Workshop Week 8: COMP20008 2019

1. What is classification? What is regression? What is the difference between the two?

2. What is the difference between training data and testing data?

3. Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

   - Write a formula for the information gain when splitting on feature A.
   - Write a formula for the information gain when splitting on feature B.
   - Which feature would the decision tree induction algorithm choose?

4. Consider the following simple dataset

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

   - Classify the point x=5.0 according to its 1-, 3-, 5- and 9-nearest neighbors.
   - How does the parameter k affect the k-NN classifier? What would be the behaviour as $k \to \infty$?

5. The algorithm discussed in lectures for using a decision tree to classify an instance, did not consider the situation where the test instance may having missing feature values. Describe two ways one could use a decision tree to make a classification in this situation.

6. (Harder) Suppose Alice takes a dataset D with 100 instances, 4 features, plus a class label feature. She computes the correlation of each of the 4 features with the class label using mutual information and discards the two features with lowest correlation. She

now has a processed version D' of the dataset (2 features, class label feature and 100 instances). She splits D' into two - 80% training (80 instances) and 20% testing (20 instances). She learns a decision tree model on the training set and evaluates the model accuracy on the testing set. She reports the accuracy as being 90%. Why might this estimate of 90% accuracy be over-optimistic? Give reasons.

7. Load the workshop-week8-2019.ipynb jupyter notebook and complete the two practical exercises.