CST8390 - Lab 5
Clustering by k-Means

Name: Min Li   **Id:** 040930563

Due Date: Week 6 in corresponding lab sessions (*after 1st Assignment*)

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 4** 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 **Labs 3, 4**).

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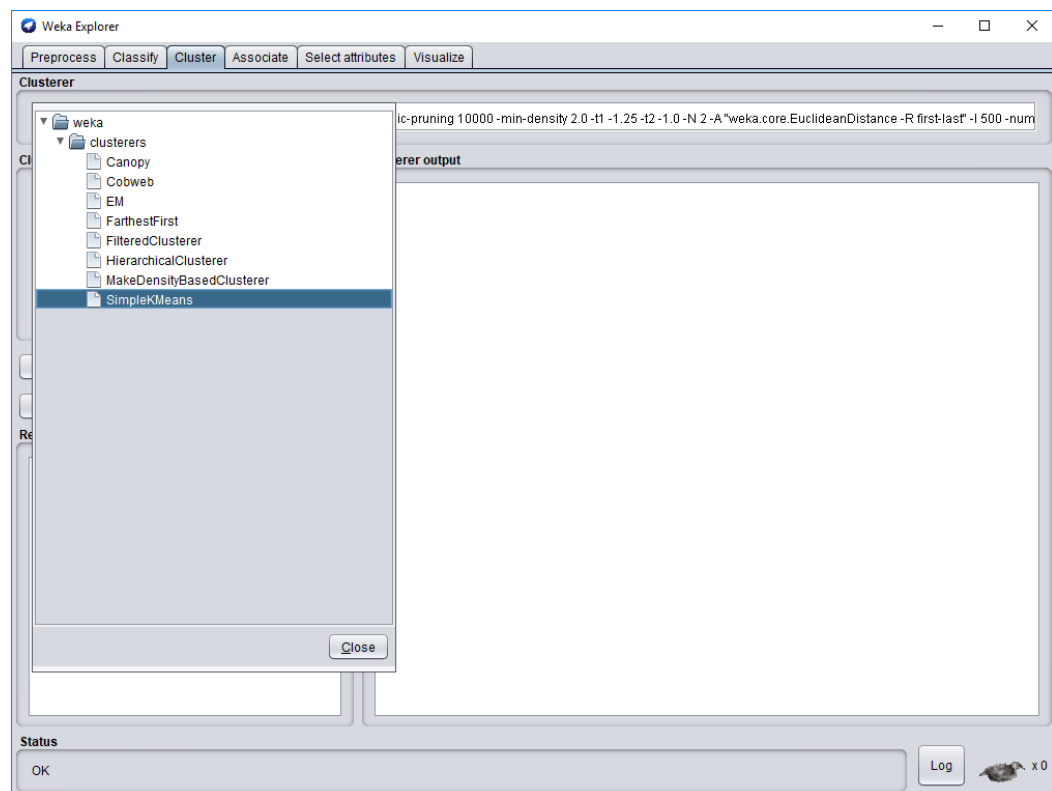
4. Click on the selected k-Means textbox to open the window with **parameter list**. As we know that wine dataset has 3 classes (1, 2 and 3), set numClusters to **3** (marked in red). As we need to see the **standard deviation**, set displayStdDevs to **True** (marked in blue). Close the window.

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? 8.

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

    c. How many instances are there in clusters 0, 1, and 2? 60,55,63.

    d. What are the average Alcohol levels and the corresponding standard deviations for all the clusters? For each cluster, write in the format "average +/- sd". Example: 13.7193 +/- 0.4921. 13.7193+/-0.4921, 13.0998+/-0.5297,   12.2295+/-0.5421.

    e. Compare the full data column with clustered data. Which cluster has below average Alcohol level? Cluster 2.

    f. Look at the bottom of the result window and find the number of incorrectly classified instances. 10

    g. Which classes of wine were misclassified? Class 1 and class 2

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

        Class 1 – Cluster 0.

        Class 2 – Cluster 2.

        Class 3 – Cluster 1.

6. Record the **initial centroids** of all clusters for attributes Malic Acid and Magnesium 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
Malic Acid	Cluster 0	3.26	1.72	1.17	3.55	5.65
	Cluster 1	3.03	1.29	1.77	0.98	3.8
	Cluster 2	1.35	1.35	1.01	2.16	3.86
Magnesium	Cluster 0	107	94	78	106	95
	Cluster 1	97	92	107	99	102
	Cluster 2	122	94	78	99	85

Show your answers to the lab professor when you are done.

FOR YOUR ANALYSIS:

* Option 1: *Explain with your own words the algorithm k-Means and in which situations you should choose it.*

* Option 2: *Change one (or more property) and describe how they can affect the results.*

Ottawa, Feb 2020.

Option 1: We have three kinds of color of rice, one is blue, one is red, one is yellow, the three kinds of rice are putting randomly, Now I have another rice, this rice is also putting randomly, I need to calculate this rice's k-Means, to know which color does it belong to.

In another way, in my opinion, choosing the algorithm k-Means to calculate the centroid, and then to see the which class does this point belongs to.

Option 2:   There are three properties "K" parameter, "seed"parameter, and "fold"parameter.