP Rate  FP Rate  Precision  Recall  F-Measure  Class
    1         0         1          1       1          Iris-setosa
    1         0.063    0.905     1       0.95      Iris-versicolor
    0.882    0         1          0.882  0.938     Iris-virginica
    
```

Result list (right-click for options):
 

- 11:49:05 - trees.j48J48

Status: OK

Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options:
 

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:
 

```

    Time taken to build model: 0.24 seconds
    === Evaluation on test split ===
    === Summary ===
    Correctly Classified Instances      49      96.0784 %
    Incorrectly Classified Instances    2       3.9216 %
    Kappa statistic                    0.9408
    Mean absolute error                 0.0396
    Root mean squared error            0.1579
    Relative absolute error             8.8979 %
    Root relative squared error        33.4091 %
    Total Number of Instances          51

    === Detailed Accuracy By Class ===
    TP Rate  FP Rate  Precision  Recall  F-Measure  Class
    1         0         1          1       1          Iris-setosa
    1         0.063    0.905     1       0.95      Iris-versicolor
    0.882    0         1          0.882  0.938     Iris-virginica
    
```

Result list (right-click for options):
 

- 11:49:05 - trees.j48J48

Status: OK

Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options:
 

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

Classifier output:
 

```

    Time taken to build model: 0.24 seconds
    === Evaluation on test split ===
    === Summary ===
    Correctly Classified Instances      49      96.0784 %
    Incorrectly Classified Instances    2       3.9216 %
    Kappa statistic                    0.9408
    Mean absolute error                 0.0396
    Root mean squared error            0.1579
    Relative absolute error             8.8979 %
    Root relative squared error        33.4091 %
    Total Number of Instances          51

    === Detailed Accuracy By Class ===
    TP Rate  FP Rate  Precision  Recall  F-Measure  Class
    1         0         1          1       1          Iris-setosa
    1         0.063    0.905     1       0.95      Iris-versicolor
    0.882    0         1          0.882  0.938     Iris-virginica
    
```

Result list (right-click for options):
 

- 11:49:05 - trees.j48J48

Context menu:
 

- View in main window
- View in separate window
- Save result buffer
- Load model
- Save model
- Re-evaluate model on current test set
- Visualize classifier errors
- Visualize tree**
- Visualize margin curve
- Visualize threshold curve
- Visualize cost curve

Status: OK

Weka Knowledge Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Weka Classifier Tree Visualizer: 11:49:05 - trees.j48J48 (iris)

Tree View:

```

    graph TD
      A[petalwidth] -- "<= 0.6" --> B[Iris-setosa (50.0)]
      A -- "> 0.6" --> C[petalwidth]
      C -- "<= 1.7" --> D[petalwidth]
      C -- "> 1.7" --> E[Iris-virginica (46.0/1.0)]
      D -- "<= 4.9" --> F[Iris-versicolor (48.0/1.0)]
      D -- "> 4.9" --> G[petalwidth]
      G -- "<= 1.5" --> H[Iris-virginica (3.0)]
      G -- "> 1.5" --> I[Iris-versicolor (3.0/1.0)]
    
```

Result list (right-click for options):
 

- 11:49:05 - trees.j48J48

Status: OK

## Homework #1 – Due Feb. 11

---

- Analyze the zoo dataset from the UCI repository using the Weka Explorer.
  - For each of the attributes feathers, predators, tail, and domestic, report on the types and numbers of animals having the attribute true.
  - Remove instances whose “type” attribute is larger than or equal to 4. Use the classifier J48graft to derive the corresponding decision tree. Draw the corresponding tree.
  - Use the rules classifier PART to derive the rules on the zoo dataset. List the rules obtained.
  - Remove the “type” attribute from the dataset and run the default clustering algorithm SimpleKMeans. How many clusters do you obtain? Can you relate these clusters to the initial class values?