School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2023)

Week 6

1. Given the following dataset, we wished to perform feature selection on this dataset, where the class is PLAY:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
G	o	c	n	T	?
H	s	m	h	F	?

   (i). Which of `Humi` and `Wind` has the greatest *Pointwise Mutual Information* for the class Y? What about N?

   (ii). Which of the attributes has the greatest *Mutual Information* for the class, as a whole?

2. 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	?

   (i). Treat the problem as an unsupervised machine learning problem and calculate the clusters according to k-means with k = 2, using the Manhattan distance, and instances A and F as starting seeds.

   (ii). 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.

3. 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.

4. [OPTIONAL] Using the dataset introduced in Question 2, consider the clusters, C1: {A, B, E} and C2: {C, D, F} and compare them with clusters C1': {A, B, E, F} and C2': {C, D}

   (i). Using the cohesion and separation of the clusters.

   (ii). Using the purity of the clusters.