Assignment- 5 (Due date: 14 Sep, 11:59 PM)
CSL2010: Introduction To Machine Learning
AY 2022-23

Decision Trees

General Instructions

1. You need to upload a zip <Lab5_Your_Roll_No>.zip, which contains one file for the task in <Lab5_Your_Roll_No>.py format and the report for the entire assignment in <Lab5_Your_Roll_No>.pdf format.
2. Provide your colab file link in the report. Make sure that your file is accessible.
3. Submit a single report mentioning your observations for all the tasks.
4. Report/Cite any resources you have used while attempting the assignment.
5. Plagiarism of any form will be penalized. If you take help from any source, cite it as a reference in your report.
6. Q.2 is for your practice and will not be graded.

Question 1: Decision Tree Classifier Total points: 25

The question is designed to help you understand the working of decision trees. You will be using the dataset given here to classify three species of penguins. Use entropy as the function to find splits.

a. Perform preprocessing and data visualization. You may make scatter plots or histograms and take care of missing data. [5]
b. Find ordinal/ nominal and categorical features, and convert them into numerical equivalent. [5]
c. Split the dataset into 70:20:10. Train the decision tree classifier on 70% and validate it on 20%. [10]
   Try with different values of the following parameters:
   i. max_depth
   ii. Min_samples_leaf

   Analyze how they affect the training and validation accuracy and report. Properly explain the thought process behind which hyper-parameters you vary and the expected effects in the report.

d. Choose your best model and report its testing accuracy. Also, provide graphical visualization of this tree and comment on overfitting. **[5]**

Q2. Decision Tree Regressor (NON-GRADED)

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.

The tasks for this question are the following:-

1. Preprocess the data and split it ratio is 70:20:10.
2. Write a function to train the data using a regression decision tree. Vary the hyper-parameters to find the tree that generalizes best (based on its performance on the validation set). So, you need to train on training data and check performance on the validation data. Understand which hyper-parameters the accuracy.
3. Analyze and make plots of mean squared error on the validation set to support your arguments.
4. Finally, calculate the mean squared error between the predicted and the ground-truth values in the test data for your best model.
