Data Mining - Clustering

Decision Supports Systems 2017/18, Lecture 08

Marko Tkalčič

Alpen-Adria-Universität Klagenfurt

Other Data-driven Approaches

- we used data to fit the distributions of the unknown variables

Other Data-driven Approaches

- we used data to fit the distributions of the unknown variables
- machine learning techniques are useful for making predictions
  - supervised:
    - regression (continuous)
    - classification (discrete)
  - unsupervised
    - clustering

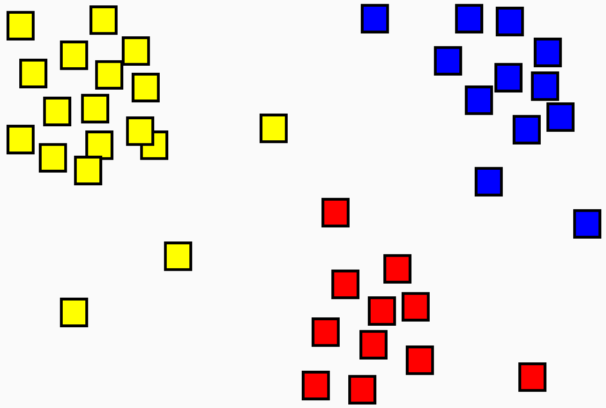
Table of Contents

Clustering

- grouping objects together into **clusters** (groups)
- object in the same cluster should be **more similar** than objects outside of the cluster
- **similarity** is key and requires contextual knowledge

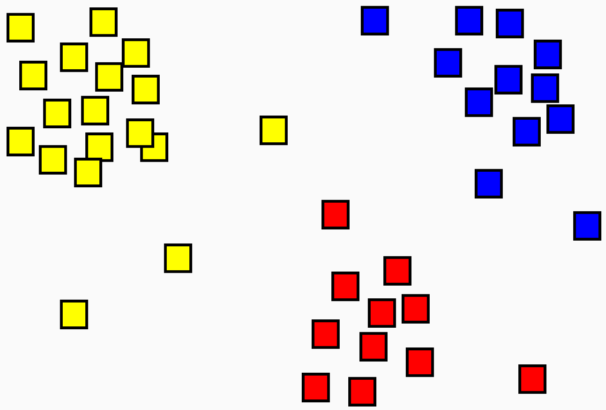
Clustering

- grouping objects together into **clusters** (groups)
- object in the same cluster should be **more similar** than objects outside of the cluster
- **similarity** is key and requires contextual knowledge



Clustering

- grouping objects together into **clusters** (groups)
- object in the same cluster should be **more similar** than objects outside of the cluster
- **similarity** is key and requires contextual knowledge



- there are several algorithms to achieve this task
  - k-means
  - hierarchical models
  - distribution-based models (GMM)

Table of Contents

Iris

- iris.arff (https://archive.ics.uci.edu/ml/datasets/iris)
- 3 classes of 50 instances each, where each class refers to a type of iris plant.
- One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Iris

- iris.arff (https://archive.ics.uci.edu/ml/datasets/iris)
- 3 classes of 50 instances each, where each class refers to a type of iris plant.
- One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Attribute Information:

- 1. sepal length in cm
- 2. sepal width in cm
- 3. petal length in cm
- 4. petal width in cm
- 5. class:
  - Iris Setosa
  - Iris Versicolour
  - Iris Virginica

Running Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- it has an easy GUI
- does not require knowledge of ML
- download Weka from https://www.cs.waikato.ac.nz/ml/weka/
- install

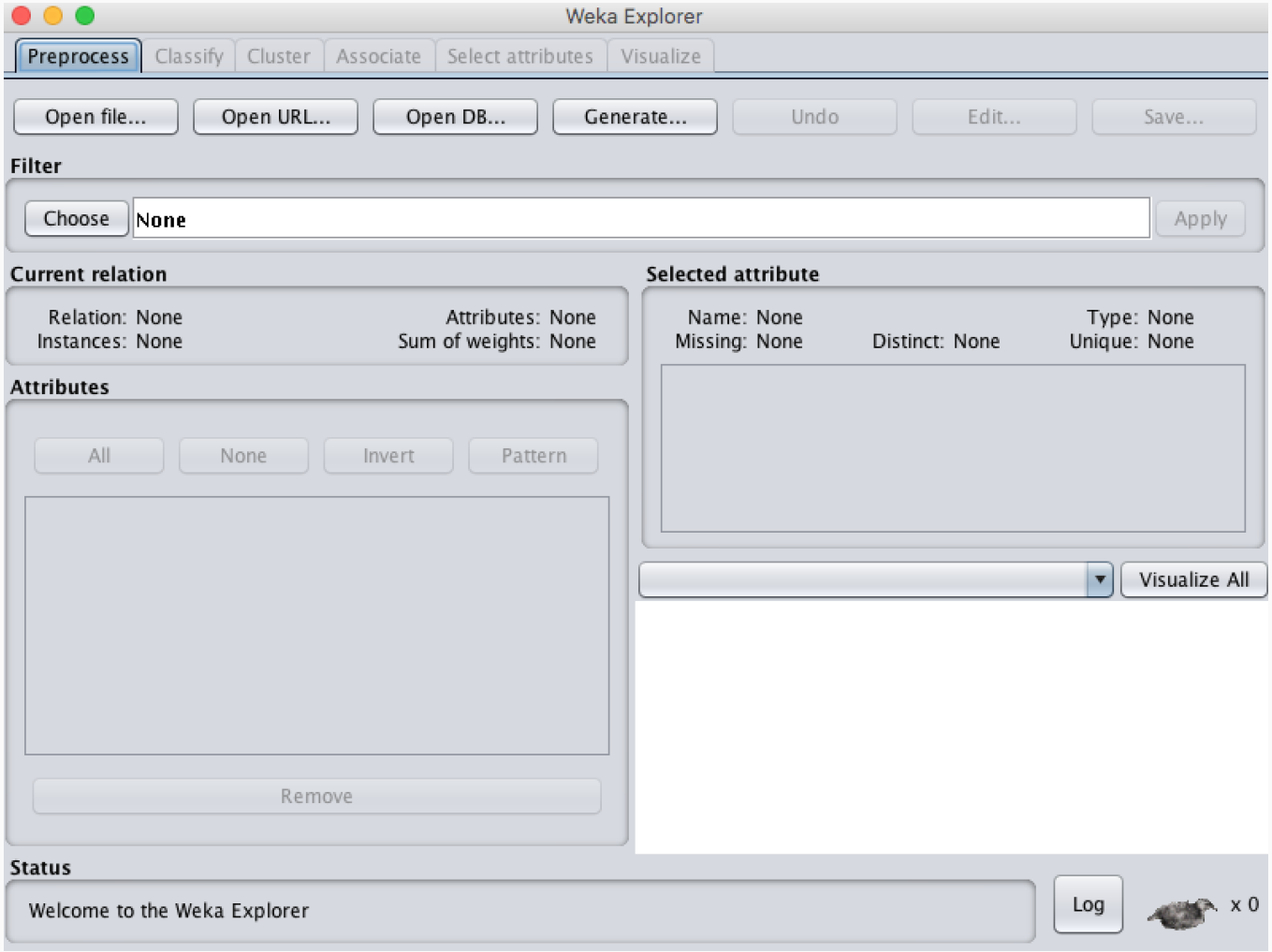
Running Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- it has an easy GUI
- does not require knowledge of ML
- download Weka from https://www.cs.waikato.ac.nz/ml/weka/
- install

- run and click *Explorer*

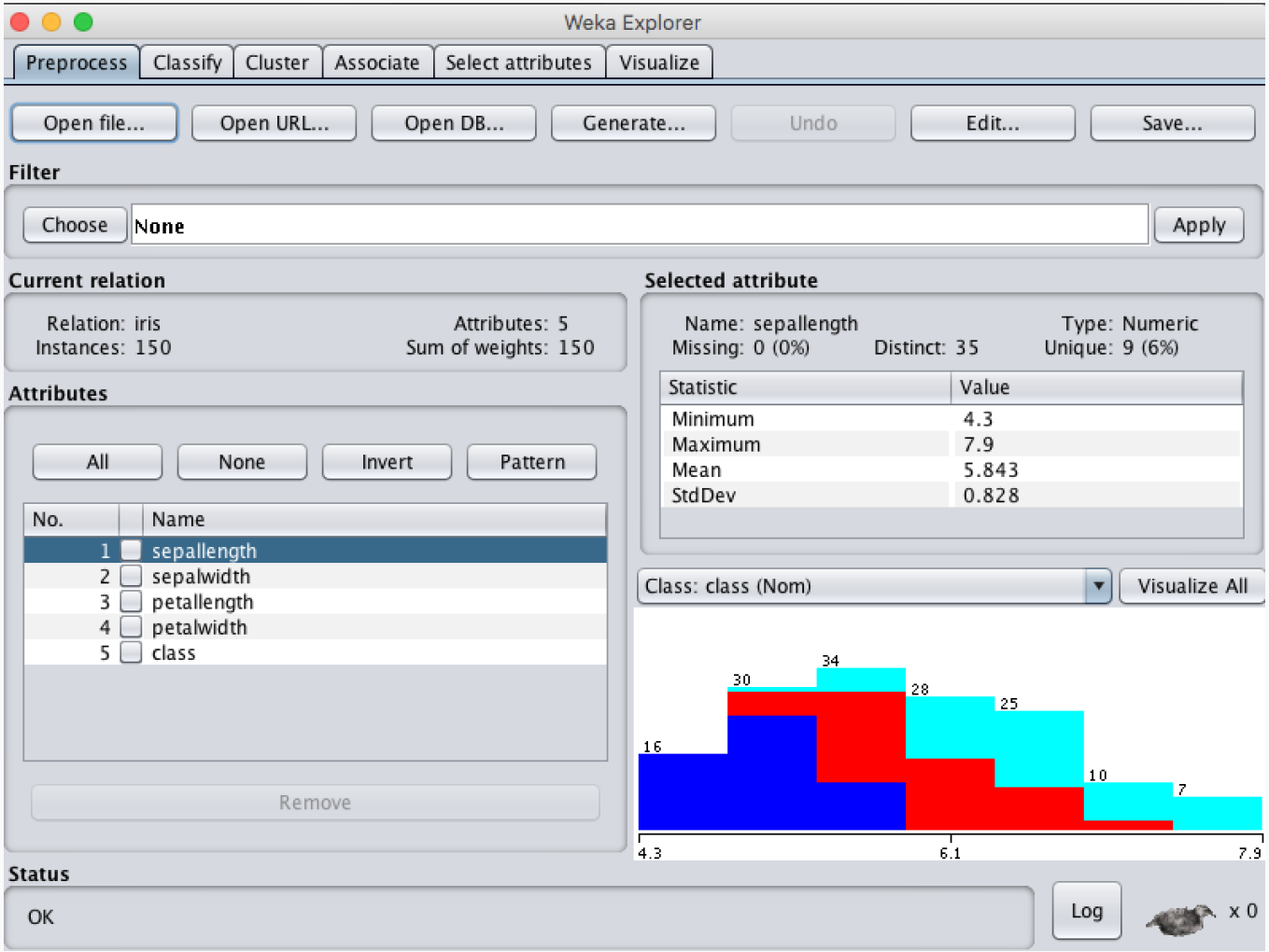
Explorer

- open file
- choose iris.arff

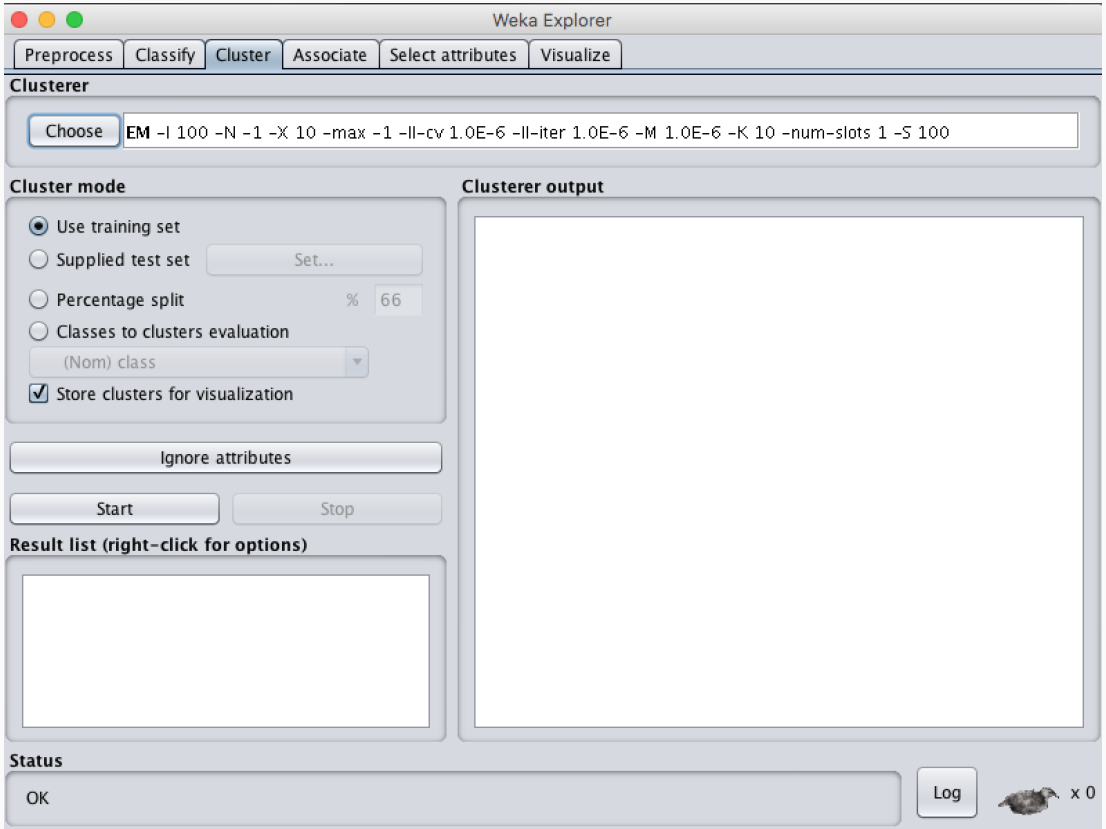


Pre-process

- you can observe the histogram of the variables
- click on Cluster

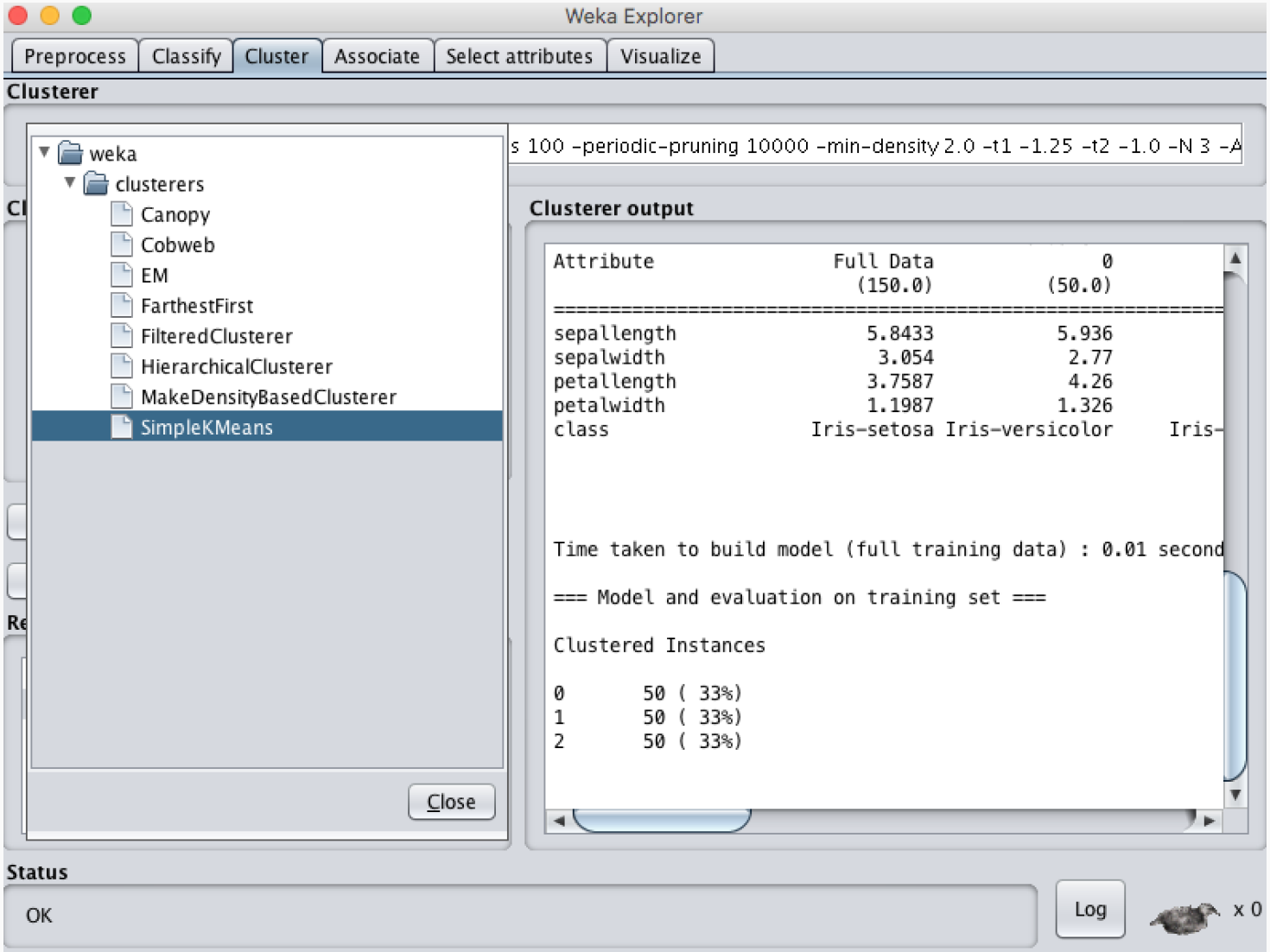


Cluster



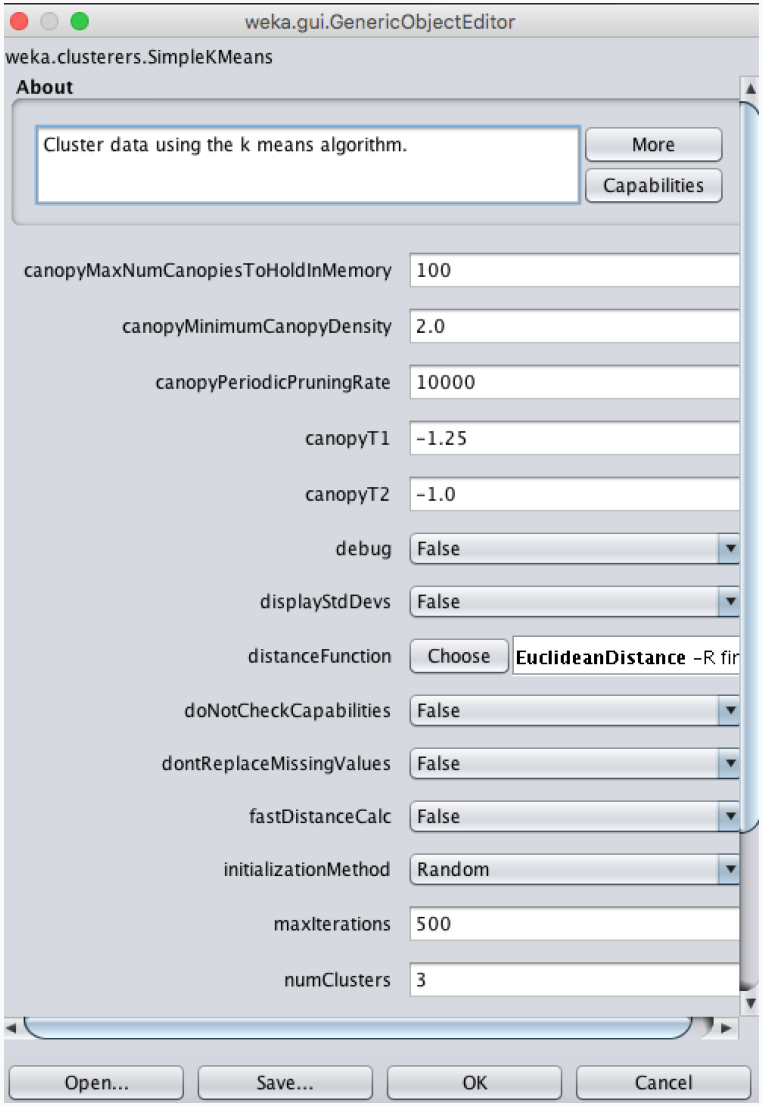
1. Choose clustering algorithm (click *choose* -> k-means)
2. Choose number of clusters (click on the long list of parameters)
3. Remove attributes for clustering (click on *Ignore attributes*)
4. Start
5. Visualize cluster assignments (right-click on the chosen result in the resultlist and click *Visualize cluster assignments*)

Algorithm

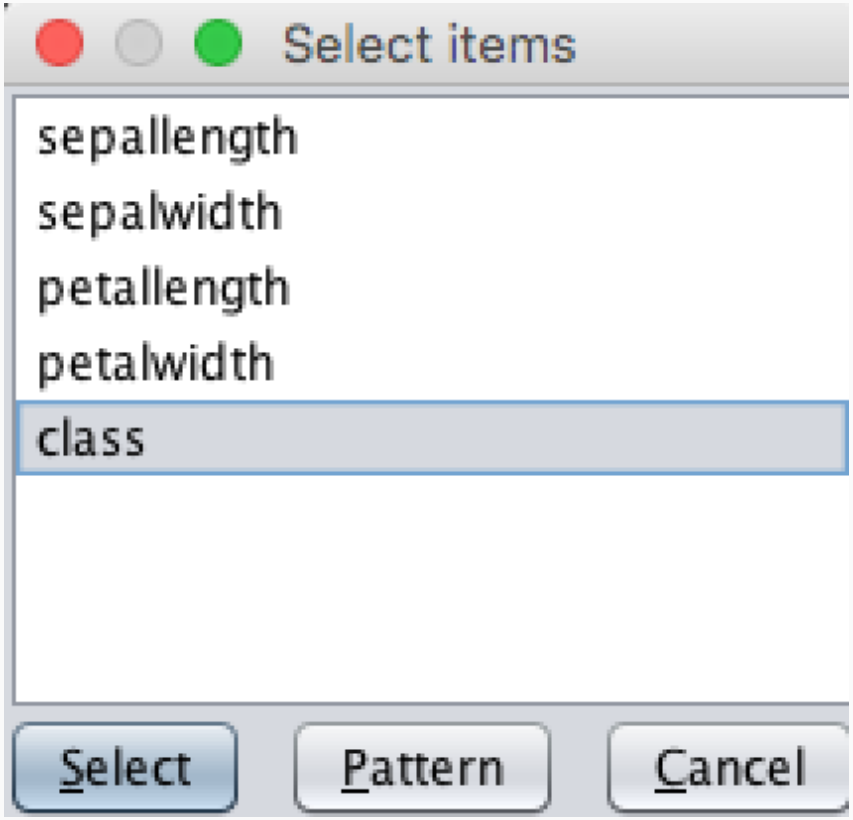

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
weka
clusterers
Canopy
Cobweb
EM
FarthestFirst
FilteredClusterer
HierarchicalClusterer
MakeDensityBasedClusterer
SimpleKMeans
s 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
Clusterer output
Attribute Full Data 0
(150.0) (50.0)
sepallength 5.8433 5.936
sepalwidth 3.054 2.77
petallength 3.7587 4.26
petalwidth 1.1987 1.326
class Iris-setosa Iris-versicolor Iris-
Time taken to build model (full training data) : 0.01 second
=== Model and evaluation on training set ===
Clustered Instances
0 50 (33%)
1 50 (33%)
2 50 (33%)
Close
Status
OK
Log
x 0

Parameters

- set number of clusters to three



Remove variables for clustering



Visualize

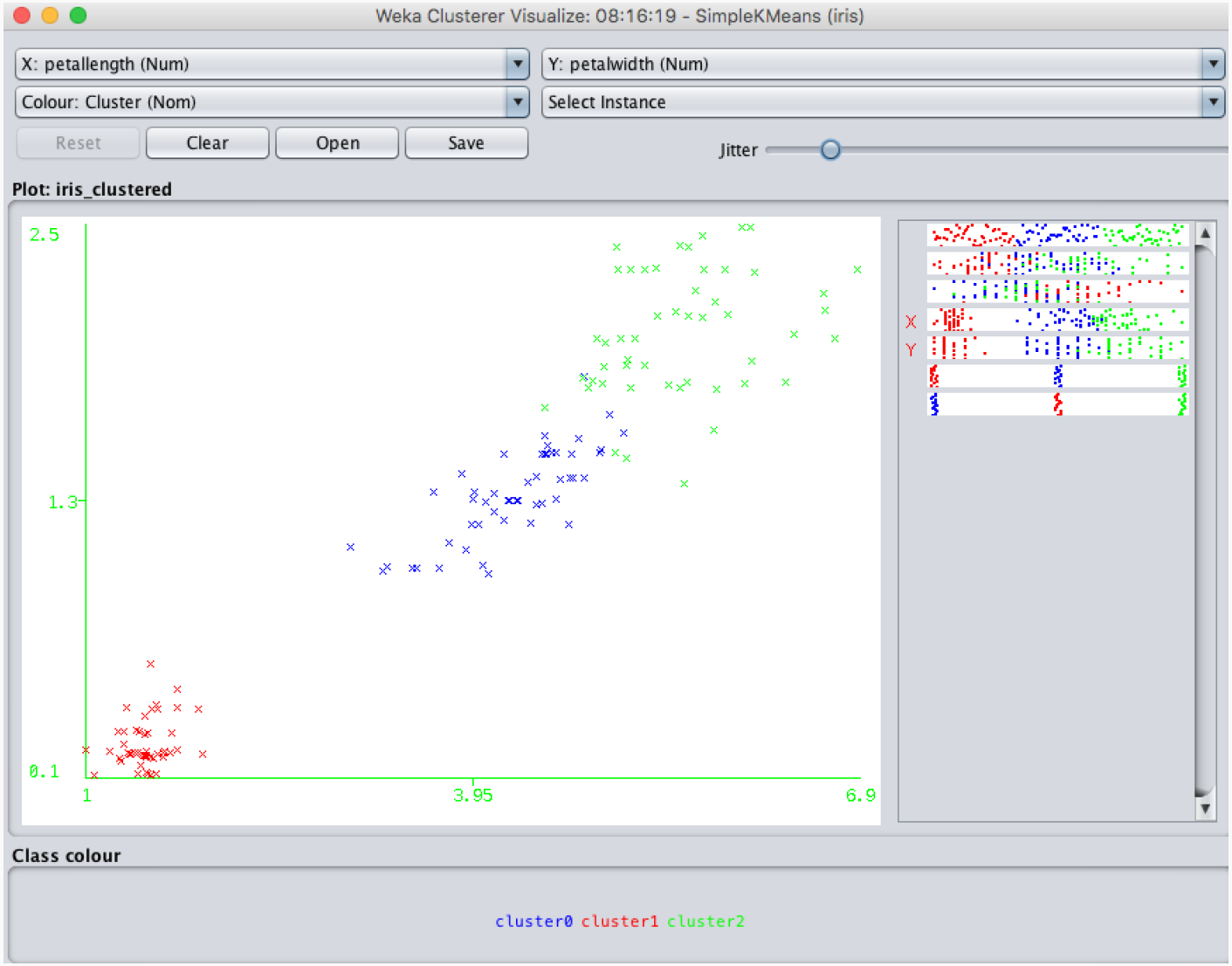

Weka Clusterer Visualize: 08:16:19 - SimpleKMeans (iris)
X: petallength (Num)
Y: petalwidth (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris_clustered
2.5
1.3
0.1
1
3.95
6.9
X
Y
Class colour
cluster0 cluster1 cluster2

Table of Contents

Business Decision

The management team of a large shopping mall would like to understand the types of people who are, or could be, visiting their mall. They have good reasons to believe that there are a few different **market segments**, and they are considering designing and positioning the shopping mall services better in order to attract mainly a few profitable market segments, or to differentiate their services (e.g. invitations to events, discounts, etc) across market segments.

Data acquisition

- gather data
- market survey with potential customers

Name	Description	Scale
V1	Shopping is fun	1-7
V2	Shopping is bad for your budget	1-7
V3	I combine shopping with eating out	1-7
V4	I try to get the best buys while shopping	1-7
V5	I don't care about shopping	1-7
V6	You can save lot of money by comparing prices	1-7
Income	The household income of the respondent	Dollars
Mall.Visits	How often they visit the mall	1-7

Data

ID	V1	V2	V3	V4	V5	V6	Income	Mall.Visits
1	6	4	7	3	2	3	60000	3
2	2	3	1	4	5	4	30000	1
3	7	2	6	4	1	3	70000	3
4	4	6	4	5	3	6	30000	7
5	1	3	2	2	6	4	60000	1
6	6	4	6	3	3	4	50000	2
7	5	3	6	3	3	4	65000	3
8	7	3	7	4	1	4	55000	4
9	2	4	3	3	6	3	70000	0
10	3	5	3	6	4	6	25000	6
...	...	...	...	...	...	...	...	...

Steps in clustering

We will take the following steps

- Select Segmentation Variables
- Define similarity measure
- Method and Number of Segments
- Profile and interpret the segments

Select Segmentation Variables

- critically important decision
- exploratory research usually helps
  - visualization of distributions
  - contextual knowledge, creativity, and experimentation/iterations are needed.
- clustering - we use only few variables (V1..V6)
- profiling - we use the remaining ones (income, numOfVisits)

Define similarity measure

- clustering = grouping objects based on how **similar** they are
- similarities:
  - Euclidian
  - Manhattan
  - Cosine
  - ...

Euclidian distance

- distance between two objects p and q, each with N variables

$$p = (p_1, p_2 ... p_N); q = (q_1, q_2, ... q_N)$$
$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_N - q_N)^2}$$

- for first 10 subjects

	1	2	3	4	5	6	7	8	9	10
1	0									
2	8	0								
3	3	8	0							
4	6	6	7	0						
5	8	3	9	7	0					
6	2	7	3	4	7	0				
7	2	6	3	5	6	1	0			
8	2	9	2	6	9	3	3	0		
9	7	3	8	6	2	6	5	8	0	
10	7	4	7	2	6	6	6	7	5	0

Method and Number of Segments

- choosing the clustering method and number of clusters:
  - statistical reasoning,
  - judgment,
  - interpretability of the clusters,
  - actionable value of the clusters found,
- In practice different algorithms and numbers of segments should be explored, and the final choice should be made based on both statistical and qualitative criteria.

Method and Number of Segments

- choosing the clustering method and number of clusters:
  - statistical reasoning,
  - judgment,
  - interpretability of the clusters,
  - actionable value of the clusters found,
- In practice different algorithms and numbers of segments should be explored, and the final choice should be made based on both statistical and qualitative criteria.

- method:
  - kMeans
  - hierarchical
- number of clusters
  - 3

Profile and interpret the segments

- interpretation of the characteristics of the clusters

	Population	Cluster 1	Cluster 2	Cluster 3
V1	3.85	5.75	1.67	3.50
V2	4.10	3.62	3.00	5.83
V3	3.95	6.00	1.83	3.33
V4	4.10	3.12	3.50	6.00
V5	3.45	1.88	5.50	3.50
V6	4.35	3.88	3.33	6.00
Income	46000.00	60000.00	42500.00	30833.33
Mall.Visits	3.25	3.25	1.00	5.50

Our example in Weka

- File open: mall.csv
- repeat steps from the iris example

References

Part of the material has been taken from the following sources. The usage of the referenced copyrighted work is in line with fair use since it is for nonprofit educational purposes.

- http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html
- wikipedia.org
- https://archive.ics.uci.edu/ml/datasets/iris