

# Attribute Discretization and Selection

## Clustering

NIKOLA MILIKIĆ

[nikola.milikic@fon.bg.ac.rs](mailto:nikola.milikic@fon.bg.ac.rs)

UROŠ KRČADINAC

[uros@krcadinac.com](mailto:uros@krcadinac.com)

# *Naive Bayes Features*

- Intended primarily for the work with nominal attributes
- In case of numeric attributes
  - Use the probability distribution of attributes (Normal distribution is default) for probability estimation for the each attribute
  - Discretize the attribute's values

# Attribute Discretization

**Discretization** is the process of transformation numeric data into nominal data, by putting the numeric values into distinct groups, which length is fixed.

Common approaches:

- Unsupervised:
  - Equal-width binning
  - Equal-frequency binning
- Supervised – classes are taken into account

# Equal-Width Binning

**Equal-width binning** divides the scope of possible values into  $N$  subsopes (bins) of the same width:

$$\text{width} = (\text{max value} - \text{min value}) / N$$

Example: If the scope of the values is between 0 and 100, we should create 5 subsopes (bins) in the following manner:

$$\text{Width} = (100 - 0) / 5 = 20$$

Subscopes (bins): [0-20], (20-40], (40-60], (60-80], (80-100]

Usually, the first and the final subscope (bin) are being expended in order to include possible values outside the original scope.

# Equal-frequency binning

**Equal-frequency binning** (or equal-height binning) divides the scope of possible values into N subsopes where each subscope (bin) carries the same number of instances:

Example: We want to put the following values in 5 subsopes (bins):

5, 7, 12, 35, 65, 82, 84, 88, 90, 95

So, each subscope will have 2 instances:

5, 7, | 12, 35, | 65, 82, | 84, 88, | 90, 95

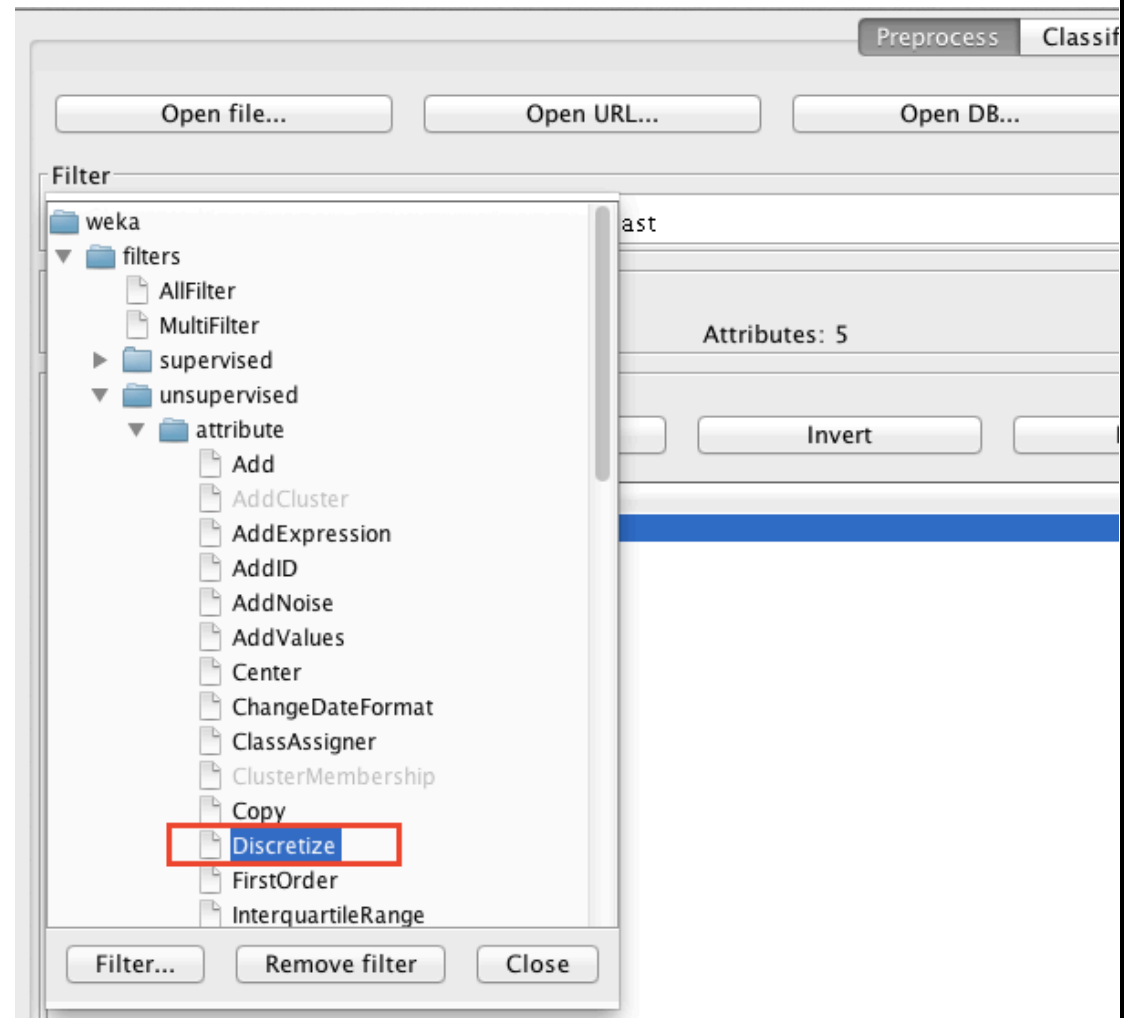
# Discretization in Weka

We apply certain *Filters* to attributes we want to discretize.

*Preprocess* tab

Option: *Choose -> Filter*

*filters / unsupervised / attribute*  
***Discretize.***

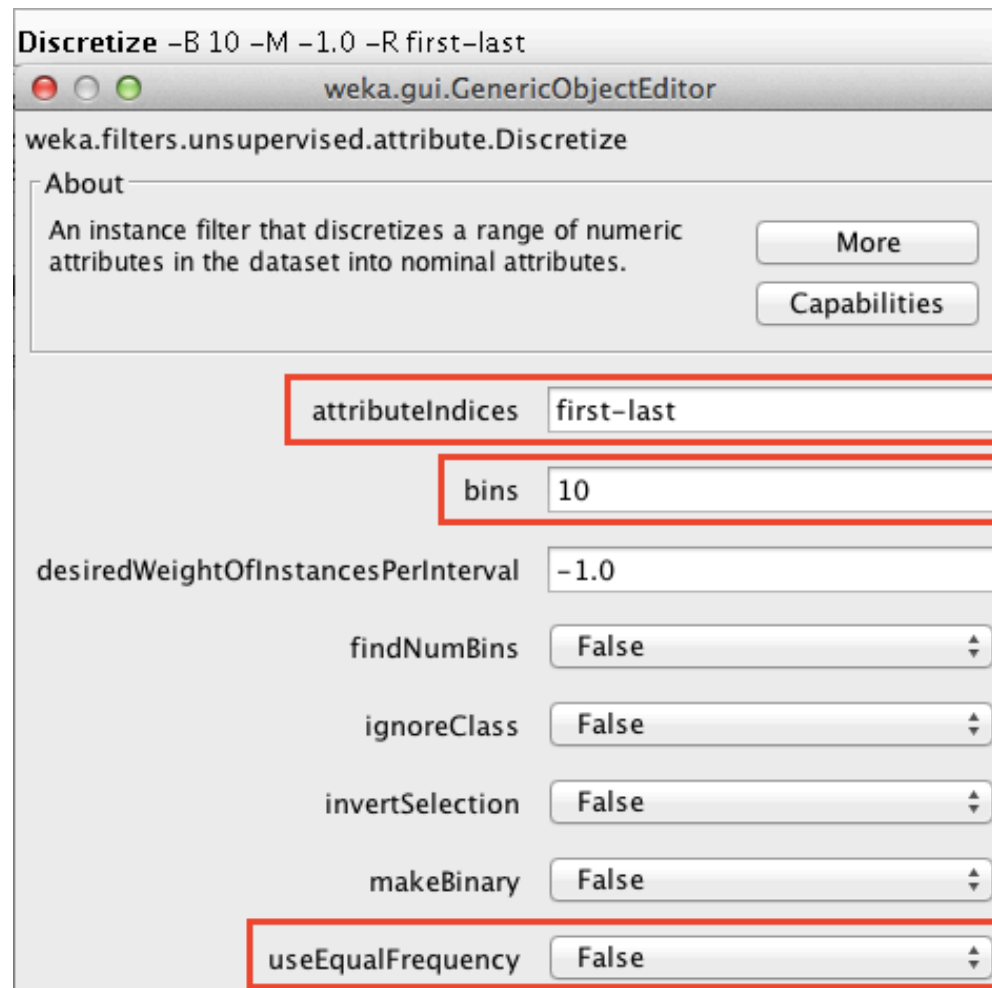


FishersIrisDataset.arff

# Discretization in Weka

Equal-width binning is the default option.

- *attributeIndices* – the *first-last* value means that we are discretizing all values. We can also name the attribute numbers.
- *bins* – the desired number of scopes (bins)
- *useEqualFrequency* – *false* by default; *true* if we use Equal Frequency binning



# Discretization in Weka

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

Open file... Open URL... Open DB... Generate... Undo

Filter: Choose **Discretize -B 10 -M -1.0 -R first-last** **Apply**

Current relation  
Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove...  
Instances: 150 Attributes: 5

Attributes  
All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	Sepal Length
<input type="checkbox"/>	Sepal Width
<input type="checkbox"/>	Petal Length
<input type="checkbox"/>	Petal Width
<input type="checkbox"/>	Species

Remove

Selected attribute  
Name: Sepal Length  
Missing: 0 (0%) Distinct: 10  
Type: Nominal  
Unique: 0 (0%)

No.	Label	Count
1	'(-inf-4.66]'	9
2	'(4.66-5.02]'	23
3	'(5.02-5.38]'	14
4	'(5.38-5.74]'	27
5	'(5.74-6.1]'	22
6	'(6.1-6.46]'	20
7	'(6.46-6.82]'	18
8	'(6.82-7.18]'	6
9	'(7.18-7.54]'	5
10	'(7.54-inf]'	6

Class: Species (Nom) Visualize All

Bin	Setosa (Red)	Versicolour (Blue)	Virginica (Cyan)	Total
1	0	0	9	9
2	0	23	0	23
3	14	0	0	14
4	0	27	0	27
5	0	22	0	22
6	0	20	0	20
7	0	18	0	18
8	6	0	0	6
9	5	0	0	5
10	6	0	0	6

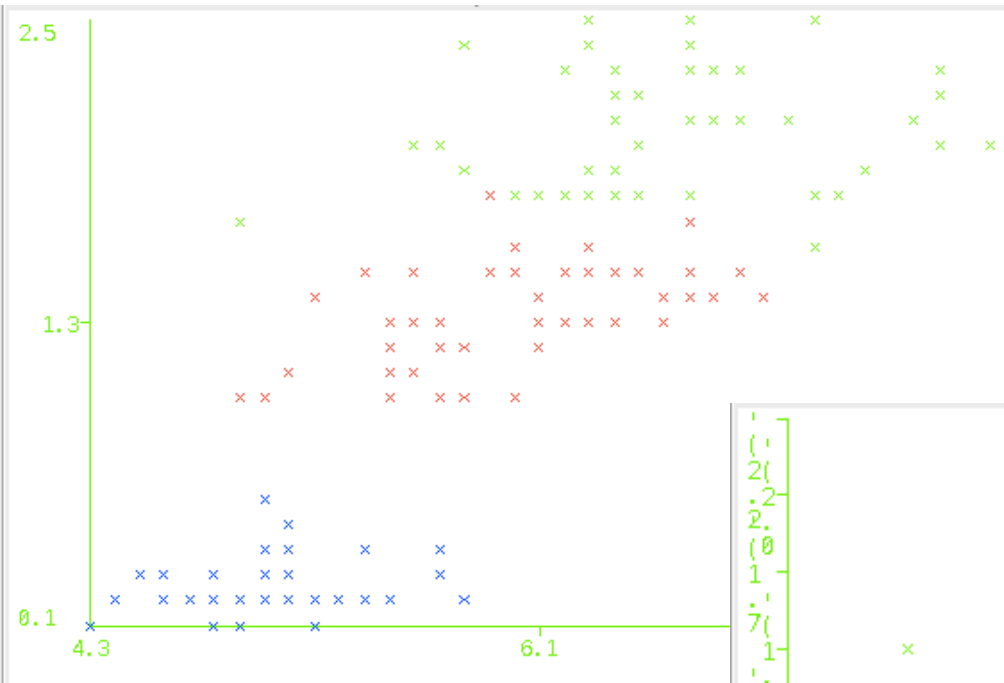
Applying the filter

The resulting subscopes (bins)

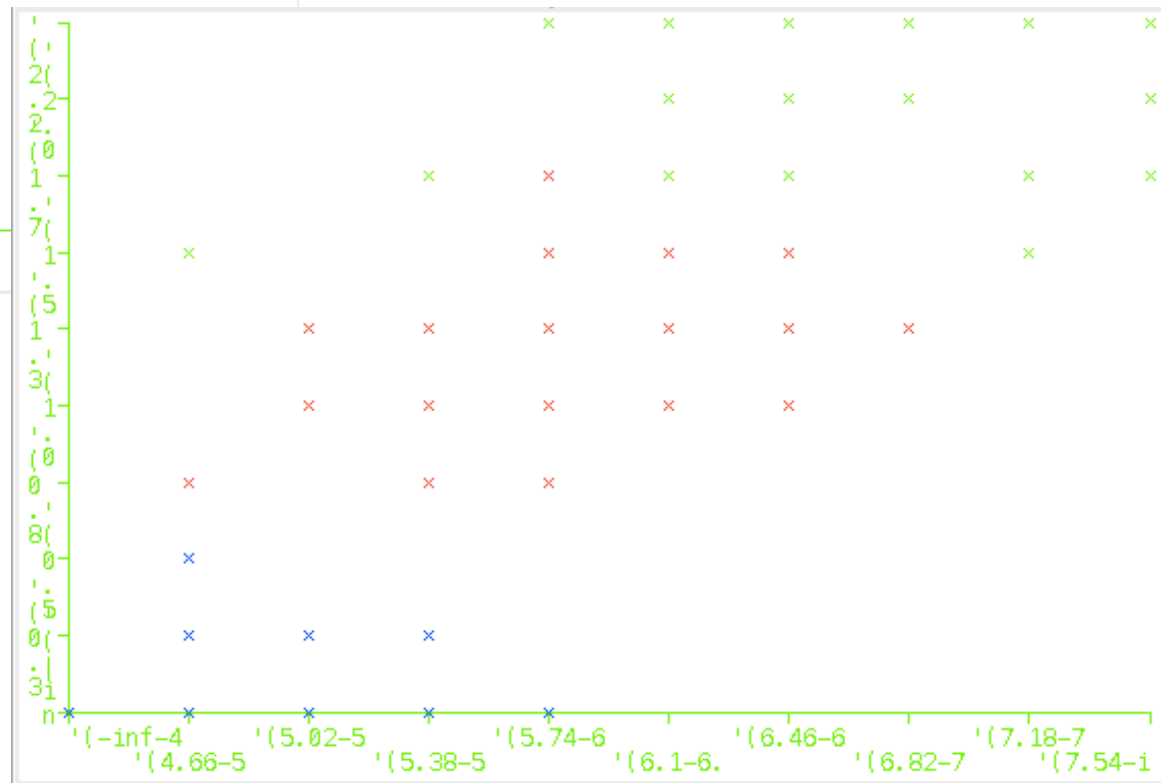
Status: OK Log x 0



# Data, before and after discretization



After



# Attribute Selection

**Attribute Selection** (or Feature Selection) is the process of choosing a subset of relevant attributes that will be used during the further analysis.

It is being applied in cases where the dataset contains attributes which are redundant and/or irrelevant.

- Redundant attributes are the ones that do not provide more information than the attributes we already have in our dataset.
- Irrelevant attributes are the ones that are useless in the context of the current analysis.

# Attribute Selection Advantages

Excessive attributes can degrade the performance of the model.

Advantages:

- Advances the readability of the model (because now the model contains only the relevant attributes)
- Shortens the training time
- Generalization power is higher because it lowers the possibility of overfitting

If the problem is well-known, the best way to select attribute is to do it manually. However, automated approaches also give good results.

# Approaches to Attribute Selection

Two approaches:

- *Filter* method – use the approximation based on the general features of the data.
- *Wrapper* method – attribute subsets are being evaluated by using the machine learning algorithm, applied to the dataset. The name *Wrapper* comes from the fact that the algorithm is wrapped within the process of selection. The chosen subset of attributes is the one for which the algorithm gives the best results.

# Attribute Selection Example

census90-income.arff

The screenshot shows the Weka Attribute Selection interface. The 'Preprocess' tab is active. The 'Filter' is set to 'None'. The 'Current relation' is '1990census' with 32561 instances and 15 attributes. The 'Attributes' list includes: 1 age, 2 workclass, 3 fnlwgt, 4 education, 5 education-num, 6 marital-status, 7 occupation, 8 relationship, 9 race, 10 sex, 11 capital-gain, 12 capital-loss, 13 hours-per-week, 14 native-country, and 15 income. The 'age' attribute is selected. The 'Selected attribute' panel shows: Name: age, Type: Numeric, Missing: 0 (0%), Distinct: 73, Unique: 2 (0%). A table of statistics for 'age' is shown below: Minimum: 17, Maximum: 90, Mean: 38.582, StdDev: 13.64. The 'Class' is set to 'income (Nom)'. A histogram shows the distribution of 'age' values, with a blue area under the bars and red bars on top. The x-axis ranges from 17 to 90. The status bar shows 'OK' and a 'Log' button.

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

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

Filter: Choose None Apply

Current relation: Relation: 1990census, Instances: 32561, Attributes: 15

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> income

Remove

Selected attribute: Name: age, Type: Numeric, Missing: 0 (0%), Distinct: 73, Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom) Visualize All

Status: OK Log x 0

# Attribute Selection Example

The screenshot displays the WEKA software interface. The top menu bar includes 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below the menu bar are buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', and 'Edit...'. The 'Filter' panel on the left shows a tree view of filters under the 'weka' folder. The 'AttributeSelection' filter is highlighted with a red box. A yellow callout box points to this filter with the text: "We want to apply the selection of attributes".

The right panel shows the 'Selected attribute' statistics for 'age':

Selected attribute	
Name: age	Type: Unique
Missing: 0 (0%)	Distinct: 73
Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Below the statistics, the class is set to 'income (Nom)'. At the bottom right, there is a bar chart showing the distribution of the selected attribute 'age' across the class 'income (Nom)'. The bars are stacked with blue and red colors, representing the distribution of 'age' values for different income levels.

# Attribute Selection Example

The screenshot displays the Weka GUI's 'Attribute Selection' configuration window. The window title is 'AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka'. The main content area shows the class 'weka.filters.supervised.attribute.AttributeSelection' with an 'About' section describing it as a supervised attribute filter. Below this, there are 'More' and 'Capabilities' buttons. A file browser is open, showing the 'weka' directory structure with 'attributeSelection' expanded. The file 'ClassifierSubsetEval' is highlighted with a red box. A yellow callout box points to this file with the text: '**ClassifierSubsetEval** is our choice for the evaluator'. The background shows a list of attributes with checkboxes, including 'age', 'work', 'fn', 'ec', 'ec', 'm', 'oc', 'relationship', 'race', 'sex', 'capital-gain', and 'capital-loss'.

Filter

Choose

AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka

weka.gui.GenericObjectEditor

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

More

Capabilities

evaluator

search

Open...

weka

- attributeSelection
  - CfsSubsetEval
  - ChiSquaredAttributeEval
  - ClassifierSubsetEval**
  - ConsistencySubsetEval
  - CostSensitiveAttributeEval
  - CostSensitiveSubsetEval
  - FilteredAttributeEval

**ClassifierSubsetEval** is our choice for the evaluator

No.	Attribute	Selected
1	age	<input type="checkbox"/>
2	work	<input type="checkbox"/>
3	fn	<input type="checkbox"/>
4	ec	<input type="checkbox"/>
5	ec	<input type="checkbox"/>
6	m	<input type="checkbox"/>
7	oc	<input type="checkbox"/>
8	relationship	<input type="checkbox"/>
9	race	<input type="checkbox"/>
10	sex	<input type="checkbox"/>
11	capital-gain	<input type="checkbox"/>
12	capital-loss	<input type="checkbox"/>

# Attribute Selection Example

The screenshot displays the Weka GUI's Attribute Selection interface. A 'Filter' window is open, showing the configuration for 'AttributeSelection'. The main window lists 15 attributes with checkboxes, including 'age', 'work-week-hours', 'fnlwgt', 'education', 'education-num', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native-country', and 'income'. A 'ClassifierSubsetEval' dialog is also open, showing the 'classifier' field set to 'NaiveBayes'. A yellow callout box points to this field with the text 'NaiveBayes classifier'.

**Filter**

Choose **AttributeSelection** -E "weka.attributeSelection.ClassifierSubsetEval" -B weka.classifiers.bayes.NaiveBayes

Current relation: weka.gui.GenericObjectEditor  
Relation: 1  
Instances: 3

**Attributes**

All

No. | Name | Selected

- 1  age
- 2  work-week-hours
- 3  fnlwgt
- 4  education
- 5  education-num
- 6  marital-status
- 7  occupation
- 8  relationship
- 9  race
- 10  sex
- 11  capital-gain
- 12  capital-loss
- 13  hours-per-week
- 14  native-country
- 15  income

Remove

**weka.gui.GenericObjectEditor**  
weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

More

Capabilities

evaluator Choose **ClassifierSubsetEval** -B weka.classifiers.bayes

search Choose

Open... Save...

**weka.gui.GenericObjectEditor**  
weka.attributeSelection.ClassifierSubsetEval

About

Classifier subset evaluator:

Evaluates attribute subsets on training data or a separate hold out testing set.

classifier Choose **NaiveBayes**

holdOutFile Click to set hold out or test instances

useTraining True

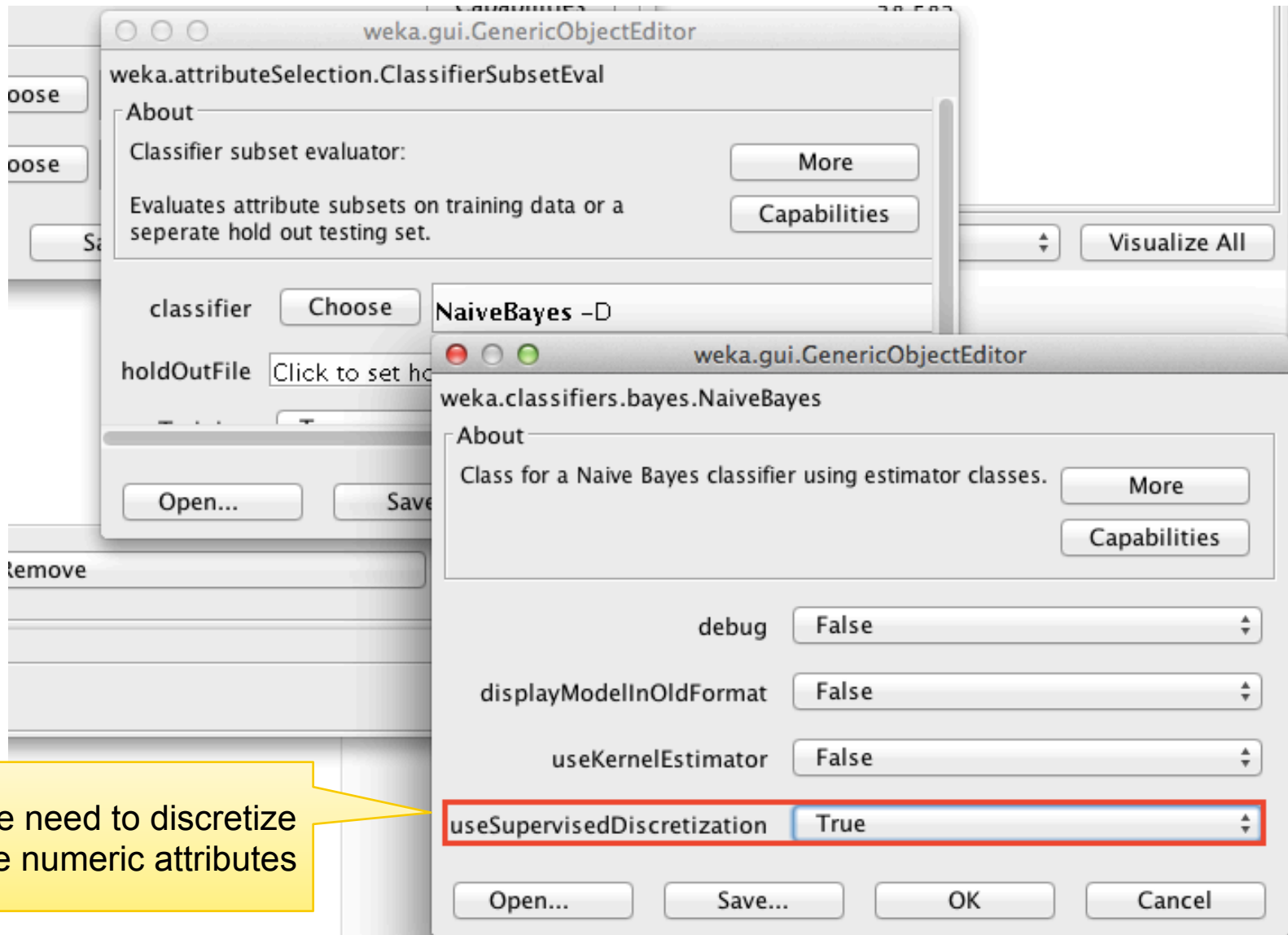
Open... Save... OK Cancel

Distinct: 73

Value
17
90
38.582
13.64



# Attribute Selection Example



We need to discretize the numeric attributes

# Attribute Selection Example

The image shows a screenshot of the Weka GUI's Attribute Selection dialog box. The dialog is titled "AttributeSelection" and is part of the "weka.gui.GenericObjectEditor". It displays the class name "weka.filters.supervised.attribute.AttributeSelection" and an "About" section describing it as a supervised attribute filter. The "evaluator" is set to "ClassifierSubsetEval" and the "search" method is set to "BestFirst -D 1 -N 5". A red box highlights the "search" field and its "Choose" button. A yellow callout box points to the "BestFirst" text, stating: "As the search method we choose the **BestFirst**". The background shows a list of attributes with checkboxes, including "No.", "No.", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", and "12".

Filter

Choose

Current relation

Relation: 1

Instances: 3

Attributes

All

No. | N

1 |  | ac

2 |  | w

3 |  | fr

4 |  | ec

5 |  | ec

6 |  | m

7 |  | oc

8 |  | relationship

9 |  | race

10 |  | sex

11 |  | capital-gain

12 |  | capital-loss

weka.gui.GenericObjectEditor

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

More

Capabilities

evaluator Choose ClassifierSubsetEval -B weka

search Choose **BestFirst -D 1 -N 5**

Open... Save... OK

As the search method we choose the **BestFirst**

Bar	Color	Height
1	Blue	Low
2	Blue	High
3	Blue	Low
4	Red	High
5	Red	High

# Attribute Selection Example

The screenshot shows the Weka Attribute Selection interface. The 'Filter' field contains the command: `AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \{Click`. The 'Apply' button is highlighted with a red box. A yellow callout bubble points to the 'Apply' button with the text: "Filter is set and can be applied".

**Current relation**  
Relation: 1990census  
Instances: 32561  
Attributes: 15

**Attributes**

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> income

**Selected attribute**  
Name: age  
Missing: 0 (0%)  
Distinct: 73  
Type: Numeric  
Union

**Statistics**  
Minimum  
Maximum  
Mean  
StdDev

**Class: income (Nom)** Visualize All

**Histogram**  
The histogram shows the distribution of the 'income' class. The x-axis ranges from 17 to 90, with a major tick at 53.5. The y-axis represents frequency. The bars are stacked with blue and red colors, showing a right-skewed distribution.

**Status**  
OK

Log x 0

# Attribute Selection Example

The screenshot displays the Weka software interface for attribute selection. The 'Preprocess' tab is active, and the 'Attribute Selection' filter is applied. The current relation is '1990census-weka.filters.supervised.attribute.Attribute...', containing 32561 instances and 7 attributes. The 'Attributes' list shows 'age' selected (indicated by a red box and a yellow callout). The 'Selected attribute' panel shows 'age' is numeric with 73 distinct values and 2 unique values. A histogram at the bottom right shows the distribution of 'age' values, ranging from 17 to 90.

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

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

Filter: Choose `AttributeSelection -E "weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H \\"Click` Apply

Current relation  
Relation: 1990census-weka.filters.supervised.attribute.Attribute...  
Instances: 32561 | Attributes: 7

Attributes  
All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> education
3	<input type="checkbox"/> relationship
4	<input type="checkbox"/> race
5	<input type="checkbox"/> capital-gain
6	<input type="checkbox"/> capital-loss
7	<input type="checkbox"/> income

The number of attributes is reduced to 7

Selected attribute  
Name: age | Type: Numeric  
Missing: 0 (0%) | Distinct: 73 | Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: income (Nom) | Visualize All

17 | 53.5 | 90

Status: OK | Log | x 0

# Clustering

**Clustering** belongs to a group of techniques of unsupervised learning. It enables grouping instances into groups, where we know which are the possible groups *in advance*.

These groups are called **clusters**.

As the result of clustering each instance is being added *a new attribute* – the cluster to which it belongs. The clustering is said to be successful if the final clusters make sense, if they could be given meaningful names.

# K-Means algorithm in Weka

FishersIrisDataset.arff

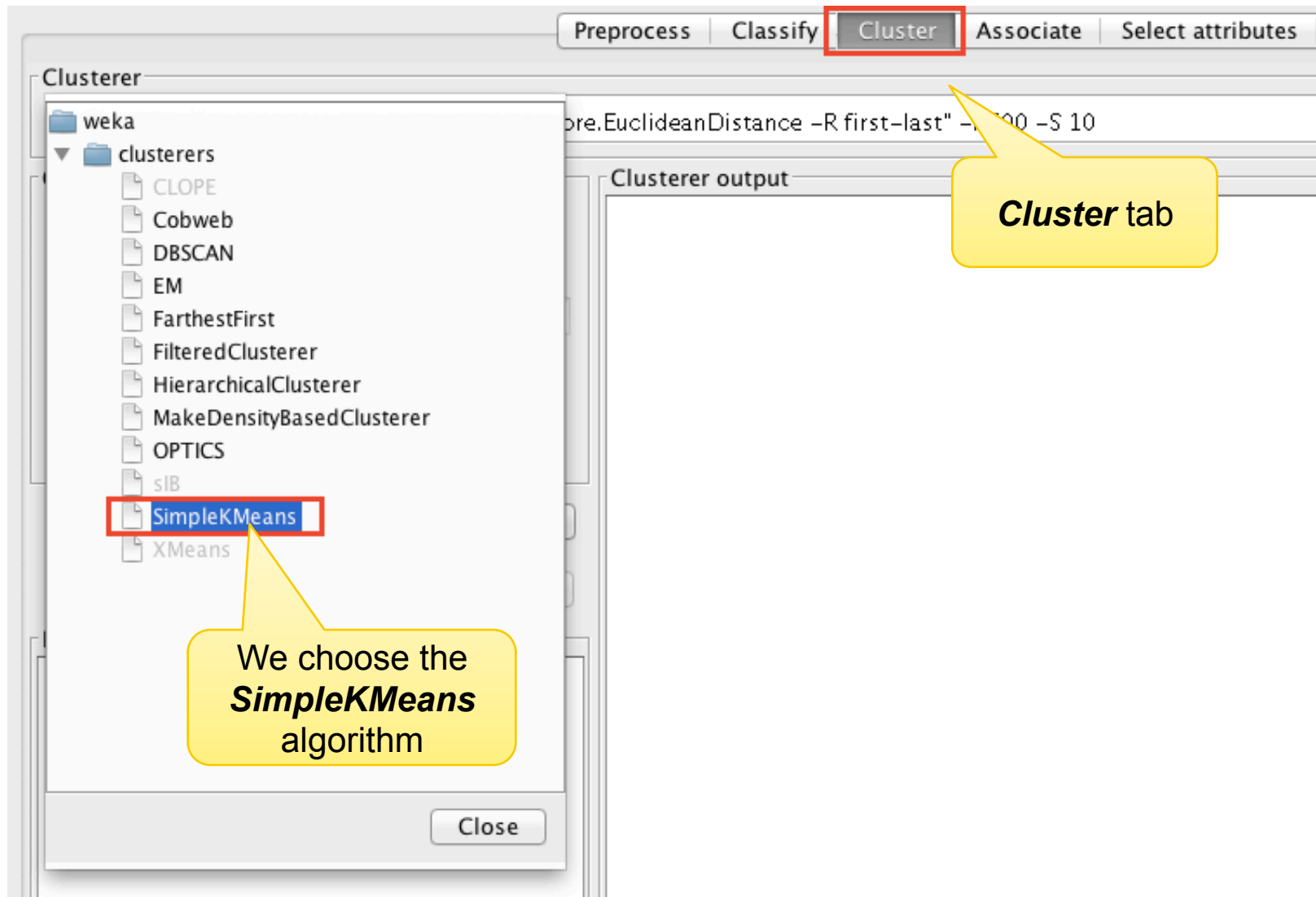
The screenshot shows the Weka software interface with the 'Preprocess' tab selected. The 'Current relation' is 'FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1' with 150 instances and 5 attributes. The 'Attributes' list includes 'Sepal Length', 'Sepal Width', 'Petal Length', 'Petal Width', and 'Species'. The 'Selected attribute' is 'Sepal Length', which is numeric with 35 distinct values and 9 unique values (6%). A histogram displays the distribution of 'Sepal Length' values, with three clusters highlighted in blue, red, and cyan. The histogram shows the following counts for each cluster: Blue (16), Red (30), and Cyan (34). The x-axis ranges from 4.3 to 7.9.

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Cluster	Count
Blue	16
Red	30
Cyan	34

Status: OK

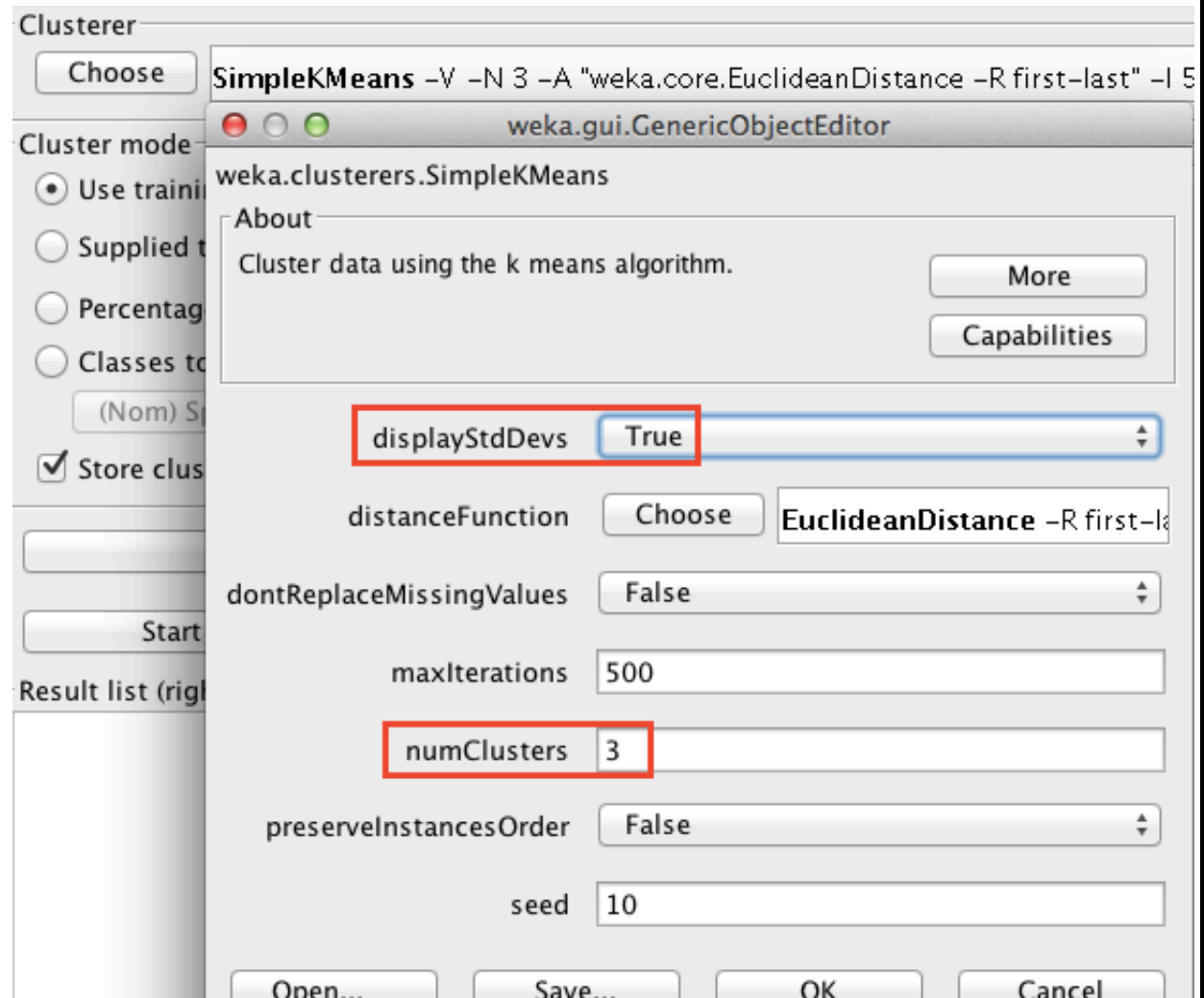
# Choosing the clustering algorithm



# Parameter settings

*numClusters* – the number of desired clusters;  
we set it to 3 because we have 3 kinds

*displayStdDevs* – if  
*true*, the standard  
deviation will be  
displayed





# Running the Clustering

The screenshot shows the Weka GUI for running a SimpleKMeans clustering. A yellow callout bubble points to the 'Use training set' radio button, stating 'Clustering over the imported data'. Another yellow callout bubble points to the 'Ignore attributes' text box, stating 'We ignore the *Species* attribute'. The 'Ignore attributes' text box contains the word 'Species'. The 'Store clusters for visualization' checkbox is checked. The 'Result list' window shows a list of attributes: Sepal Length, Sepal Width, Petal Length, Petal Width, and Species. The 'Species' attribute is highlighted in blue. The main output window shows the following text:

```
==== Run information ====  
Scheme:weka.clusterers.SimpleKMeans -V -N 3 -A "weka.core.Euc  
Relation: FishersIrisDataset-weka.filters.unsupervised.at  
Instances: 150  
Attributes: 5  
Sepal Length  
Sepal Width  
Petal Length  
Petal Width  
Ignored: Species  
Test mode:evaluate on training data  
==== Model and evaluation on training set ====  
Number of iterations: 6  
Within cluster sum of squared errors: 6.982216473785234  
Missing values globally replaced with mean/mode  
Cluster centroids:  
Attribute Full Data Cluster#  
(150) (61) (50) (39)  
-----  
Sepal Length 5.8433 5.8885 5.006 6.8462  
+/-0.8281 +/-0.4487 +/-0.3525 +/-0.5025  
Sepal Width 3.0573 2.7377 3.428 3.0821
```

# Results of Clustering

Cluster mode

Use training set  
 Supplied test set Set...  
 Percentage split % 66  
 Centroids of each cluster and their standard deviations

Ignore attributes

Start Stop

Result list (right-click for options)

15:07:08 - SimpleKMeans

Clusterer output

kMeans  
=====  
Number of iterations: 6  
Within cluster sum of squared errors: 6.982216473785234  
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (150)	Cluster#		
		0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	3.428 +/-0.3791	3.0821 +/-0.2799
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	1.462 +/-0.1737	5.7026 +/-0.5194
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Time taken to build model (full training data) : 0.04 seconds  
=== Model and evaluation on training set ===

Clustered Instances

0	61 ( 41%)
1	50 ( 33%)
2	39 ( 26%)

Number of instances in each cluster

# Evaluation of Results

The screenshot shows the Orange3 SimpleKMeans widget interface. On the left, the 'Cluster mode' section has 'Classes to clusters evaluation' selected, with '(Nom) Species' chosen in the dropdown. A red box highlights this dropdown. Below it is a 'Result list' showing two runs of SimpleKMeans. On the right, the 'Clusterer output' pane shows a table of statistics, a summary of time taken, and a 'Clustered Instances' table. A red box highlights the 'Class attribute: Species' and 'Classes to Clusters:' section, which shows a mapping of instances to clusters. A third red box highlights the 'Cluster 0 <-- versicolor' line. Yellow callout boxes provide context for these elements.

**Cluster mode**

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation**  
(Nom) Species
- Store clusters for visualization

**Result list (right-click for options)**

- 15:07:08 - SimpleKMeans
- 15:20:38 - SimpleKMeans**

**Clusterer output**

377	3.428	3.0821
34	+/-0.3791	+/-0.2799
67	1.462	5.7026
59	+/-0.1737	+/-0.5194
18	0.246	2.0795
23	+/-0.1054	+/-0.2811

Time taken to build model (full training data) : 0.02 seconds

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

**Clustered Instances**

0	61 ( 41%)
1	50 ( 33%)
2	39 ( 26%)

**Class attribute: Species**  
Classes to Clusters:

0	1	2	<-- assigned to cluster
0	50	0	setosa
47	0	3	versicolor
14	0	36	virginica

Cluster 0 <-- versicolor  
Cluster 1 <-- setosa  
Cluster 2 <-- virginica

Incorrectly clustered instances : 17.0 11.3333 %

**Annotations:**

- Select the attribute which we want to compare the results with.
- Which classes are in which clusters
- Names of classes which are given to clusters

# Visualization of Clusters

The image shows the Weka Clusterer Visualize interface. On the left, a control panel includes a checked option for "Store clusters for visualization", an "Ignore attributes" button, and a "Start" button. Below these is a "Result list" with two entries: "15:07:08 - SimpleKMeans" and "15:20:38 - SimpleKMeans", the latter being selected. A context menu is open over the selected entry, listing options such as "View in main window", "View in separate window", "Save result buffer", "Delete result buffer", "Load model", "Save model", "Re-evaluate model on current test set", "Visualize cluster assignments" (highlighted with a red box), and "Visualize tree".

The main window displays a scatter plot titled "Plot: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1". The X-axis is labeled "X: Petal Length (Num)" with values 1, 3.95, and 6.9. The Y-axis is labeled "Y: Petal Width (Num)" with values 0.1 and 1.3. The plot shows three distinct clusters of data points: cluster0 (red 'x' marks), cluster1 (blue squares), and cluster2 (green 'x' marks). A legend at the bottom of the window, titled "Class colour", shows "cluster0" in red, "cluster1" in blue, and "cluster2" in green.

Annotations include a yellow callout bubble pointing to the "Start" button with the text "Right click", and another yellow callout bubble pointing to the "Visualize cluster assignments" menu item with the text "Visual representation of clusters".

# Was clustering successful?

*Within cluster sum of squared error* gives us the assessment of quality

Cluster mode

Use training set

It is being counted as the sum of square differences between the value of the attribute of each instance and the value of the centroid of the given attribute

Start Stop

Result list (right-click for options)

- 15:07:08 - SimpleKMeans
- 15:20:38 - SimpleKMeans

Clusterer output

kMeans

====

Number of iterations: 6

**Within cluster sum of squared errors: 6.982216473785234**

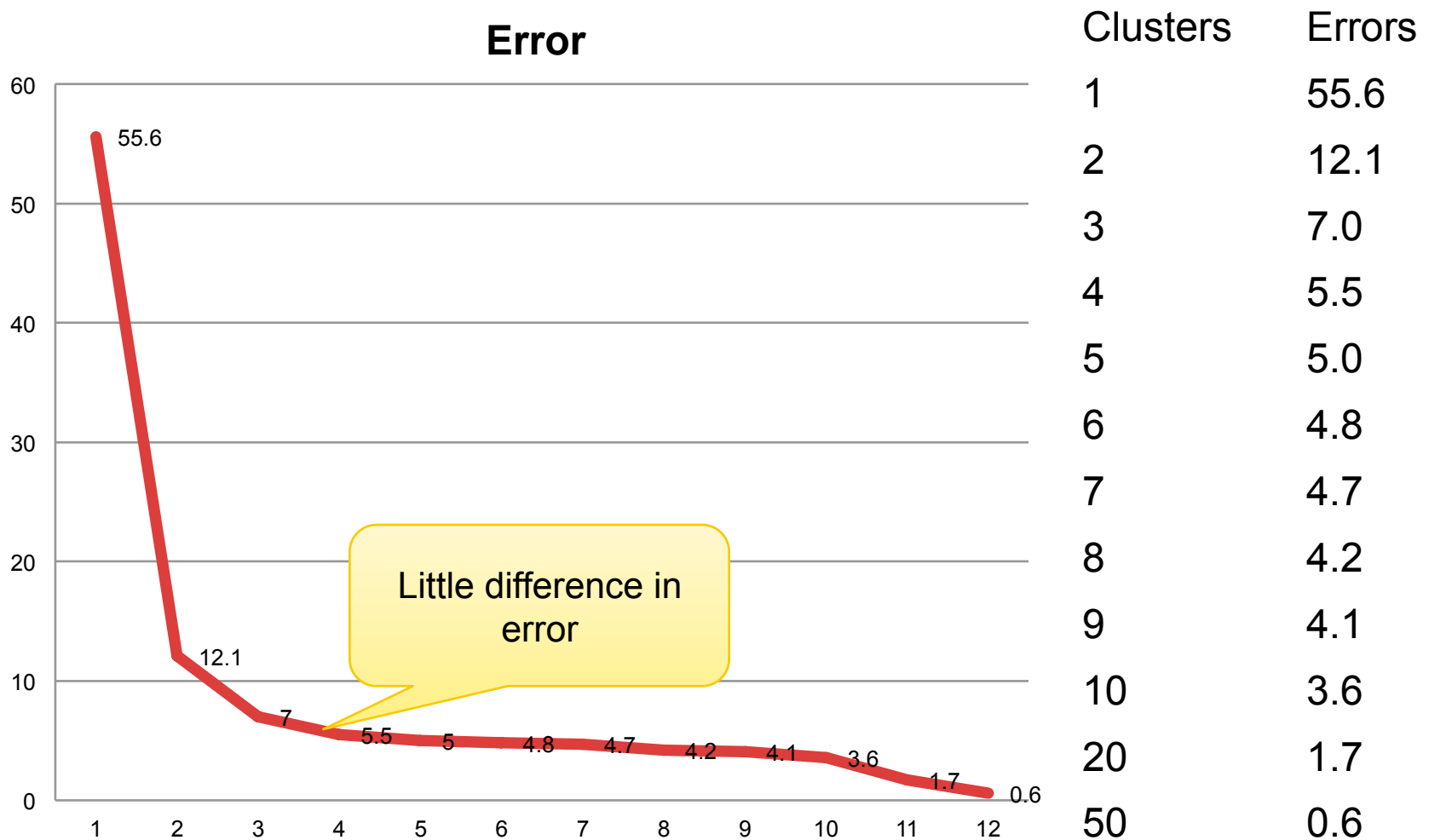
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	+/-0.3	
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	+/-0.1737	+/-0.5194
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Vrednosti centroida po svim atributima

# How to figure out the number of clusters?



# Using Clusters for Classification

The screenshot shows the Weka software interface with the 'Filter' dialog box open. The 'Filter' dialog has a tree view on the left showing the 'weka' directory structure. Under 'unsupervised', the 'attribute' folder is expanded, and 'AddCluster' is selected and highlighted with a red box. A yellow speech bubble points to 'AddCluster' with the text 'AddCluster – our choice of the filter'. In the main area of the dialog, the 'Class' dropdown menu is set to 'No class', also highlighted with a red box. A yellow speech bubble points to this dropdown with the text 'Setting no class'. The 'Filter...' button is at the bottom left of the dialog. The background shows the 'Preprocess' tab with 'SimpleKMeans' selected, and the 'Selected attribute' panel on the right showing 'Sepal Length'.

Preprocess | Classify | Cluster | Associate | Select attributes

Open file... | Open URL... | Open DB... | Generate... | Undo

Filter

- weka
  - filters
    - AllFilter
    - MultiFilter
  - supervised
  - unsupervised
    - attribute
      - Add
      - AddCluster**
      - AddExpression
      - AddID
      - AddNoise
      - AddValues
      - Center
      - ChangeDateFormat
      - ClassAssigner
      - ClusterMembership
      - Copy
      - Discretize
      - FirstOrder
      - InterquartileRange

SimpleKMeans -N 2 -A {"weka.core.EuclideanDistance -R first-last}"

Selected attribute  
Name: Sepal Length  
Missing: 0 (0%)

Statistic  
Minimum  
Maximum  
Mean  
StdDev

Invert | Pattern

**No class**

Filter... | Remove filter | Close

30 | 34

# Using Clusters for Classification

The image shows a screenshot of the Weka GUI with two windows. The main window is titled 'weka.gui.GenericObjectEditor' and shows the configuration for a 'clusterer' attribute. The 'clusterer' dropdown is set to 'SimpleKMeans'. The 'ignoredAttributeIndices' field is set to '5'. A yellow callout bubble points to the 'SimpleKMeans' selection with the text: 'We choose the *SimpleKMeans* as the clustering algorithm'. Another yellow callout bubble points to the 'ignoredAttributeIndices' field with the text: 'In terms of clustering, we ignore the attribute 5 (Speices)'. The second window is titled 'weka.gui.GenericObjectEditor' and shows the configuration for 'weka.clusterers.SimpleKMeans'. The 'numClusters' field is set to '3'. A red box highlights the 'numClusters' field.

Filter

Choose **AddCluster** -V "weka.clusterers.SimpleKMeans -V -N 3 -A {"weka.core.EuclideanDistance -R first-last}" -

Current relation: weka.gui.GenericObjectEditor

Relationship: Attribute.AddCluster

Attributes: Attribute representing distance by the specified

clusterer: Choose **SimpleKMeans** -V

ignoredAttributeIndices: **5**

Open... Save... OK

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs: True

distanceFunction: Choose Eu

dontReplaceMissingValues: False

maxIterations: 500

**numClusters: 3**

preserveInstancesOrder: False

seed: 10

Open... Save... C



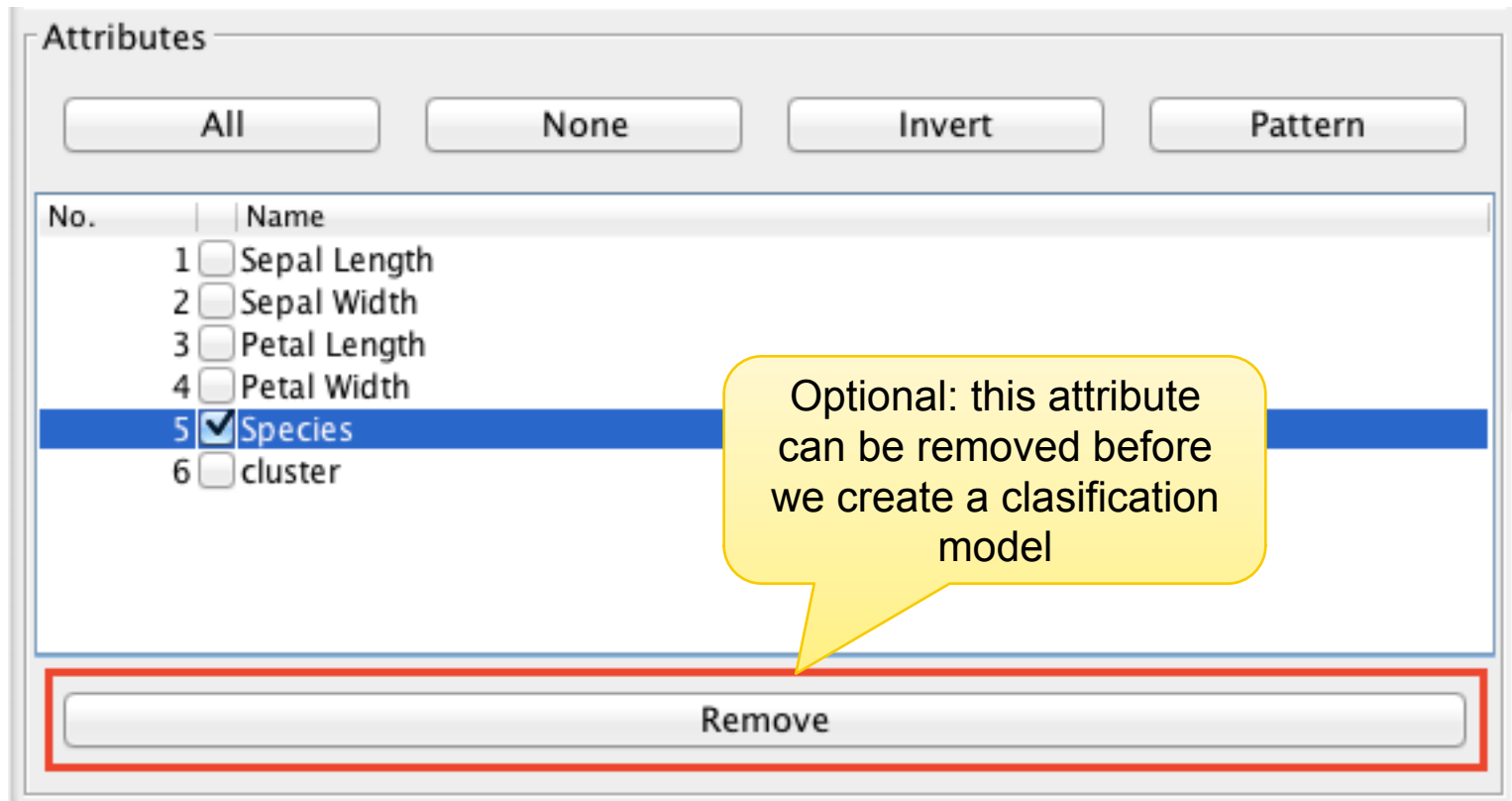
# Using Clusters for Classification

The screenshot shows the Weka software interface with the 'Cluster' tab selected. The 'Filter' field contains the command: `AddCluster -W "weka.clusterers.SimpleKMeans -V -N 3 -A {"weka.core.EuclideanDistance -R first-last}" -I 500 -S 10" -I 5`. The 'Apply' button is highlighted with a red box. The 'Current relation' is 'FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1-wek...' with 150 instances and 6 attributes. The 'Attributes' list shows 'cluster' selected, also highlighted with a red box. A yellow callout box points to the 'cluster' attribute with the text: 'After the filter is being applied (**Apply**) we add the new attribute by the name of **cluster**'. The 'Selected attribute' table shows the distribution of the cluster attribute:

No.	Label	Count
1	cluster1	61
2	cluster2	50
3	cluster3	39

The bar chart below the table visualizes the counts for each cluster: cluster1 (61), cluster2 (50), and cluster3 (39).

# Using Clusters for Classification



# Using Clusters for Classification

The screenshot shows the Orange3 software interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'NaiveBayes'. Under 'Test options', 'Use training set' is selected. The 'Cluster' dropdown is set to '(Nom) cluster'. The 'Result list' shows a single entry: '16:18:49 - bayes.NaiveBayes'. The main window displays the following statistics:

Pe  
mean 1.4108 0.2766 2.0806  
std. dev. 0.2837 0.1074 0.2717  
weight sum 61 50 39  
precision 0.1143 0.1143 0.1143

Time taken t

Correctly C 98.6667 %  
Incorrectly 1.3333 %  
Kappa stati .9796  
Mean absolu .0206  
Root mean s .0851  
Relative abs 4.7192 %  
Root relative squared error 18.209 %  
Total Number of Instances 150

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.968	0.022	0.968	1			cluster1
1	0	0	1				cluster2
0.949	0	1					cluster3
Weighted Avg.	0.987	0.009	0.987				

=== Confusion Matrix ===

a	b	c	<-- classified as
61	0	0	a = cluster1
0	50	0	b = cluster2
2	0	37	c = cluster3

Yellow callout boxes provide the following annotations:

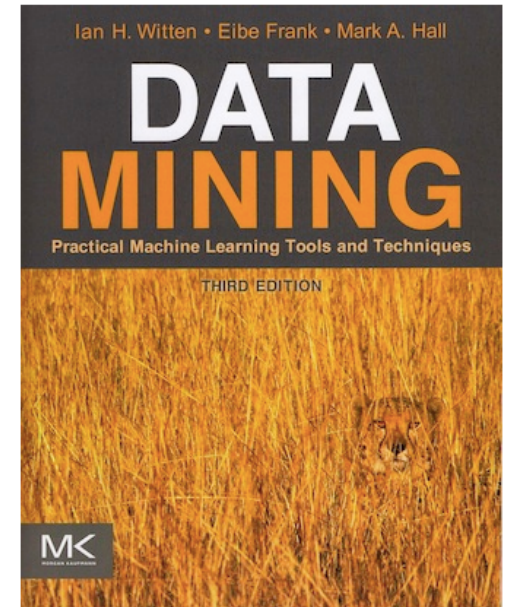
- "We use the **NaiveBayes** classifier" (pointing to the Classifier dropdown)
- "We do the classification according to the **cluster** attribute" (pointing to the Cluster dropdown)
- "The confusion matrix" (pointing to the Confusion Matrix table)

# Thank you notes

Weka Tutorials and Assignments @ The Technology Forge

- Link: <http://www.technologyforge.net/WekaTutorials/>

Witten, Ian H., Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.



A survey for you, to judge us :)

<http://goo.gl/cqdp3I>

# Any questions?

NIKOLA MILIKIĆ

[nikola.milicic@fon.bg.ac.rs](mailto:nikola.milicic@fon.bg.ac.rs)

UROŠ KRČADINAC

[uros@krcadinac.com](mailto:uros@krcadinac.com)