Name: Tuheena Singh Lab #5: Clustering

Submit **a scanned copy of answered this Lab #5** to the Lab #5 DropBox by Sunday, 11/8 @11:59PM.

# Introduction:

Clustering involves the identification of natural groups across a dataset. It is an unsupervised learning problem that deals with identifying structure in a collection of unlabeled data.

In this lab session, you will experiment with two clustering algorithms. You will use Weka for these experiments.

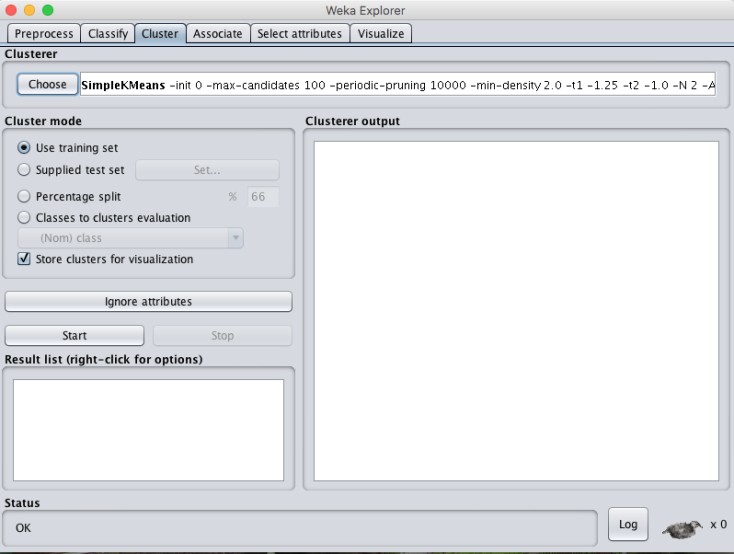
**Background reading:** Basic clustering concepts like k-means & Hierarchical Agglomerative Clustering (HAC)

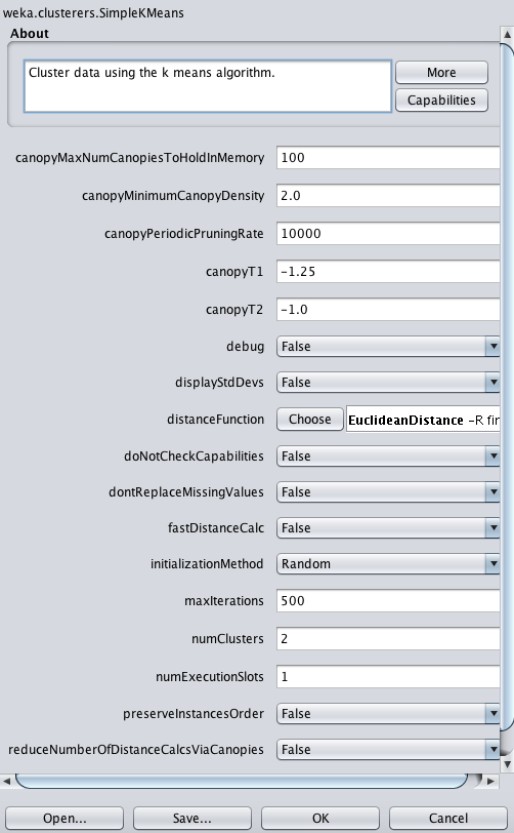
# Part 1. Partitional Clustering: K-Means

**Dataset used: *iris*** dataset (check Weka installation folder)

The *k*-means algorithm is a simple, straightforward algorithm to assign instances to clusters. A cluster centroid defines each cluster, and instances belong to the cluster for which their Euclidian distance to the centroid is the smallest. For each cluster a new centroid is found by taking the average over the cluster instances, leading to shifts of instances between clusters. This iterative process ends when the centroids stop changing.

In Weka Explorer, load the training file iris.arff. Navigate to the cluster tab and select the clustering algorithm SimpleKMeans.



Now click anywhere in the textbox next to the choose button.

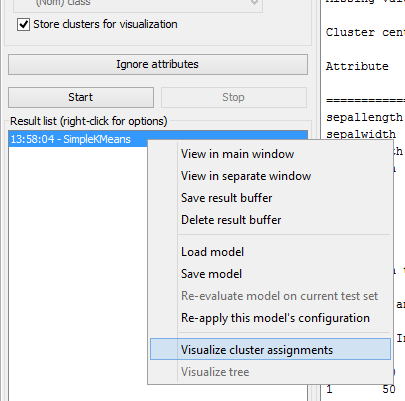
Click more to get details about each parameter. The important ones to look for are distanceFunction, maxIterations, and numClusters (value of k). Set numClusters to 3. Select “Classes to cluster evaluation” and click the “Ignore attributes” and select “class.”

Then start the clustering.

|  |  |  |
| --- | --- | --- |
| Q1.1 | How many instances were clustered incorrectly? | 17.0 ; 11.3333 % |
| Q1.2 | How many instances are in cluster2? | 39 |
|  | How many of these instances were incorrectly clustered?  And which cluster should they belong to? | 3  0 |

**Visualization:**

To look at the graphical representation, right-click on the model and select “Visualize cluster assignments.”



Q1.3 Set the x-axis to instance number and y-axis to sepallength. Change the color to class. Which type of iris flower has all instances correctly clustered? Iris-Setosa

**Evaluation:**

The way Weka evaluates the clustering depends on the cluster mode you select. Four different cluster modes are available (as buttons in the Cluster mode panel):

1. **Use training set** (default). After generating the clustering Weka classifies the training instances into clusters according to the cluster representation and computes the percentage of instances falling in each cluster.
2. In the **Supplied test set** or **Percentage split,** Weka can evaluate clusterings on separate test data if the cluster representation is probabilistic (e.g., for EM).
3. **Classes to clusters evaluation.** In this mode, Weka first ignores the class attribute and generates the clustering. It assigns classes to the clusters during the test phase, based on the majority value of the class attribute within each cluster. Then it computes the classification error based on this assignment and also shows the corresponding confusion matrix

## Q1.4 Load the iris.arff again. Remove the class attribute using the pre-processing dialog box. Cluster data using SimpleKMeans using “**Use training set(default)**.” Experiment with k=3, 5, 7, and 8 and compare the results using the most common measure discussed in the lecture.

## K=3 : Sum of squared errors: 6.998

## K=5 : Sum of squared errors: 5.130

## K=7 : Sum of squared errors: 3.757

## K=8 : Sum of squared errors: 3.407

## Increasing the value of K, shows the decrease in the size of the clusters created. For example when K=8:

## Clustered Instances

## 0 22 ( 15%)

## 1 19 ( 13%)

## 2 25 ( 17%)

## 3 13 ( 9%)

## 4 16 ( 11%)

## 5 18 ( 12%)

## 6 20 ( 13%)

## 7 17 ( 11%)

## And when K=3 :

## Clustered Instances

## 0 61 ( 41%)

## 1 50 ( 33%)

## 2 39 ( 26%) We can observe from the SSE’s , for increasing K values the SSE is decreasing because the data points keep getting closer to cluster centroid with each iteration.

**Part 2. Hierarchical Clustering**

Use the **similarity** matrix(1= two points are precisely the same; 0= completely different) in the following Table to perform both single link and complete link hierarchical clustering algorithms. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. Give the updated similarity matrix after each merge.

Similarity Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| P1 | 1 | 0.1 | 0.41 | 0.55 | 0.35 |
| P2 | 0.1 | 1 | 0.64 | 0.47 | 0.88 |
| P3 | 0.41 | 0.64 | 1 | 0.44 | 0.85 |
| P4 | 0.55 | 0.47 | 0.44 | 1 | 0.76 |
| P5 | 0.35 | 0.88 | 0.85 | 0.76 | 1 |

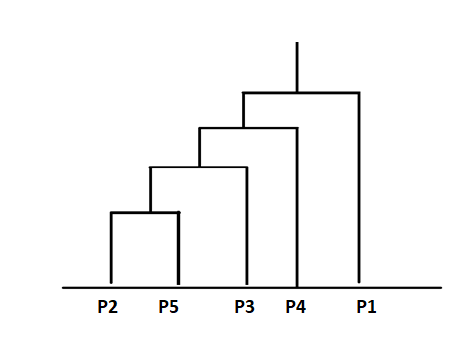
1. Single link hierarchical clustering : Shortest Distance between the clusters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | P1 | P2,P5 | P3 | P4 |
| P1 | 1 | 0.35 | 0.41 | 0.55 |
| P2,P5 | 0.35 | 1 | 0.85 | 0.76 |
| P3 | 0.41 | 0.64 | 1 | 0.44 |
| P4 | 0.55 | 0.47 | 0.44 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | P1 | P2,P5,P3 | P4 |
| P1 | 1 | 0.41 | 0.55 |
| P2,P5,P3 | 0.41 | 1 | 0.76 |
| P4 | 0.55 | 0.76 | 1 |

|  |  |  |
| --- | --- | --- |
|  | P1 | P2,P5,P3,P4 |
| P1 | 1 | 0.41 |
| P2,P5,P3,P4 | 0.55 | 1 |

Dendrogram :



Similarity Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| P1 | 1 | 0.1 | 0.41 | 0.55 | 0.35 |
| P2 | 0.1 | 1 | 0.64 | 0.47 | 0.88 |
| P3 | 0.41 | 0.64 | 1 | 0.44 | 0.85 |
| P4 | 0.55 | 0.47 | 0.44 | 1 | 0.76 |
| P5 | 0.35 | 0.88 | 0.85 | 0.76 | 1 |

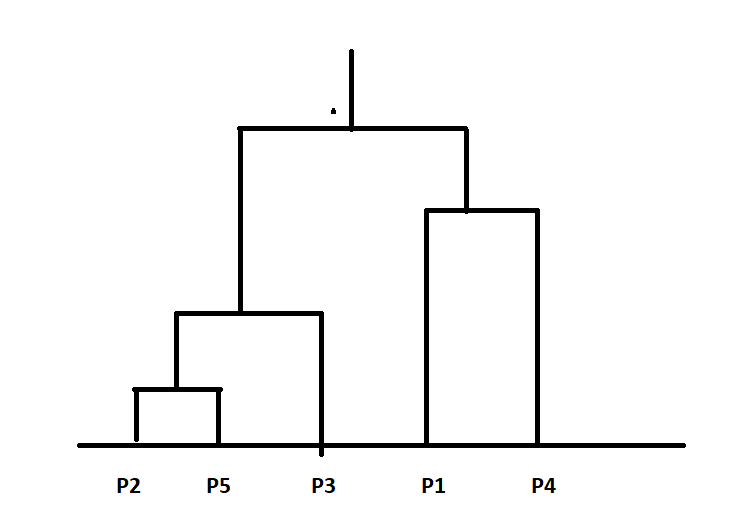
1. Complete link hierarchical clustering : Widest distance between two clusters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2,P5 | P3 | P4 |  |
| P1 | 1 | 0.1 | 0.41 | 0.55 |  |
| P2,P5 | 0.1 | 1 | 0.64 | 0.47 |  |
| P3 | 0.41 | 0.64 | 1 | 0.44 |  |
| P4 | 0.55 | 0.47 | 0.44 | 1 |  |
|  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | P1 | P2,P5,P3 | P4 |
| P1 | 1 | 0.1 | 0.55 |
| P2,P5,P3 | 0.1 | 1 | 0.44 |
| P4 | 0.55 | 0.44 | 1 |

|  |  |  |
| --- | --- | --- |
|  | P1,P4 | P2,P5,P3 |
| P1 | 1 | 0.1 |
| P2,P5,P3 | 0.1 | 1 |

Dendrogram :



## LAB 5 Gradesheet Name:

|  |  |  |  |
| --- | --- | --- | --- |
| Questions | Max. Points | Points Earned | Comments |
| Part 1 |  |  |  |
| Q1.1 | 5 |  |  |
| Q1.2 | 15 |  |  |
| Q1.3 | 5 |  |  |
| Q1.4 | 25 |  |  |
| Part 2 |  |  |  |
| a) Single Link |  |  |  |
| Updated Similarity Matrices | 15 |  |  |
| Dendrogram | 10 |  |  |
| b) Complete Link |  |  |  |
| Updated Similarity  Matrices | 15 |  |  |
| Dendrogram | 10 |  |  |
| TOTAL | 100 |  |  |