RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Tutorial Problems Week 2

1. In the context of classification, what is the purpose of

   (a) The training set

   (b) The test set

   (c) Suppose you had a data set of 100,000,000 examples. How would you split them into training and test?

2. Consider the following output from a run of the OneR classifier:

```
=== Run information ===
Scheme:       weka.classifiers.rules.OneR -B 10
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
petallength:
        < 2.45  -> Iris-setosa
        < 4.95  -> Iris-versicolor
        >= 4.95 -> Iris-virginica
(142/150 instances correct)
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         139               92.6667 %
Incorrectly Classified Instances        11                7.3333 %

Total Number of Instances              150

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 43  7 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

(a) Explain the classification rule.

(b) How would the following unknown instance be classified by this rule: SepalLength 5.4, SepalWidth 3.9, PetalLength 1.3, PetalWidth 0.4?

(c) In the first part of the output 142/150 instances are given as correct. Later 139 instances are given as correct. Explain this discrepancy.

(d) What does the entry "7" in the confusion table mean?

3. Apply the OneR algorithm to the animal data of last week (tutorial01.pdf, question 4). Use only the attributes Body Temperature, Skin Cover and Gives Birth. What is the rule and what is its accuracy on the training data?

4. Develop a OneR algorithm for numeric attributes. [Hint: Take the first column of the babies Feb data from tutorial01, write the numbers in sorted order and next to each number put the corresponding class. Construct a rule from this.]

5. It is clear how the distance between two numeric instances can be calculated. Is it possible to have a distance measure between two symbolic examples, for example, two instances from the mushroom data set?

6. Consider the application of the nearest neighbour classifier to this data. Will there be a problem applying the nearest neighbour technique to this data? If so, what can be done about it?

Age	WhiteBloodCell	Class
20	0.00005	OK
50	0.00002	ILL
90	0.00004	OK
.	.	.
.	.	.

7. Suppose you run a classifier on a two-class data set and the training accuracy is 99% and the test accuracy is 70%. What would you conclude?

8. Suppose you run a classifier on a two-class data set and the training accuracy is 53% and the test accuracy is 52%. What would you conclude?

9. Is it possible for the test error to be smaller than the training error?