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

Tutorial exercises: Week 6

ID	A (°C)	B (mm)	C (hPa)	CLASS
1	22.5	4.6	1021.2	AUT
2	16.7	21.6	1027.0	AUT
3	29.6	0.0	1012.5	SUM
4	33.0	0.0	1010.4	SUM
5	13.2	16.4	1019.5	SPR
6	14.9	8.6	1016.4	SPR
7	18.3	7.8	995.4	WIN
8	16.0	5.6	1012.8	WIN

1. What is **Discretisation**, and where might it be used?

   (a) Summarise some approaches to **supervised** discretisation.

   (b) Discretise the above dataset according to the (unsupervised) methods of **equal width**, **equal frequency**, and **k-means** (breaking ties where necessary).

2. Find the (sample) **mean** and (sample) **standard deviation**[1] for the attibutes in the above dataset:

   (a) In its entirety, and;

   (b) For each individual class[2].

   (c) How could we use this information when building a classifier over this data?

   Given the following dataset:

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

3. If we wished to perform **feature selection** (or **feature weighting**) on this dataset, where the class is PLAY:

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

   (b) Which of the attributes has the greatest **Mutual Information** for the class, as a whole? (Note that we need to extend the lecture definition to handle non–binary attributes.)

[1] n.b. You might need a calculator.

[2] We would ideally do this with more instances!