School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2017)
Workshop exercises: Week 8

1. What is **overfitting**? What does it mean for a classifier to **generalise**?

2. A **confusion matrix** is an indication of the performance of a classifier over a set of test data, by counting the various output 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

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

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

3. How is **holdout** evaluation different to **cross-validation** evaluation?

4. Revise **linear regression**.

   (a) What are we attempting to model with linear regression? Why do we minimise "RSS"? What assumptions are we making?

5. For the following dataset:

apple	*ibm*	*lemon*	*sun*	class
Training Instances				
4	0	1	1	fruit
5	0	5	2	fruit
2	5	0	0	computer
1	2	1	7	computer
Test Instances				
2	0	3	1	?
1	0	1	0	?

   (a) Using the **Euclidean distance** measure, classify the test instances using the 1-NN method.

   (b) Using the **cosine similarity** measure, classify the test instances using the 3-NN method. Extend this to the **weighted** 3-NN method.

6. For the following dataset:

apple	*ibm*	*lemon*	*sun*	CLASS
Training Instances				
Y	N	Y	Y	FRUIT
Y	N	Y	Y	FRUIT
Y	Y	N	N	COMPUTER
Y	Y	Y	Y	COMPUTER
Test Instances				
Y	N	Y	Y	?
Y	N	Y	N	?

   Use the method of **Naive Bayes** classification, as shown in lectures, to classify the test instances. Revise some of the assumptions that are built into the model.

7. [EXTENSION] Revise the **multinomial distribution**. Naive Bayes can be extended to account for integer frequencies in the data (like in Question 2) using this model. Read up on so-called **multinomial Naive Bayes**.