**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.

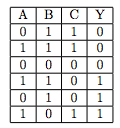
Please submit all the following documents as a single zip file named StudentID-Name-H2.zip:

1. Completed Word file named as StudentID-Name-H2.docx (with all results)
2. Print preview of ipynb file named as StudentID-Name-H2.pdf (with results)
3. Working ipynb file named as StudentID-Name-H2.ipynb
4. 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.

**Ans 1:** The formula to calculate the total number of decision trees would be **22^n** and I calculated this formula as follows:

* **Step 1:** Total number of binary attributes = 2n as 21 = 2 values for 1 binary attribute, then 22 = 4 values for 2 binary attributes, then 23 = 8 values for 3 binary attributes and so on. Thus, if the number of binary attributes = n, then the total number of binary attributes will be 2n as 2,4,8 etc.… work in powers of 2.
* **Step 2:** As each of these values are mapped onto 2 outputs and in the previous examples, it was mentioned that for 1 binary attribute, it has 21 = 2 values which in turn has 4 decision trees which means that 22^1 = 22 = 4 decision trees. Then, for 2 binary attributes, it has 22 = 4 values which in turn has 16 decision trees which means that 22^2 = 24 decision trees. In the lecture then, for 6 binary attributes, it means that there will be 26 = 64 values and if they are mapped to 2 outputs, it means that there will be 22^6 = 264 = 1.84 X 1019 decision trees as mentioned in the A08 decision trees lecture video.
* **Hence, the formula for the number of decision trees for n possible attributes is 22^n.**



1. Consider the following training set with features A, B, C, and target/label Y.
2. What is the entropy of the output Y?
3. Using the information gain criterion, what is the first node you would split at? Explain clearly why?
4. 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.

**Ans 2:**

1. The entropy of the output Y is calculated using the following formula from the lecture notes:

A picture containing text, clock, watch

Description automatically generated

Hence, the calculated value is: –(0.5log20.5)-(0.5log20.5) = **1**

1. Here are my calculations for the Information Gain (IG) for A, B & C respectively using Method 2:

= x log2 + x log2  + x log2 + x log2

= x log2 + x log2  + x log2 + x log2

= x log2 + x log2  + x log2 + x log2

Since both A and C has the same information gain i.e., 0.0817, we can choose to split at either A or C. In this case, if there is a tiebreaker, we can choose the least value lexicographically i.e., in the alphabetical order, so we can choose A in this case.

1. I would like to point out that the decision tree will **NOT** be unique as A and C has the same information gain. In addition to that detail, both B and the other attribute have the same information gain too when A is chosen first. If I start dividing the root with B or C, I can have different trees. The decision trees if split by A & C is shown below.

A piece of paper with writing on it

Description automatically generated with medium confidence

1. 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 | **1.0** | **0.9359649122807017** | **0.9117556838453199** | **0.91953125** |
| DT2 | **1.0** | **0.9388888888888889** | **0.9135427054253397** | **0.92578125** |