Name: EE4211 Homework H2, Fall 2020

Instructions: Please complete and submit your work to the appropriate folder in LumiNUS. You may work in study groups, but each student must be responsible for their own submission.

Please submit all the following documents as a single zip file named StudentID-Name-HW1.zip:

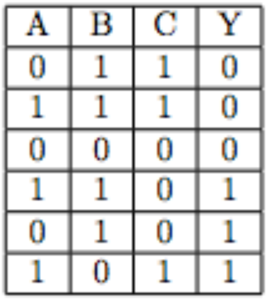
(i) Completed Word file named as StudentID-Name-HW2.docx (with all results)

(ii) Print preview of ipynb file named as StudentID-Name-HW2.pdf (with results)

(iii) Working ipynb file named as StudentID-Name-HW2.ipynb

1. I mentioned in lecture that the number of possible decision trees is very large. How many decision trees exist with n binary attributes? Here is way to think about the problem.
   - Suppose you have one binary attribute. Then there are 2^1=2 possible values for the attribute and each of those values can be mapped to 2 outputs, so there are 4 decision trees.
   - Suppose you have two binary attributes. Then there are 2^2=4 possible values for the pair of attributes, and each value can be mapped to 2 outputs, so there are 2^4=16 decision trees.
   - Now suppose you have n attributes. How many possible decision trees are there? Please justify your answer.

2. Consider the following training set with features A, B, C, and target/label Y.
   a. What is the entropy of the output Y?
   b. Using the information gain criterion, what is the first node you would split at? Explain clearly why?
   c. Using the information gain criterion, complete the learning of the decision tree for this dataset. Draw the decision tree and comment if the tree is unique.



A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

3. In this problem, we will look at the Breast Cancer Wisconsin (Diagnostic) Data Set available UCI Machine Learning Repository. Please use the wdbc.data dataset from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

   - Compute the performance of the DT algorithm on this dataset for predicting the whether the cancer is malignant or benign. Use a random train/test data split of 70%/30%. Repeat this process 20 times and compute the average performance.
   - Please evaluate the following algorithms:
     - DT1: DT with Information Gain (IG)
     - DT2: DT with IG & limited tree size, vary the number of levels and try to beat DT1
   - Please compute the following metrics and fill in the table below.
     - Training Accuracy and Test Accuracy
     - Precision and Recall (which are important metrics that complement Accuracy)
     - You can read about performance metrics at: https://en.wikipedia.org/wiki/Confusion_matrix
     - SKLearn contains functions to compute these metrics: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

	Accuracy		Precision	Recall
	Train	Test		
DT1				
DT2				