# Lab Tutorial
# K-Means and EM Clustering

Although clustering is used on unlabelled datasets, we will still use many of the same datasets that we used for classification. But in the case of clustering, the class label attribute is treated just like any other attribute of the data if you just want to explore the data set, or you can exclude the label attribute and let the clustering algorithm "predicts" the labels.

In this class, we will work on some unlabeled data from the US Census Bureau. Start Weka and import the US census data in the "Preprocessing" panel, and remove the "case ID" attribute. The file name is "UScensus_abbr_arff.arff".

1. Here is a **short introduction to the US Census data** for you to learn about its attributes and interpret results:

Attributes of the raw data is discretized to have less attribute values, which is the data we are seeing now. Attributes description of the raw data attributes is at:

http://archive.ics.uci.edu/ml/databases/census1990/USCensus1990raw.attributes.txt

Some attributes are kept the same from raw data set to the current data set, with an "i" attached to the front of current attribute name indicating it's unchanged; the discretized attributes of raw data set are named with a "d" added in front of their original names. For example, in current data set, attribute "dAge" is discretized from raw data set, and its description should be "AAGE" in the raw data description (Age); "iAvail" means the attribute values is not changed from its raw values, and its corresponding attribute is "AVAIL" in raw data description (Available for work).

For more information, the mapping functions from raw attributes to current attributes can be found here:

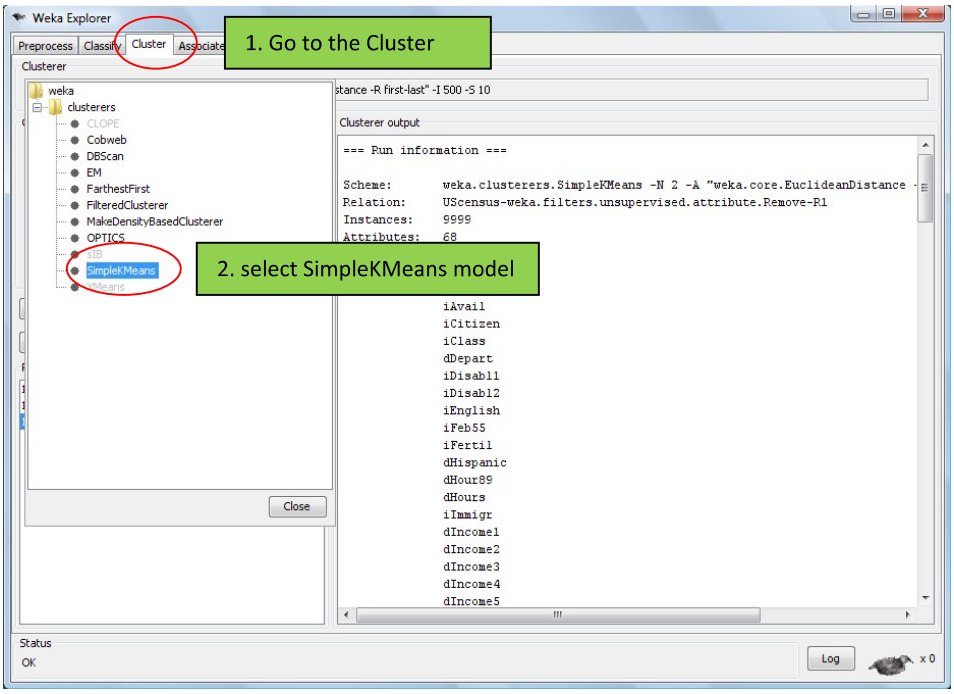
http://archive.ics.uci.edu/ml/databases/census1990/USCensus1990.mapping.sql

The file used in this lab is an abbreviated version of the data set, obtaining the first 9,999 instances out of 2,458,285.

**2. Simple K-means**

**2.1 Perform Simple K-means clustering**

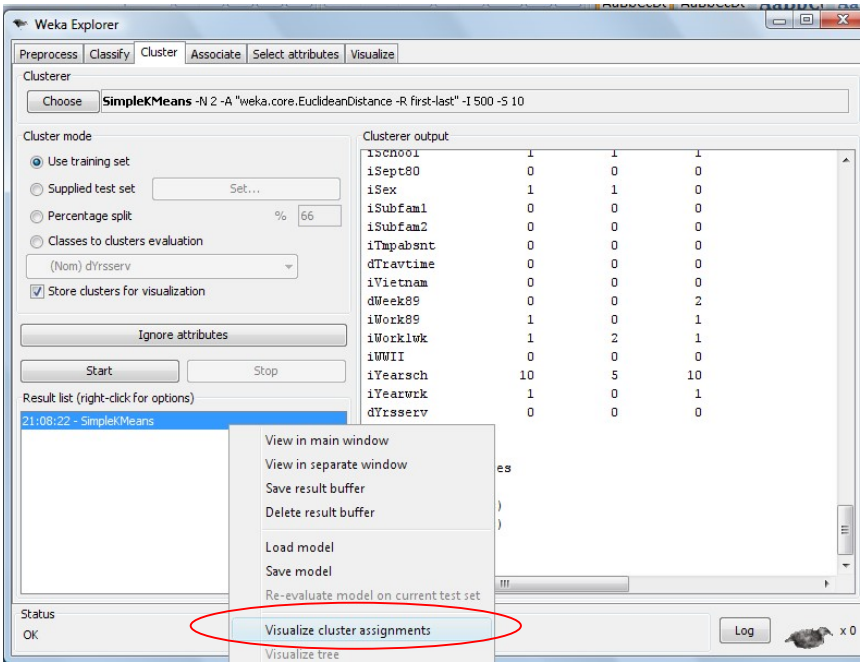
Go to the "Cluster" panel. Click the "Choose" button on the top left, and select "weka->clusterers->SimpleKMeans". Keep the parameter settings as default.




Click on the “Start” button and the result comes up in a while.

### 2.2 Visualize clustering result

Right click on the SimpleKMeans model in the “Result list” section, and select the “Visualize cluster assignments” option.



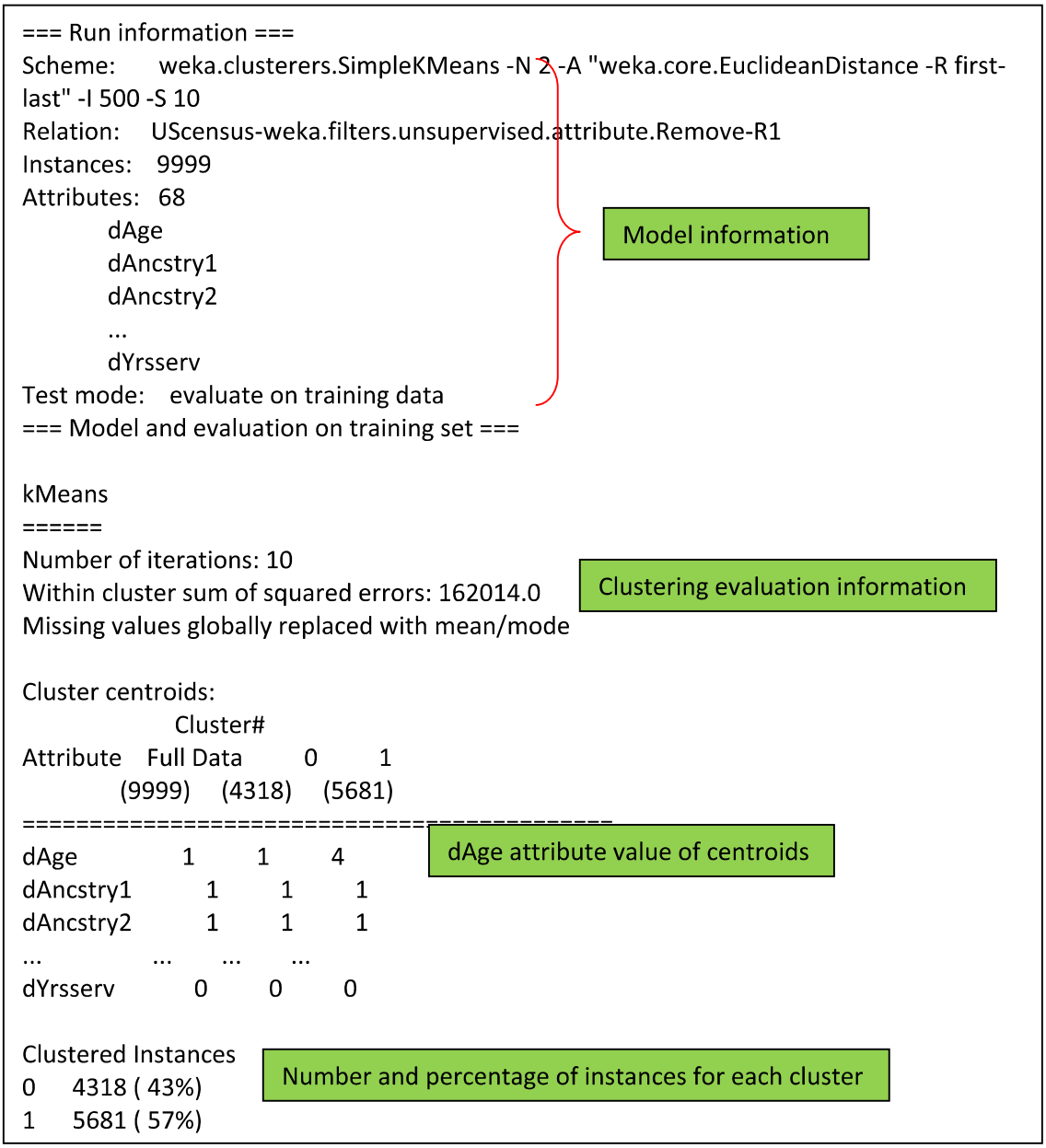

In the pop-up window, you can select X axis, Y axis, etc. to visualize clustering results. By default we are given this:


The plot shows instance number (X axis) and its value on the dAge attribute. As indicated below, color blue represents cluster0 and red represents cluster1. We can see that when Y=1 (dAge=1, meaning 0<age<13), the crosses (representing instances) are almost blue. That is to say, for people who are under 13 years old, they belong to cluster 0. You can observe more characteristics of clusters on attributes by changing Y axis options to other attributes.

If you want to see exactly how many instances are on Y=1, on the “Cluster” panel you can set the parameter “displayStdDevs” to true. It will display number of instances for a specific attribute value for nominal values or standard deviation for numeric values.

### 2.3 Explaining the result

Below shows SimpleKMeans result and illustrates its meanings.




"Within cluster sum of squared errors" measures sum of distances from an instance to the centroid. Put in plain words, we want this value to be as small as possible, to make sure instances are close to centroids within its cluster. You can tune parameters to reduce this value in the exercise.

The "cluster centroids" tells the value of centroid instances on a specific attribute. For example, the first line "1, 1, 4" tells us the dAge value is 1 for the centroid of full data, the dAge value is 1 for the centroid of cluster0, and the dAge value is 4 for the centroid of cluster1.

“Clustered instances” shows number and percentage of instances that each cluster owns. In this example, cluster0 has 43% of all instances and cluster1 has 57%.

### 2.4 Tune parameters

You can try adjusting these parameters:

distanceFunction: the function of computing instance distances

numClusters: number of clusters you want the data set to be grouped to

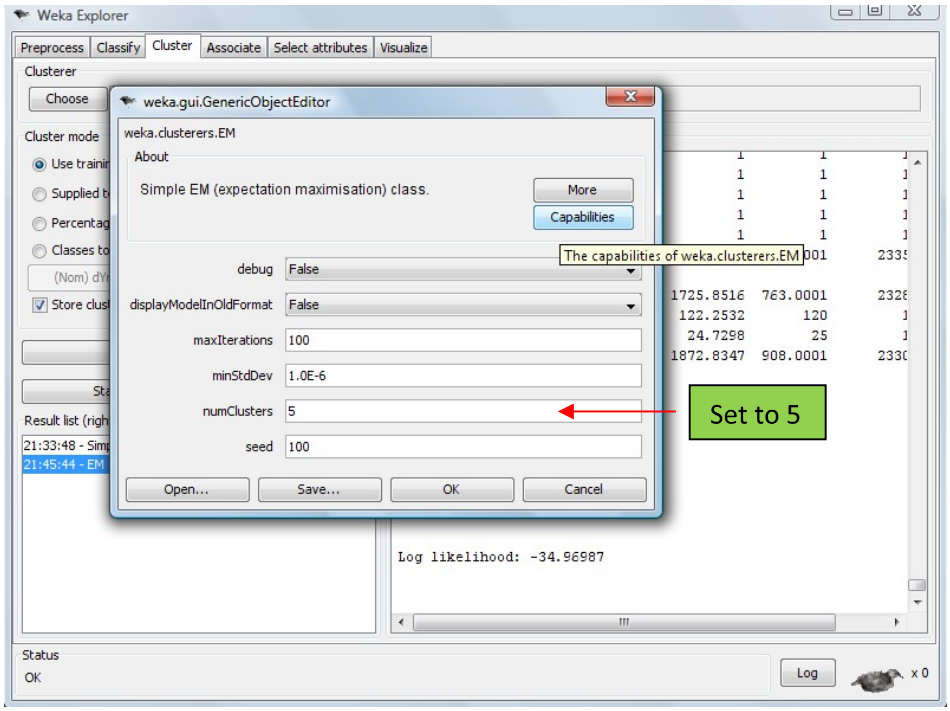
seed: the random seed number to decide the initial centroids

You may want to delete some items on your Result list as you work in order to reclaim memory used for those models. If Weka runs out of memory during your experiments, you can just restart it, or you can try to expand the memory using the instructions given for libsvm.

## 3. Another clustering model: EM

### 3.1 Perform EM clustering

Still use the US census data, and go to the “Cluster” panel. Click the “Choose” button on the top left, and select “weka->clusterers->EM”. Click on the parameter textbox on the right of the “Click” button. In the pop-up “weka.gui.GenericObjectEditor”, set the “numClusters” to 5. Keep other parameters as default.



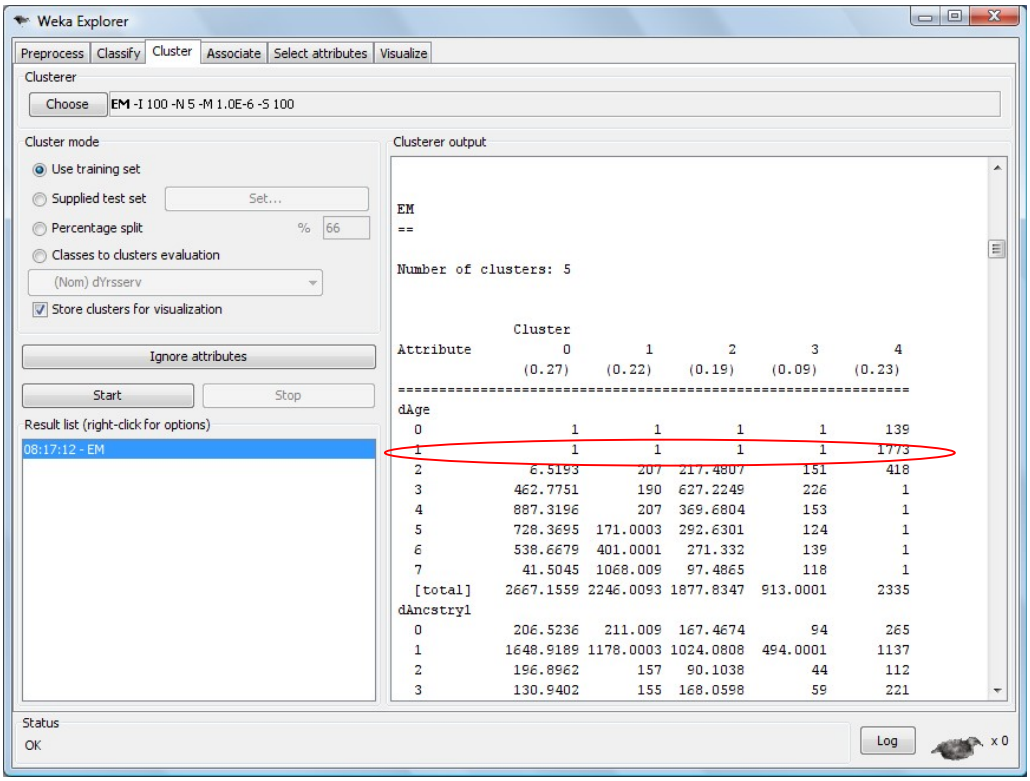
Click on the “Start” button to start clustering.

### 3.2 Visualizing the result

This is the same as section 2.2.

### 3.3 Explaining the result

The below section of clustering output shows for each attribute how it is distributed on clusters. For example, in the red circle, we can see for attribute dAge when its value is 1 (meaning 0<age<13) there is 1 instance in cluster0, 1 instance in cluster1, 1 instance in cluster2, 1 instance in cluster2, 1 instance in cluster3, and 1773 instances in cluster4. Since the EM clustering algorithm is computing probabilities of clusters, for nominal value counts, it will sometimes add small fractions in order to avoid 0 probabilities.




3.4 Tune parameters

Some parameters that you can try:

numClusters: number of clusters that you want to group the data set in

seed: the random number seed (an integer) to decide the initial centroids