COMP20008 2020 SM1 Workshop Week 9: Classification

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

2. 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?

3. 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 neighbours.
   - How does the parameter k affect the k-NN classifier? What would be the behavior as $k \to \infty$

4. 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.

5. Load the **2020SM1-workshop-week9-classification.ipynb** jupyter notebook and complete the two practical exercises.