School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Tutorial exercises: Week 8

1. Revise the difference between **supervised** and **unsupervised** machine learning.

   Then, 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	?

2. Treat the problem as an unsupervised machine learning problem (excluding the *id* and LABEL attributes), and calculate the clusters according to (hard) k-**means** with $k = 2$, using the Manhattan distance:

   (a) Using seeds A and D.

   (b) Using seeds A and F.

3. Repeat the previous question using "soft" k-means, with a "stiffness" $\beta = 1$.

4. What is logic behind the **EM algorithm**, when used for clustering?

   (a) Explain the significance of the "E" step, and the "M" step.

5. What is **semi–supervised learning**, and when is it desirable?

   (a) What is **self training**?

   (b) What is the logic behind **active learning**, and what are some methods to choose instances for the **oracle**?