MSc Data Science and Advanced Computing
CSMDM21 - Data Analytics and Mining

Practical 3: K-Means Clustering

Setup:
For the following tasks, you will use KNIME and need to install a new KNIME node, which provides a slightly modified version of the K-Means algorithm compared to the standard K-Means node in KNIME:

- Download from Blackboard the file "uk.ac.reading.cs.knime.kmeans_1.0.4.jar" and copy it into the folder "dropins" of your KNIME installation. For windows "dropins" folder can be at the path `C:\Program Files\KNIME\dropins` or at the path `C:\KNIME\dropins`
- Start KNIME.
- Verify that when you execute KNIME, the "KMeansWSS" node is available under the category (folder) "UoR" at the bottom of the "Node Repository" as shown in Figure 1. Instead of the standard node "*k-Means*" which is located in the category (folder) "/Analytics/Mining/Clustering", you should use to use this "*KMeansWSS*" node for this exercise: it provides the sum of squared errors measures WSS and BSS (cohesion and separation) in one of the output ports as well as in the console. (The standard k-Means node does not provide this information.)

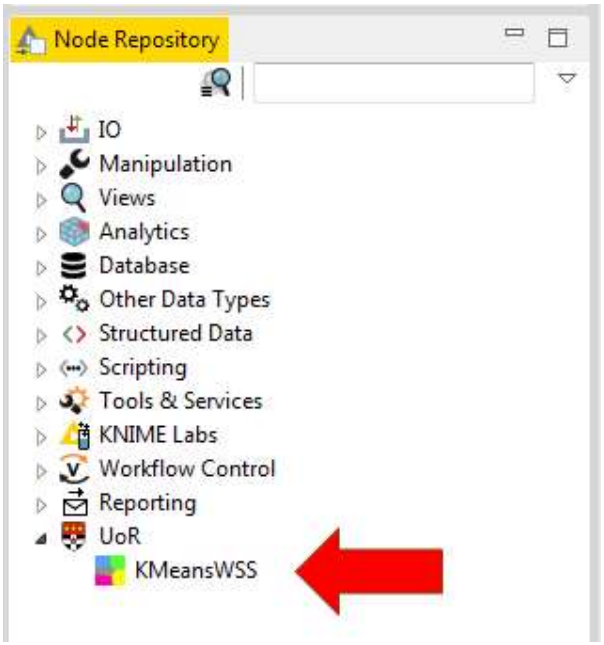

Figure 1: the new KNIME node KMeansWSS

Workflow activities/tasks:
Create a KNIME workflow to read the file iris.csv (download from Blackboard).

Perform a cluster analysis of the iris dataset using different numbers (K) of clusters and compare the results:

1. Use the node "KMeans" to compute the cluster assignments and the centroids with K=3, 12, 90, 150. The total Within-cluster Sum of Squares (WSS or total SSE) is displayed in the KNIME console.
2. Compute the Between-clusters Sum of Squares (BSS) for the same cases.
3. Report the WSS and BSS in the table below. Compare and comment the results.

K	WSS	BSS
3		
12		
90		
150		

Comments on the comparison:

4. For the cases K=3 and K=6: report the number of patterns of each cluster which belong to each class in the table below. Also report the Entropy for each cluster and the overall Entropy. Finally, compare and discuss the results.

K=3

*Cluster / **Class*** ➔	***versicolor***	***virginica***	***setosa***	Entropy
0				
1				
2				

Total Entropy:	

K=6

*Cluster / **Class*** ➔	***versicolor***	***virginica***	***setosa***	Entropy
0				
1				
2				
3				
4				
5				

Total Entropy:	

Discussion on the Entropy:

Solutions:

Sample solutions to these exercises are available on Blackboard (Bb) in two forms: images of the workflows and the actual KNIME workflows (as a single zip archive). You should first try to build your own workflows for each exercise. During the practical session, you may use the images to see the proposed solutions and to reconstruct them. During or at the end of the practical session, the archives with the actual KNIME workflows will be made available for you to import and test them. These are KNIME archives (file extension ". knwf ") that can be imported in KNIME. Note that after importing the KNIME archives, you may need to change the file locations of the source data file and the output file destination folder.