Universidade do Minho
Departamento de Informática

FE07

Course: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Knowledge Discovery

Exercise Sheet FE07	
Teacher	Cristiana Neto
Theme	RapidMiner – K-Means Clustering
Class	PL
Year	2019-20 – 2nd Semester
Duration	2 hours

1. Part I

[1] What does the k in k-Means clustering stand for?

[2] How are clusters identified? What process does RapidMiner use to define clusters and place observations in a given cluster?

[3] What does the Centroid Table tell the data miner? How do you interpret the values in a Centroid Table?

[4] How might the presence of outliers in the attributes of a data set influence the usefulness of a k-Means clustering model? What could be done to address the problem?

2. Part II

[1] Think of a problem that can be solved by grouping observations into clusters. Search the internet for a dataset that can be used and applied to a k-Means model. Suggestion: go to the UCI - Machine Learning Repository website and choose a dataset whose Default Task is Clustering.

(a) Import the data into RapidMiner. Do not forget to ensure that these are in CSV format. Perform the Data Understanding step.

(b) Perform the Data Preparation step. It can include data inconsistency components, missing values, or changing the data type.

(c) Add a k-means clustering operator to the dataset in RapidMiner and change the parameters as needed (especially the k value, to suit the problem in question);

(d) Study the Centroid Table, Folder View, and other assessment tools;

(e) Report all previous steps and the evidence found, as well as how your findings allow you to respond to the initial problem.

[2] Try the same dataset with different k-Means operators like Kernel or Fast. How they differ from the original model. Do these operators change the original clusters? If so, to what extent?