School of Computing and Information Systems
The University of Melbourne
COMP90049 INTROCTION to MACHINE LEARNING
(Semester 2, 2020)

Tutorial exercises: Week 4

NOTE: *Questions marked OPTIONAL will only be covered during the workshop if time permits. They are added to the workshop as an additional opportunity for individual practice. The solution to all questions and a recorded sample workshop will be published on Fridays after the last workshop.*

1. 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

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

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

2. For the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	n	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
TEST INSTANCES					
G	o	m	n	T	?
H	?	h	?	F	?

   (i). Classify the test instances using the method of **0-R**.

   (ii). *[OPTIONAL]* Classify the test instances using the method of **1-R**. (for H assume *Outl* = *s*)

3. Given the above dataset, build a Naïve Bayes model for the given training instances.

4. Using the Naïve Bayes model that you developed in question 4, classify the given test instances.

   (i). No smoothing.

   (ii). Using the "epsilon" smoothing method.

   (iii). *[OPTIONAL]* Using "Laplace" smoothing ($\alpha = 1$)

5. *[OPTIONAL]* How is **holdout** evaluation different to **cross-validation** evaluation? What are some reasons we would prefer one strategy over the other?