

## CST8390 - Lab 5

### *Clustering by k-Means*

Name:  - Id:

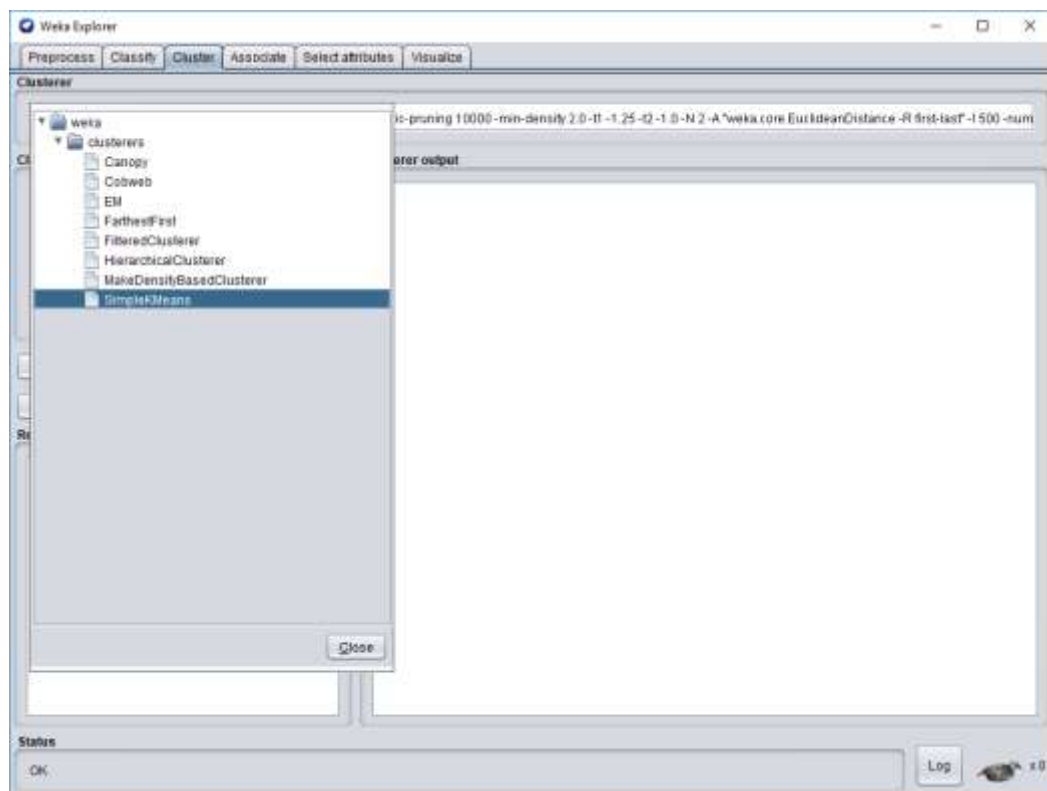
**Due Date:** Week 6 in corresponding lab sessions (*after 1<sup>st</sup> Assignment*)

#### 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.

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

**About**

Cluster data using the k means algorithm. More Capabilities

canopyMaxNumCanopiesToHoldInMemory 100

canopyMinimumCanopyDensity 2.0

canopyPeriodicPruningRate 10000

canopyT1 -1.25

canopyT2 -1.0

debug False

**displayStdDevs True**

distanceFunction Choose EuclideanDistance -R first-l

doNotCheckCapabilities False

dontReplaceMissingValues False

fastDistanceCalc False

initializationMethod Random

maxIterations 500

**numClusters 3**

numExecutionSlots 1

preserveInstancesOrder False

reduceNumberOfDistanceCalcsViaCanopies False

seed 10

Open... Save... OK Cancel

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.

- How many iterations were needed for the centroid convergence? .
- What method was used to replace missing values globally? .
- How many instances are there in clusters 0, 1, and 2? .
- 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. .
- Compare the full data column with clustered data. Which cluster has below average Alcohol level? .
- Look at the bottom of the result window and find the number of incorrectly classified instances. .
- Which classes of wine were misclassified? .
- 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 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					
	Cluster 1					
	Cluster 2					
Magnesium	Cluster 0					
	Cluster 1					
	Cluster 2					

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

**REMEMBER:**

Show your answers to the lab professor when you are done (in [Weka](#) and document).

***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.

---