K-Means Clustering

By Susan L. Miertschin

Data Mining - Task Types

- Classification
- Clustering
- Discovering Association Rules
- Discovering Sequential Patterns – Sequence Analysis
- Regression
- Detecting Deviations from Normal

Data Mining - Task Types

- Classification
- Clustering
  - Divide data into groups with similar characteristics - *Larson*
  - Find clusters of data objects similar in some way to one another – *Oracle book* (http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/clustering.htm)
- Discovering Association Rules
- Discovering Sequential Patterns – Sequence Analysis
- Regression

Clustering

- Find customers similar to each other based on geographical distance to nearest store-front location, number of small dogs owned, number of cats owned, and number of children in household
  - Purpose? Target niche markets, plan new stores
- Find cardiologists who are similar with respect to likelihood of prescribing a certain class of medication for treatment of congestive heart failure (based on hospital patient records) and patient mix demographics

Clustering

- Partitional
- Hierarchical

Clustering Algorithms

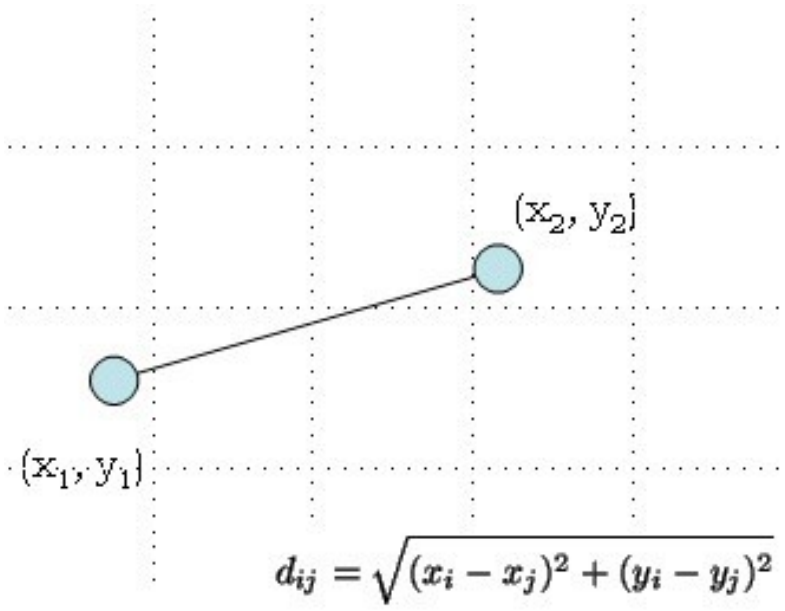
- Group the data based on a criterion
- Look for improvements in the grouping
- If improvement is possible – then revise the groups
- iterate

K-Means Clustering Algorithm

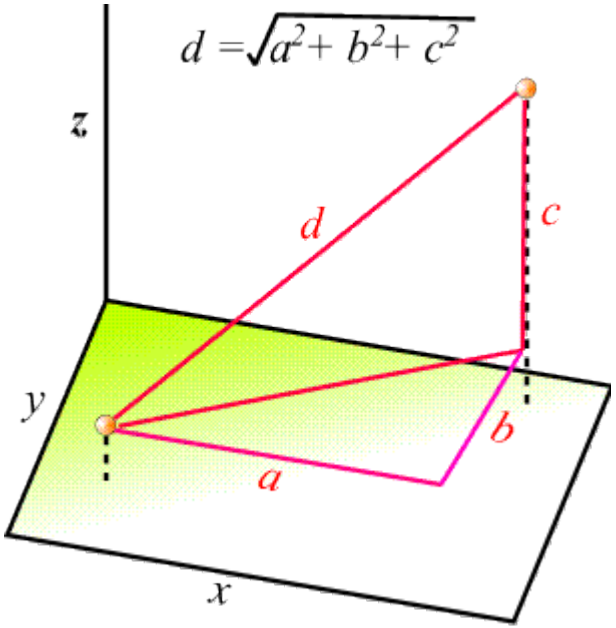
- Choose a value for K – the number of clusters the algorithm should create
- Select K cluster centers from the data
  - Arbitrary as opposed to intelligent selection for “raw” K-means
- Assign the other instances to the group based on “distance to center”
  - Distance is simple Euclidean distance
- Calculate new center for each cluster based on mean values of instances included
- Evaluate to look for possible improvement

Euclidean Distance

- 2 dimensions



- 3 dimensions



Restrictions/Considerations

- Euclidean distance can only be calculated with real numbers
- Categorical data must be converted to numbers
  - There are issues associated with this conversion process
  - If the categorical data is ordinal (i.e., an order can be established for the categories, e.g. win/place/show is an ordered set of categories) – then the conversion is better
  - If the categorical data is nominal – then the conversion is not true to meaning of the

Example – Credit Card Promotion Data Descriptions

Attribute Name	Value Description	Numeric Values	Definition
Income Range	20-30K, 30-40K, 40-50K, 50-60K	20000, 30000, 40000, 50000	Salary range for an individual credit card holder
Magazine Promotion	Yes, No	1, 0	Did card holder participate in magazine promotion offered before?
Watch Promotion	Yes, No	1, 0	Did card holder participate in watch promotion offered before?
Life Ins Promotion	Yes, No	1, 0	Did card holder participate in life insurance promotion offered before?
Credit Card Insuranc	Yes, No	1, 0	Does card holder have credit card insurance?

Sample of Credit Card Promotion Data (from Table 2.3)

Income Range	Magazine Promo	Watch Promo	Life Ins Promo	CC Ins	Sex	Age
40-50K	Yes	No	No	No	Male	45
30-40K	Yes	Yes	Yes	No	Female	40
40-50K	No	No	No	No	Male	42
30-40K	Yes	Yes	Yes	Yes	Male	43
50-60K	Yes	No	Yes	No	Female	38
20-30K	No	No	No	No	Female	55
30-40K	Yes	No	Yes	Yes	Male	35
20-	No	Yes	No	No	Male	27

See data handout.

Sample of Numerical Credit Card Promotion Data (from Table 2.3)

Income Range	Magazine Promo	Watch Promo	Life Ins Promo	CC Ins	Sex	Age
40000	1	0	0	0	1	45
30000	1	1	1	0	0	40
40000	0	0	0	0	1	42
30000	1	1	1	1	1	43
50000	1	0	1	0	0	38
20000	0	0	0	0	0	55
30000	1	0	1	1	1	35
20000	0	1	0	0	1	27
30000	1	0	0	0	1	43
30000	1	1	1	0	0	41

See data handout.

Implementing K-Means Algorithm in Excel

- There is a link to the Excel file used to create the data handout in Blackboard
- Download the .zip archive using the link, extract the .csv file, and open it in Excel
- Follow along with the slides - using Excel

	A	B	C	D	E	F	G
1							
2	Income Range	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Age	Life Ins Promotion
3	40000	1	0	0	1	45	0
4	30000	1	1	0	0	40	1
5	40000	0	0	0	1	42	0
6	30000	1	1	1	1	43	1
7	50000	1	0	0	0	38	1
8	20000	0	0	0	0	55	0
9	30000	1	0	1	1	35	1
10	20000	0	1	0	1	27	0
11	30000	1	0	0	1	43	0
12	30000	1	1	0	0	41	1
13	40000	0	1	0	0	43	1
14	20000	0	1	0	1	29	1
15	50000	1	1	0	0	39	1
16	40000	0	1	0	1	55	0
17	20000	0	0	1	0	19	1
18							

K-Means Algorithm Steps in Excel

- Set the number of clusters
  - K = 4 (arbitrary)
  - Select K centers
    - Select first points that represent 4 different income ranges = Instances 1,2, 5, 6 (this is slightly less arbitrary)

19						
20	Set the value of K	4				
21	Select K centers	Instances 1, 2, 5, 6				

K-Means Algorithm Steps in Excel

- Compute distance to each center from every other instance (point)
- Use the distance formula
- Each instance in this data set is a 7-tuple
  - E.g. (40000,1,0,0,1,45, 0)

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

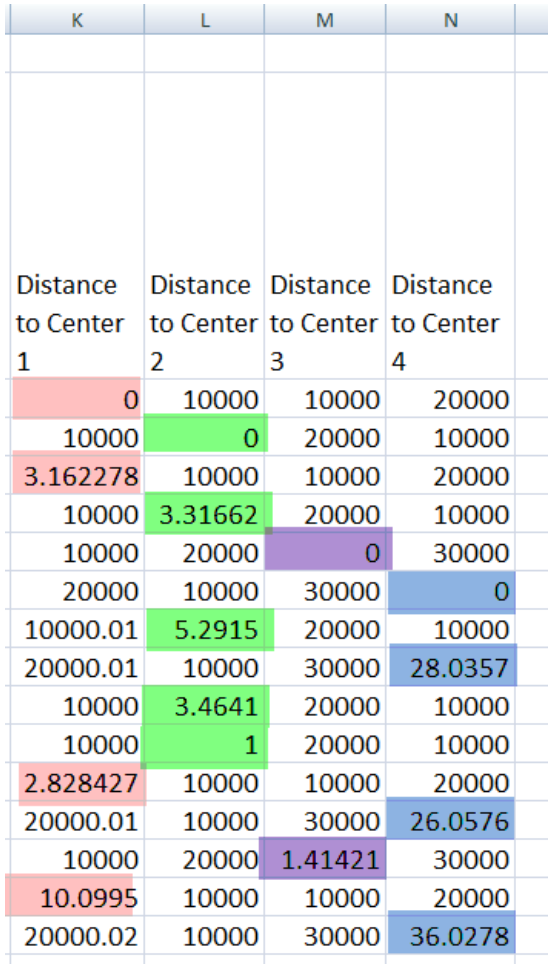
K-Means Algorithm Steps in Excel

- Here is what your result should look like
- The cells that contain 0 correspond to the distance between a chosen center point and itself

K	L	M	N
Distance to Center 1	Distance to Center 2	Distance to Center 3	Distance to Center 4
0	10000	10000	20000
10000	0	20000	10000
3.162278	10000	10000	20000
10000	3.31662	20000	10000
10000	20000	0	30000
20000	10000	30000	0
10000.01	5.2915	20000	10000
20000.01	10000	30000	28.0357
10000	3.4641	20000	10000
10000	1	20000	10000
2.828427	10000	10000	20000
20000.01	10000	30000	26.0576
10000	20000	1.41421	30000
10.0995	10000	10000	20000
20000.02	10000	30000	36.0278

K-Means Algorithm Steps in Excel

- For each instance there are four distance values
- Choose the minimum distance to associate the instance with the center of the cluster
- Do you see any problems with the way these "distances" are



K	L	M	N
Distance to Center 1	Distance to Center 2	Distance to Center 3	Distance to Center 4
0	10000	10000	20000
10000	0	20000	10000
3.162278	10000	10000	20000
10000	3.31662	20000	10000
10000	20000	0	30000
20000	10000	30000	0
10000.01	5.2915	20000	10000
20000.01	10000	30000	28.0357
10000	3.4641	20000	10000
10000	1	20000	10000
2.828427	10000	10000	20000
20000.01	10000	30000	26.0576
10000	20000	1.41421	30000
10.0995	10000	10000	20000
20000.02	10000	30000	36.0278

K-Means Algorithm Steps in Excel

- Transformed Data Values

	A	B	C	D	E	F	G
1							
2	Income Range	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Age	Life Ins Promotion
3	40	1	0	0	1	45	0
4	30	1	1	0	0	40	1
5	40	0	0	0	1	42	0
6	30	1	1	1	1	43	1
7	50	1	0	0	0	38	1
8	20	0	0	0	0	55	0
9	30	1	0	1	1	35	1
10	20	0	1	0	1	27	0
11	30	1	0	0	1	43	0
12	30	1	1	0	0	41	1
13	40	0	1	0	0	43	1
14	20	0	1	0	1	29	1
15	50	1	1	0	0	39	1
16	40	0	1	0	1	55	0
17	20	0	0	1	0	19	1
18							

- New Distances Calculated

K	L	M	N
Distance to Center 1	Distance to Center 2	Distance to Center 3	Distance to Center 4
0	11.3137	12.2882	22.4054
11.31371	0	20.1246	18.1108
3.162278	10.3923	10.9087	23.8747
10.34408	3.31662	20.6882	15.7797
12.28821	20.1246	0	34.5109
22.40536	18.1108	34.5109	0
14.21267	5.2915	20.2731	22.4499
26.94439	16.4924	32.0156	28.0357
10.19804	3.4641	20.664	15.6844
10.90871	1	20.2485	17.2916
2.828427	10.4881	11.2694	23.3666
25.671	14.9332	31.3688	26.0576
11.78983	20.025	1.41421	34.0441
10.0995	18.1108	19.8242	20.0499
32.86335	23.3238	35.5387	36.0278

K-Means Algorithm Steps in Excel

- New clusters

K	L	M	N
Distance to Center 1	Distance to Center 2	Distance to Center 3	Distance to Center 4
0	11.3137	12.2882	22.4054
11.31371	0	20.1246	18.1108
3.162278	10.3923	10.9087	23.8747
10.34408	3.31662	20.6882	15.7797
12.28821	20.1246	0	34.5109
22.40536	18.1108	34.5109	0
14.21267	5.2915	20.2731	22.4499
26.94439	16.4924	32.0156	28.0357
10.19804	3.4641	20.664	15.6844
10.90871	1	20.2485	17.2916
2.828427	10.4881	11.2694	23.3666
25.671	14.9332	31.3688	26.0576
11.78983	20.025	1.41421	34.0441
10.0995	18.1108	19.8242	20.0499
32.86335	23.3238	35.5387	36.0278

K-Means Algorithm Steps in Excel

- Identify the instances that belong to the minimum distance values

25		Income Range	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Age	Life Ins Promotion
26	Cluster 1	40	1	0	0	1	45	0
27	Cluster 2	30	1	1	0	0	40	1
28	Cluster 1	40	0	0	0	1	42	0
29	Cluster 2	30	1	1	1	1	43	1
30	Cluster 3	50	1	0	0	0	38	1
31	Cluster 4	20	0	0	0	0	55	0
32	Cluster 2	30	1	0	1	1	35	1
33	Cluster 2	20	0	1	0	1	27	0
34	Cluster 2	30	1	0	0	1	43	0
35	Cluster 2	30	1	1	0	0	41	1
36	Cluster 1	40	0	1	0	0	43	1
37	Cluster 2	20	0	1	0	1	29	1
38	Cluster 3	50	1	1	0	0	39	1
39	Cluster 1	40	0	1	0	1	55	0
40	Cluster 2	20	0	0	1	0	19	1
41								

K-Means Algorithm Steps in Excel

- Calculate means of attribute values by cluster to determine the cluster center
- Sort by cluster to aid in calculation
- If calculated center = former center (to a certain precision)
  - then terminate the algorithm

		Income Range	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Age	Life Ins Promotion
25								
26	Cluster 1	40	1	0	0	1	45	0
27	Cluster 1	40	0	0	0	1	42	0
28	Cluster 1	40	0	1	0	0	43	1
29	Cluster 1	40	0	1	0	1	55	0
30	Cluster 2	30	1	1	0	0	40	1
31	Cluster 2	30	1	1	1	1	43	1
32	Cluster 2	30	1	0	1	1	35	1
33	Cluster 2	20	0	1	0	1	27	0
34	Cluster 2	30	1	0	0	1	43	0
35	Cluster 2	30	1	1	0	0	41	1
36	Cluster 2	20	0	1	0	1	29	1
37	Cluster 2	20	0	0	1	0	19	1
38	Cluster 3	50	1	0	0	0	38	1
39	Cluster 3	50	1	1	0	0	39	1
40	Cluster 4	20	0	0	0	0	55	0
41	Cluster 1 Center	40	0.25	0.5	0	0.8	46	0.25
42	Cluster 2 Center	26.25	0.625	0.625	0.375	0.6	35	0.75
43	Cluster 3 Center	50	1	0.5	0	0	39	1
44	Cluster 4 Center	20	0	0	0	0	55	0

K-Means Algorithm Steps in Excel

- Continue iteration using the new centers
- Yields new clusters
- Either
  - terminate if new centers = previous centers
  - OR
  - Continue iterations

Distance to Center 1	Distance to Center 2	Distance to Center 3	Distance to Center 4
1.581139	17.2649	12.0208	22.4054
11.87434	6.62028	20.0624	18.1108
4.301163	15.655	10.7471	23.8747
10.63015	9.22378	20.5548	15.7797
13.0384	24.012	0.70711	34.5109
21.85177	21.356	34.271	0
15.13275	3.91511	20.3593	22.4499
27.76689	9.92865	32.1792	28.0357
10.55936	9.25084	20.5548	15.6844
11.37981	7.45507	20.1618	17.2916
3.464102	16.1347	11.0227	23.3666
26.42915	8.46039	31.504	26.0576
12.4298	24.1677	0.70711	34.0441
8.774964	24.6085	19.3778	20.0499
33.83785	16.8769	35.812	36.0278

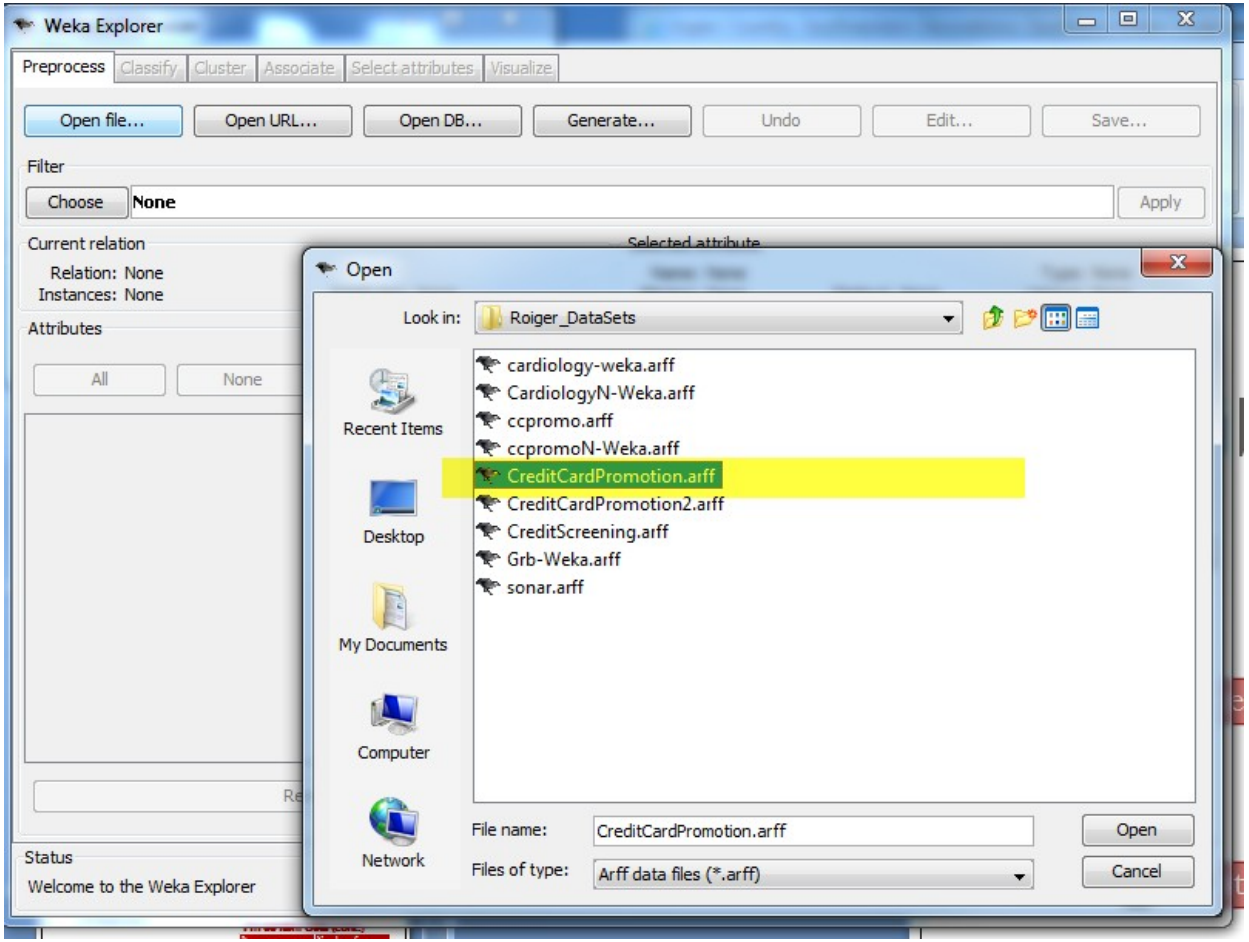
Computation Question #10 (p. 103, Roiger)

- Perform the third iteration of the K-Means algorithm for the example given here in the slides
- What are the new cluster centers?
- Save your Excel workbook with your organized work relating to K-Means clustering and submit it in the dropbox named IC 0809 K-Means in Balckboard

Use WEKA

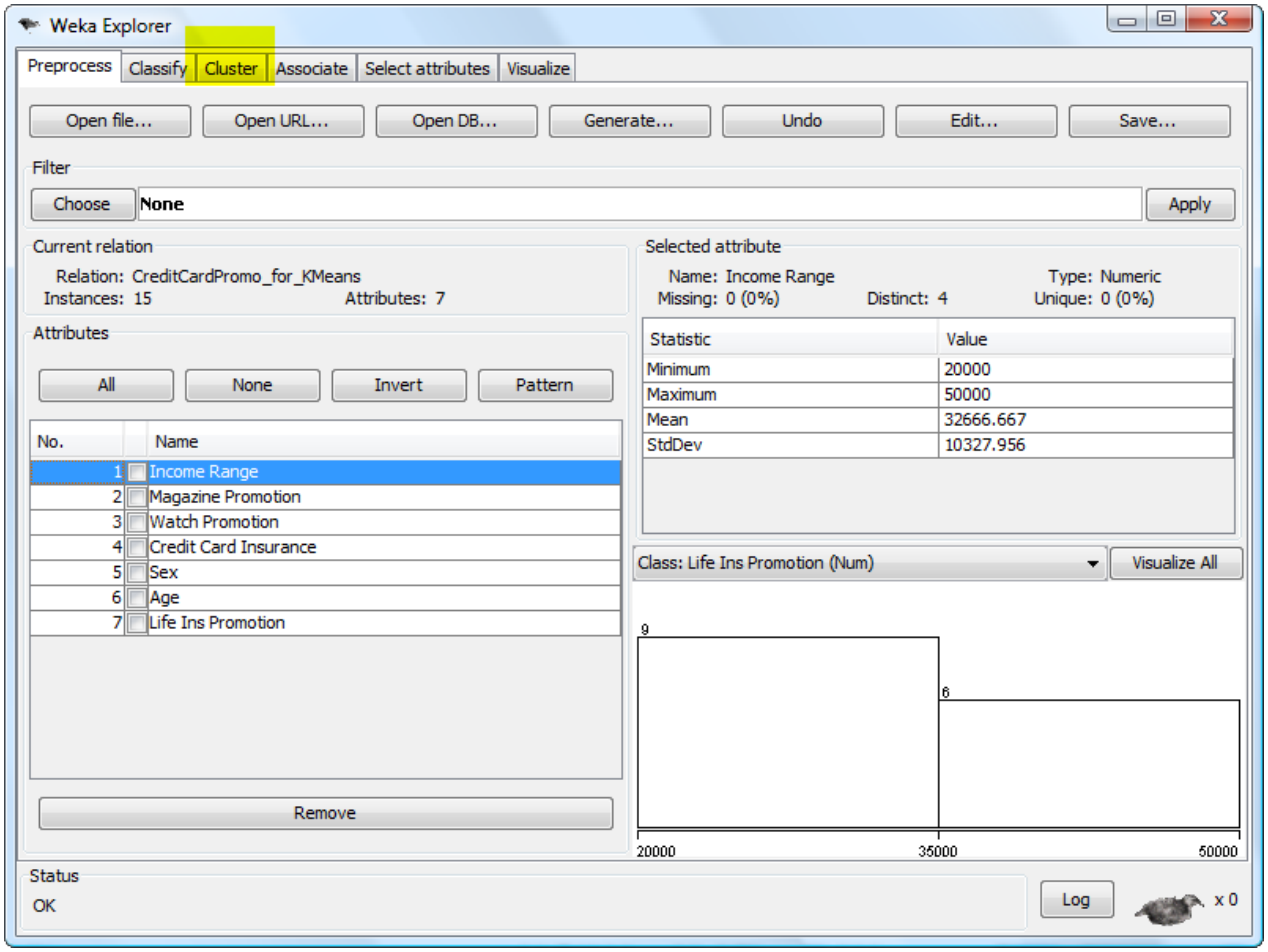

Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.6.3
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

Use WEKA

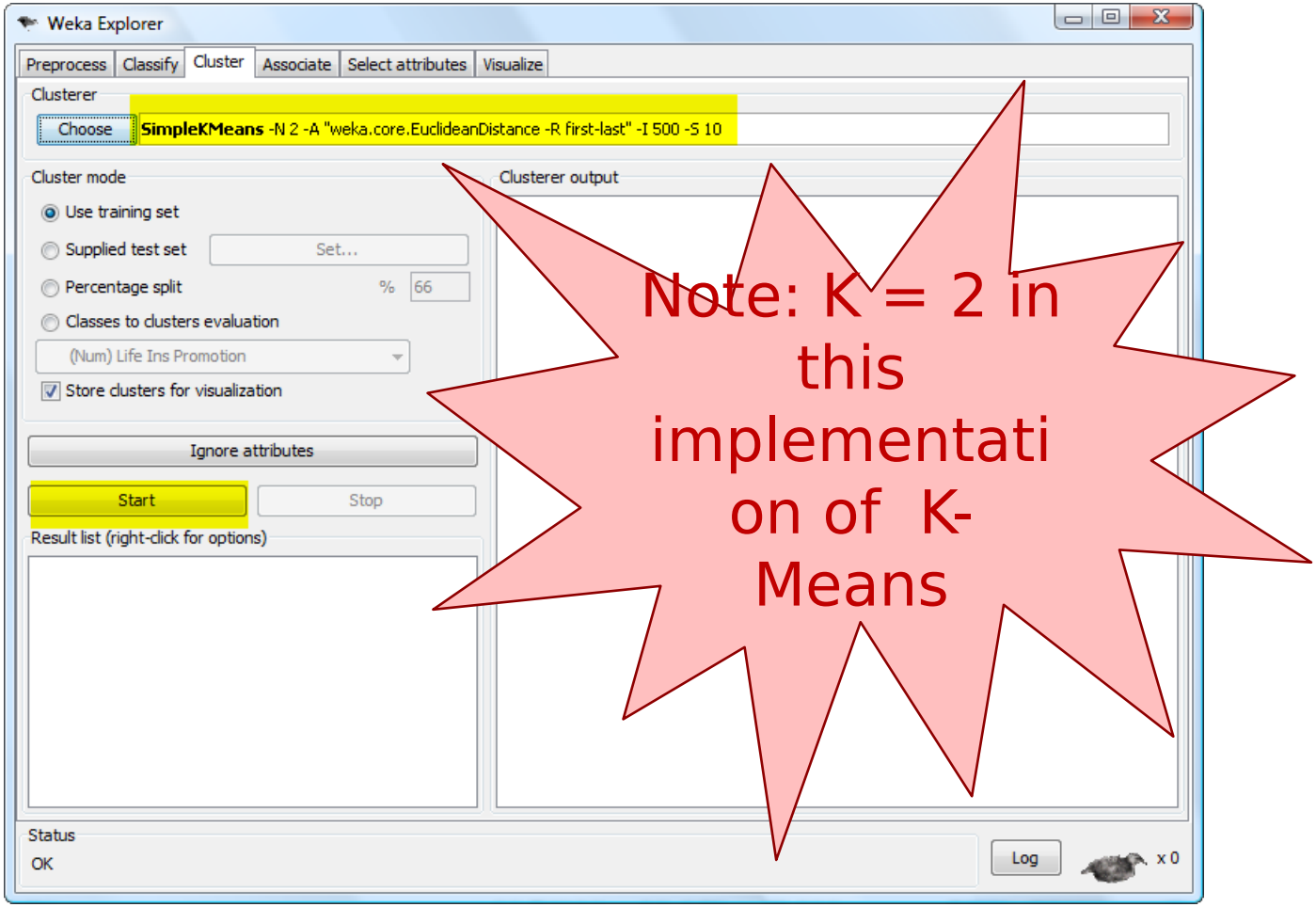


Open the data file you downloaded and used for the Excel exercise. If you open this file in WEKA and then save it With WEKA

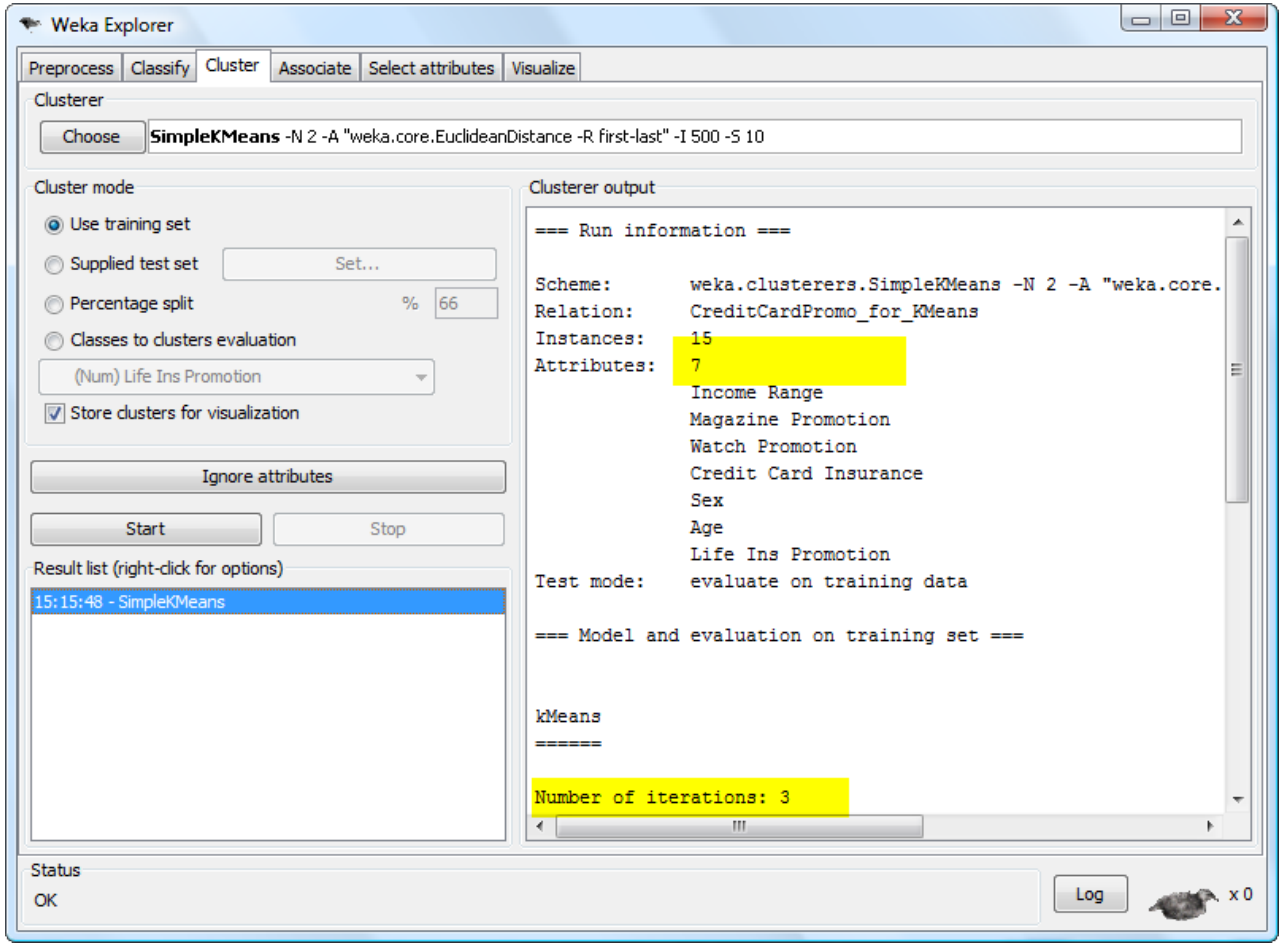
Use WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: CreditCardPromo_for_KMeans
Instances: 15
Attributes: 7
Attributes
All
None
Invert
Pattern
No.
Name
1 Income Range
2 Magazine Promotion
3 Watch Promotion
4 Credit Card Insurance
5 Sex
6 Age
7 Life Ins Promotion
Remove
Selected attribute
Name: Income Range
Type: Numeric
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)
Statistic
Value
Minimum 20000
Maximum 50000
Mean 32666.667
StdDev 10327.956
Class: Life Ins Promotion (Num)
Visualize All
9
6
20000
35000
50000
Status
OK
Log
x 0

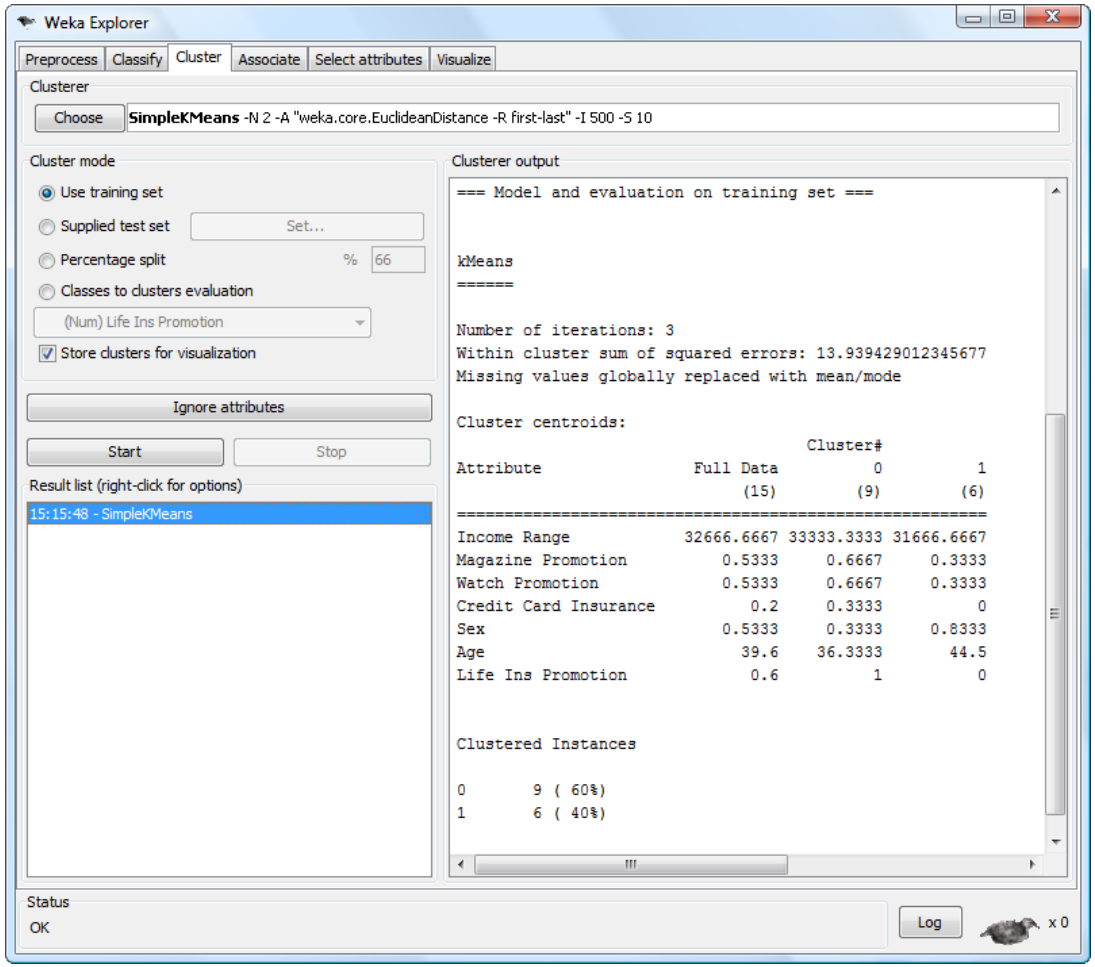
Use WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Life Ins Promotion
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Clusterer output
Status
OK
Log
x 0
Note: K = 2 in this implementation of K-Means

Use WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Life Ins Promotion
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
15:15:48 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.
Relation: CreditCardPromo_for_KMeans
Instances: 15
Attributes: 7
Income Range
Magazine Promotion
Watch Promotion
Credit Card Insurance
Sex
Age
Life Ins Promotion
Test mode: evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 3
Status
OK
Log
x 0

Use WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) Life Ins Promotion
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
15:15:48 - SimpleKMeans
Clusterer output
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 13.939429012345677
Missing values globally replaced with mean/mode
Cluster centroids:
Cluster#
Attribute Full Data 0 1
(15) (9) (6)
Income Range 32666.6667 33333.3333 31666.6667
Magazine Promotion 0.5333 0.6667 0.3333
Watch Promotion 0.5333 0.6667 0.3333
Credit Card Insurance 0.2 0.3333 0
Sex 0.5333 0.3333 0.8333
Age 39.6 36.3333 44.5
Life Ins Promotion 0.6 1 0
Clustered Instances
0 9 (60%)
1 6 (40%)
Status
OK
Log
x 0

K-Means Clustering

By Susan L. Miertschin