THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331: Introduction to Data Mining
Fall 2021 Assignment 2
Due time and date: 11:59pm, November 6 (Sun), 2022

IMPORTANT NOTES

- **Your grade will be based on the correctness and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **ZERO-Tolerance on Plagiarism: All involved parties will get zero mark.**

Q1. In this question, you will run the decision tree on the following dataset (a csv version is available on canvas). The first 4 samples are used for testing, while the rest are used for training.

type	**difficulty**	**learned_before**	**completeness**	class: **proper_question**
math	hard	yes	poor	no
chemistry	hard	no	good	no
language	hard	yes	good	yes
language	medium	no	poor	yes
math	easy	yes	poor	yes
language	easy	no	good	no
chemistry	easy	yes	poor	yes
math	medium	no	poor	no
math	easy	yes	poor	yes
chemistry	medium	no	poor	yes
language	medium	yes	good	yes
math	medium	no	poor	yes
chemistry	hard	yes	good	yes
math	medium	yes	poor	no

(a) Using sklearn.tree.DecisionTreeClassifier, learn a decision tree using the **information gain** with min_samples_split (the minimum number of samples required to split an internal node) equals 4. Always set random_state to 1 so as to obtain a deterministic behavior. Show (i) the tree obtained using plot_tree; (ii) training accuracy; and (iii) testing accuracy.

(b) Vary min_samples_split in the range $\{2, 3, 4, 5, 6, 7, 8, 9\}$. Plot the training and testing accuracies (on the y-axis) versus the value of min_samples_split (on the x-axis). What is the **min_samples_split** value with the best testing accuracy.

(c) Repeat part (a) by using the **gini** index.

(d) Consider the use of **information gain**, work out by hand which attribute is to be selected at the root of the decision tree.

(e) Repeat part (d) with the **gain ratio**.

(f) Repeat part (d) with the **gini index**.

Q2. In this question, you are required to build a naive Bayes classifier to classify the sentiment of the movie review as positive or negative. The data set Q2.csv can be downloaded from canvas. Class label 0 means positive sentiment, while class label 1 means negative sentiment. Parts of the code are provided in the template.

(a) Using sklearn.naive_bayes and the training/testing splits in the template, learn a naive Bayes classifier **without** Laplace correction. Show the (i) training accuracy; and (ii) testing accuracy.

(b) What is the probability of seeing the word "excellent" given that the label is positive? What is the probability of seeing the word "excellent" given that the training label is negative? Repeat the above for the word "terrible". Do the probabilities obtained agree with what you expect?

Submission Guidelines

Please submit

(i) a report report.pdf includes your the results for Q1 and Q2.

(ii) a python notebook assignment2.ipynb for your code. Note that you should use the provided template.

Zip all the files to YourStudentID_assignment2.zip (e.g., 12345678_assignment2.zip). Please submit the assignment by uploading the compressed file to Canvas. Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. **Plagiarism will lead to zero point on this assignment.**