School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2019)
Workshop exercises: Week 8

Consider the following dataset:

id	*apple*	*ibm*	*lemon*	*sun*	**label**
A	4	0	1	1	fruit
B	5	0	5	2	fruit
C	2	5	0	0	comp
D	1	2	1	7	comp
E	2	0	3	1	?
F	1	0	1	0	?

1. Treat the problem as an unsupervised machine learning problem (excluding the *id* and *label* attributes), and calculate the clusters according to ***k*-means** with $k = 2$, using the Manhattan distance:
   (a) Starting with seeds A and D.
   (b) Starting with seeds A and F.

2. Perform **agglomerative clustering** of the above dataset (excluding the *id* and *label* attributes):
   (a) Using the *Euclidean distance* and calculating the group average as the *cluster centroid*.
   (b) Do you expect to observe a different dendrogram if we were instead using the *single link* method?

&

3. Using the same dataset
   (a) Classify the test instances (E and F) using the **1-NN** method. (using the Euclidean distance measure)
   (b) Classify the test instances (E and F) using the **weighted 3-NN** method. (using the Manhattan distance measure)

4. For the following dataset:

apple	*ibm*	*lemon*	*sun*	CLASS
TRAINING INSTANCES				
Y	N	Y	Y	FRUIT
Y	N	Y	Y	FRUIT
Y	Y	N	N	COMPUTER
Y	Y	Y	Y	COMPUTER
TEST INSTANCES				
Y	N	Y	Y	?
Y	N	Y	N	?

   Use the method of **Naive Bayes** classification, as shown in lectures, to classify the test instances. Revise some of the assumptions that are built into the model.