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, Nov 4 (Thur), 2021.

IMPORTANT NOTES

- **Your grade will be based on the correctness, efficiency 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.**

In this assignment, you will try both the decision tree and naive Bayes classifier on the following dataset (a csv version is available on canvas). The first 4 samples are used for testing while the others are used for training.

age	income	student	credit_rating	class_buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Q1. Decision tree using sklearn.tree.DecisionTreeClassifier (always set random_state to 1 so as to obtain a deterministic behavior).

(a) Learn a decision tree using the **information gain**, with min_samples_split (the minimum number of samples required to split an internal node) equals 2. Show (i) the tree obtained; (ii) training accuracy; and (iii) testing accuracy.

(b) Consider adding the following noisy sample to the training set.

age	income	student	credit_rating	class_buys_computer
middle_aged	low	no	fair	no

Repeat part (a) with the expanded training set. Compare and contrast with the solution you obtained in part (a).

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

Q2. Naive Bayes classifier using sklearn.naive_bayes (do not use the Laplace correction).

(a) Similar to Q1, part (a), learn a naive Bayes classifier on the original data set. Show the (i) training accuracy; and (ii) testing accuracy.

(b) Similar to Q1, part (b), learn a naive Bayes classifier on the noisy data set. Show the (i) training accuracy; and (ii) testing accuracy.

Submission Guidelines

Please submit
(i) a report report.pdf includes your results for Q1 and Q2.
(ii) a python notebook assignment2.ipynb for your code. You may use the 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.**