Pattern Recognition and Machine Learning
(Winter 2022)
Assignment 2: Decision Trees

Early Bird Submission Deadline: **Jan 30**, 2022, 23:59 (No penalty)
Late Submission Deadline: **Feb 1**, 2022, 23:59 (20% penalty)
Final deadline: **Feb 2**, 2022, 23:59 (addl. 10% penalty, total penalty = 30%)

Guidelines for submission

1. Perform all tasks in a single colab file.
2. Create a report regarding the steps followed while performing the given tasks. The report should not include excessive unscaled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Submit a zip with the colab file [.ipynb] and report [.pdf] on the classroom
5. Submit the .py file on the floated form for the laboratory
6. Plagiarism will not be tolerated

Question 1. **[70 marks]**

This question is designed to help you understand the working of a decision tree (DT). You will find a dataset, here, containing information used to classify penguins into 3 species.
You need to implement a classification decision tree (DT) from *scratch [you **are not permitted** to use any 3rd-party library's function for the classifier e.g. scikit. You may, however, use built-in functions for auxiliary tasks like train/test split, etc.].*

- With respect to the cost function to be used to find splits, everywhere in this question, you have to implement one of the following based on your roll number:
    - Odd roll numbers - Gini index
      Even roll numbers - entropy

The implementation includes the following tasks. -

1. Perform pre-processing and visualization of the dataset. Perform categorical encoding wherever applicable and split the data into train and test sets - **[7 marks].**
2. Implement the cost function as per your roll num. (details above) - **[5 marks].**
3. In order for the decision tree to work successfully, continuous variables need to be converted to categorical variables first. To do this, you need to implement a decision function that makes this split. Let us call that **cont_to_cat().** The details of the function are the following. **[10 marks]:-**
    a. Assume that the continuous variables are independent of each other i.e. assuming 2 continuous variables A and B, the split of A does not in any way affect the split you will perform in B.

    b. The continuous variables should only be split into **2 categories**, and the **optimal split** is one that divides the samples the best, based on the value of the function you have been allotted (as per your roll number).
4. After step 2, all the attributes would have categorical values, so now you can go ahead and implement the training function. This would include implementing the following helper functions: **[25 marks]**
    a. Get the attribute that leads to the best split
    b. Make that split
    c. Repeat these steps for the newly-created split
5. The DT should also include the following properties in the **train** function
    a. There should be a max depth that should be defined i.e. a depth after which the tree shouldn't be allowed to grow **[5 marks]**
    b. The algorithm should self-identify when there is no information gain being done, i.e. the model has plateaued in it's training and shouldn't grow further. **[5 marks]**
6. Write a function which is responsible for **classification** (i.e. at test time)**. [10 marks]**
7. Find out the accuracy you get on the test data (overall and class-wise). **[3 marks]**

Note: *Implementing the wrong cost function will attract a loss of credit in the above question.*

Question 2. **[30 marks]**

This question involves performing regression using a decision tree. You are permitted to use the *sklearn* library for this question. The dataset involves energy analysis, and this problem is centered around understanding how machine learning is applied in the industry.

The dataset contains eight attributes (X1, …, X8) (representative of different properties of buildings like height, roof area, etc.) and one response (Y1) (the heating load for the building). The aim is to use the eight features to predict Y1. *(You **are permitted** to use 3rd-party libraries for this question).*

The tasks for this question are the following **:-**

1. Preprocess the data. Split it using a 70:10:20 ratio, which represents training:validation:testing. - **[5 marks]**
2. Write a function to train the data using a regression decision tree. The function varies hyper-parameters to find the tree that generalizes best (based on its performance on the validation set). So, you need to train on the 70% training data and check performance on the 10% validation data.
   Properly explain the thought process behind which hyper-parameters you vary and the expected effects in the report. Make plots of validation MSE to support your arguments. **[20 marks]**
   Note: Here we are evaluating the method involved in the generalization rather than the achieved accuracy. Please refrain from using any inbuilt scikit library functions for hyper-parameterization (eg:- grid search).
3. Perform ***5-fold cross-validation*** using the **optimal** hyper-parameters decided in the previous question. Finally, calculate the mean squared error between the predicted and the ground-truth values in the ***test data*** for your best model. Also, plot the decision tree created **[5 marks].**

For more information about the datasets, you can refer to their sources:

- For problem 1: Penguins dataset
- For problem 2: Energy Efficiency dataset

Guidelines for the Report

1. The visualization of the dataset required in problem 1 should be computed as subplots in the colab file and *any 2 pairs* of relevant features should be added in the pdf.
2. The report should be to the point. Justify the space you use!
3. Explanations for each task should be included in the report. You should know the '*why*' behind whatever you do.
