School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2018)
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), using the Euclidean distance and calculating the **group average** as the cluster centroid. Do you expect to observe a different dendrogram if we were instead using the cosine similarity?

&

3. What is **overfitting**? What does it mean for a classifier to **generalise**?

4. A **confusion matrix** is a summary of the performance of a (supervised) classifier over a set of development ("test") data, by counting the various instances:

		Actual			
		a	*b*	*c*	*d*
Classified	*a*	10	2	3	1
	b	2	5	3	1
	c	1	3	7	1
	d	3	0	3	5

   (a) Calculate the classification **accuracy** of the system. Find the **error rate** for the system.

   (b) Calculate the **precision**, **recall**, **F-score** (where $\beta = 1$), **sensitivity**, and **specificity** for class *d*. (Why can't we do this for the whole system? How can we consider the whole system?)

5. How is **holdout** evaluation different to **cross-validation** evaluation?