CST8390 - Lab 6
Clustering by k-Means

Due Date: Week 9 in corresponding lab sessions

Introduction

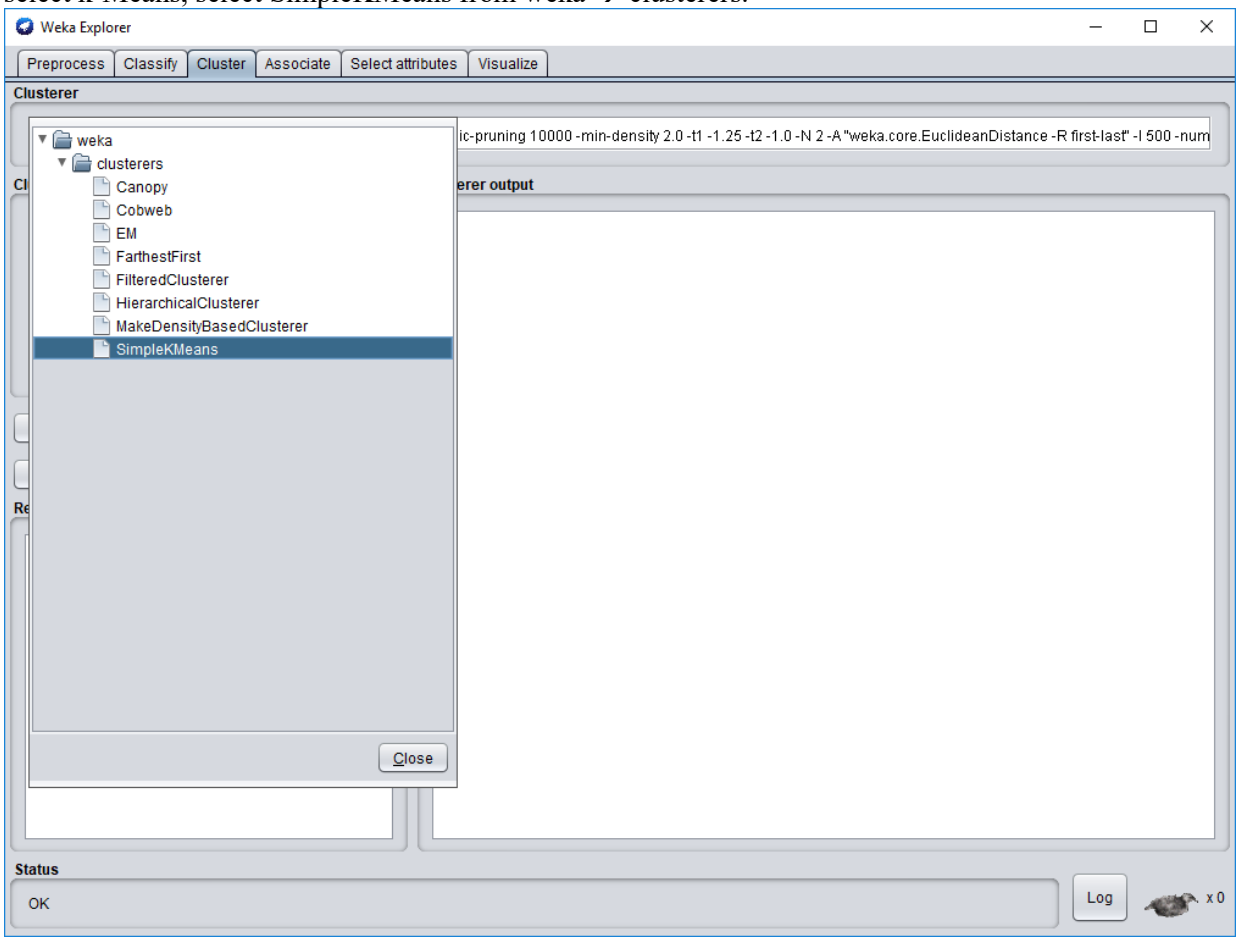
The goal of this lab is to perform clustering on wine dataset using kMeans.

Steps:

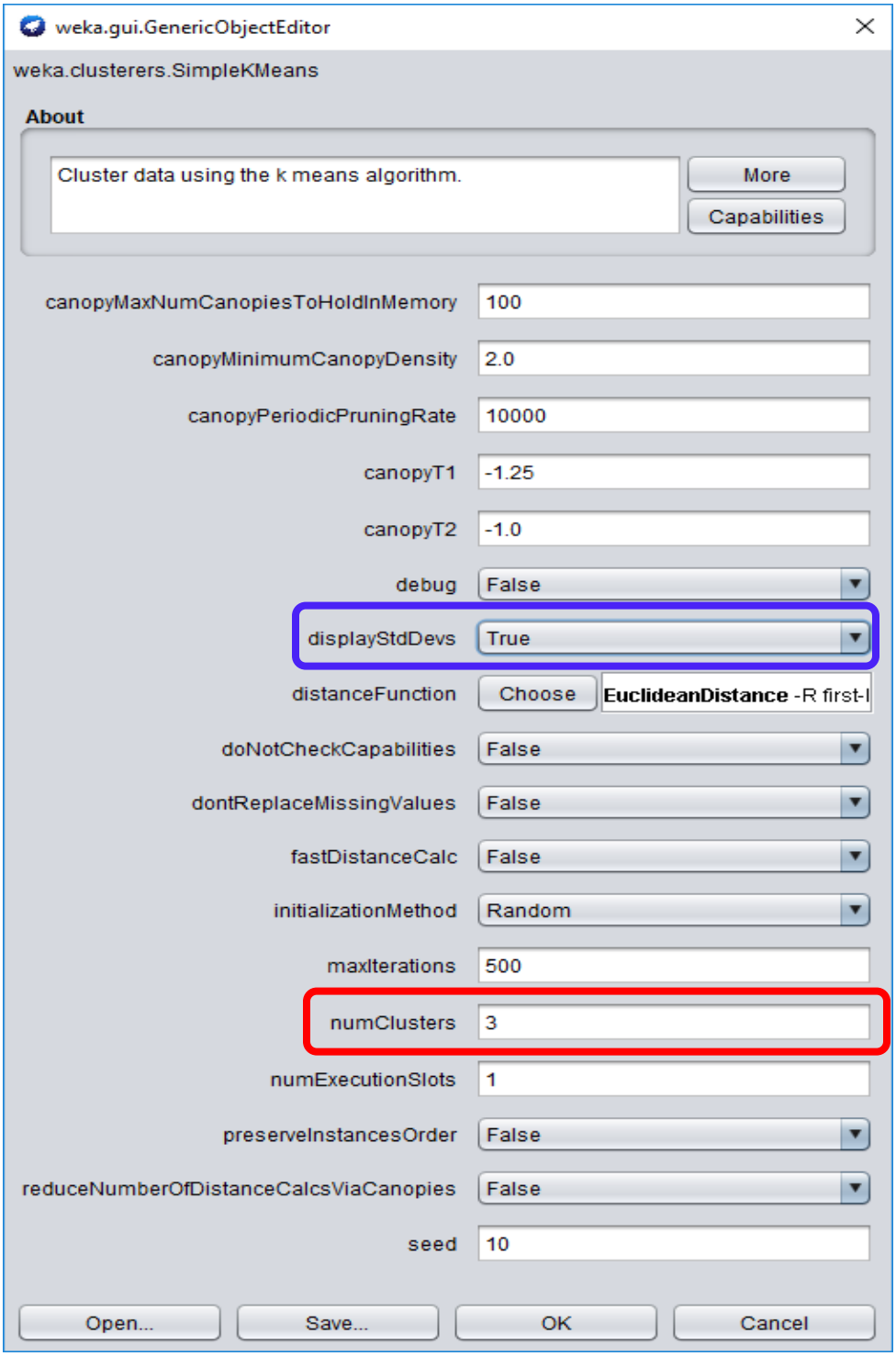
1. Load the Wine dataset that we used for lab 3 to Weka.

2. Check how various attributes are converted in Weka. Make sure that class attribute is nominal and all other attributes are numeric. If not, convert them using filters (refer lab 3).

3. Now, we need to perform clustering using k-Means method. For that, click on “Cluster” tab. To select k-Means, select SimpleKMeans from weka → clusterers.



4. Click on the selected k-Means to open the window with parameter list. As we know that wine dataset has 3 classes (1, 2 and 3), set numClusters to 3 (marked in red). As we need to see the standard deviation, set displayStdDevs to True (marked in blue) Close the window.



5. For “Cluster mode”, select “Classes to clusters evaluation” and select (Nom) Class (or name that you used for the first attribute which is the class). Now click “Start” to run the algorithm.

    a. How many iterations were needed for the centroid convergence?

    b. What method was used to replace missing values globally?

    c. How many instances are there in clusters 0, 1, and 2?

    d. What are the average Magnesium levels and the corresponding standard deviations for all the clusters? For each cluster, write in the format “average +/- sd”.
    (Example: 13.7193 +/- 0.4921)

    e. Compare the full data column with clustered data. Which cluster has below average Alcohol level?

    f. Look at the bottom of the result window and find the number of incorrectly classified instances.

    g. Which classes of wine were misclassified?

h. Which classes (1, 2, 3) of wine are represented by clusters 0, 1 and 2?

Class 1 –

Class 2 –

Class 3 –

6. Record the **initial centroids** of all clusters for attributes Alcohol and Color Intensity in the following table. Repeat clustering for seeds 5, 10, 15, 20, and 25.

Attribute		**Seed = 5**	**Seed = 10**	**Seed = 15**	**Seed = 20**	**Seed = 25**
Alcohol	Cluster 0					
	Cluster 1					
	Cluster 2					
Color Intensity	Cluster 0					
	Cluster 1					
	Cluster 2					

7. Record the initial and final centroids for Proline in the following table. Repeat clustering for seeds 5, 10, 15, 20, and 25.

Seed	Initial	Final

In order to get grades,

1. **Fill in your answers in the answer document available in Brightspace and upload it to Brightspace before submission due date.**
2. **You should be ready with all executions in Weka Explorer.**

Both demo during lab hours and submission in Brightspace are required to get credits for the lab.