School of Computing and Information Systems The University of Melbourne
COMP30027 Machine Learning (Semester 1, 2021)

Tutorial: Week 12

1. [OPTIONAL] 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	?

   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, and instances A and F as the seeds.

2. Repeat the previous question using "soft" k-means, and the "stiffness" $\beta = 1$.

3. What is logic behind the EM algorithm, when used for clustering?
   (i) Explain the significance of the "E" step, and the "M" step.
   (ii) Identify the "E" and "M" steps in GMM methods.

4. Revise the concept of Unsupervised and Supervised evaluation for clustering evaluation
   (i) Explain the two main concepts that we use to measures the goodness of a clustering structure without respect to external information.
   (ii) Explain the two main concepts that we use to measure the how well do cluster labels match externally supplied class labels.

5. Revise the difference between **supervised**, **semi-supervised** and **unsupervised** machine learning. When do we use semi-supervised learning?
   (i) What is self-training?
   (ii) What is the logic behind active learning, and what are some methods to choose instances for the oracle?

6. One of the strategies for Query sampling was query-by-committee (QBC). Using the equation below, which captures vote entropy, determine the instance that our active learner would select first.

$$x^*_{VE} = \underset{x}{\text{argmax}}(-\sum_{y_i} \frac{V(y_i)}{C} log_2 \frac{V(y_i)}{C})$$

   Respectively y_i, $V(y_i), and\ C$ are the possible labels, the number of "votes" that a label receives from the classifiers, and the total number of classifiers.

classifier	Instance 1			Instance 2			Instance 3		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
C_1	0.2	0.7	0.1	0.2	0.7	0.1	0.6	0.1	0.3
C_2	0.1	0.3	0.6	0.2	0.6	0.2	0.21	0.21	0.58
C_3	0.8	0.1	0.1	0.05	0.9	0.05	0.75	0.01	0.24
C_4	0.3	0.5	0.2	0.1	0.8	0.1	0.1	0.28	0.62