COMP 4331: Spring 2020
Assignment 2

Deadline: 23:59 7th of April, 2020

1 Submission Guidelines

- Assignments (including all attachments) should be emailed to comp4331spring2020@gmail.com before the deadline. All submissions after the deadline will incur a penalty of 20% of your marks for the first day (24 hours) and a further 10% for each hour it is late beyond the first day.

- Your final submission, titled A2_stuid.zip (where stuid is your itsc student id i.e. atzhou [**Note: this is not your student number**]) should be zipped file composed of two files:

  1. **A2_stuid_report.pdf** or **A2_stuid_report.docx**: Your report detailing your environment and experimental outcomes. This file **must** be either a pdf or docx file.

  2. **A2_stuid_code.zip**: A zipped file containing all source code used to complete this assignment. All code should be compilable and well-commented. Please either provide a separate file for each technique as mentioned in section 2.

- By submitting any work, you are acknowledging that you are the sole contributor of the work unless specified and properly referenced. **All plagiarism will not be tolerated and incur a mark of 0.**

- Your work will be graded on correctness, efficiency and clarity of communication.

- To inquire about any questions you may have regarding the assignment, please email:

  atzhou@connect.ust.hk or jfangak@connect.ust.hk

2 Classifier Implementation

Using the dataset (train.txt) found "Files → Assignment 2", you are to implement and train the following classifier models. Your implementation should then predict the X value for each input in the test dataset (test.txt) also found in "Files → Assignment 2".

2.1 Decision Tree Classifier

You are to implement two Decision Tree Classifiers in Python 2/3. The two Decision Trees you should implement are:

- **ID3 Decision Tree:** A decision tree which uses the ID3 Impurity Measurement ($Gain(A,T) = Info(T) - Info(A,T)$) (**15 Marks**)

- **C4.5 Decision Tree:** A decision tree which uses the C4.5 Impurity Measurement ($Gain(A,T) = \frac{Info(T)-Info(A,T)}{SplitInfo(A)}$) (**15 Marks**)

Your two implementation may share all other code except for the requisite impurity measurement. You may also use additional .py files and call functions from them to avoid redundancy.

Make sure to report each of predicted values (from the test set) for both classification methods in your report.

2.2 Naive Bayes Classifier

You are to implement a Naive Bayes Classifier in Python 2/3. Your method should the posterior probabilities of each permutation of attributes from the data. Once again, make sure to report the predicted values (from the test set) in your report. (**30 Marks**)

2.3 Report

You are required to report the **running time** of each method. Also, the **environment** you use (System, CPU, RAM, etc) should be provided.
Discuss all differences between predictions of the multiple classifiers, giving sound reasoning regarding the occurrences (or in the case of no differences between predictions, discuss why this has occurred). Having observed the predicted values from all three classifiers, please write down what you believe to be the predicted output for each data point in the test file. How did you reach this conclusion? (**40 Marks**)

2.4 Data Description

"train.txt" contains the training set, and "test.txt" contains the test set. The first line in "train.txt" and "test.txt" is the headers. The training set contains 12950 instances, and the test set contains 10 instances. Every instance has 8 attributes (parents, has_nurs, form, children, housing, finance, social, health) and 1 class value (NURSERY). The detailed data description can be found in "dataAttributes.txt".