Pattern Recognition and Machine Learning

Lab - 2 Assignment

Early Bird Submission Deadline: Jan 27, 11:59 PM
Late Submission Deadline: Jan 28, 2023, 12:00 Midnight (20% penalty)
Late Submission Deadline: Jan 29, 2023, 23:59 (addl. 10% penalty, total penalty = 30%)
Final deadline: Jan 30, 2023, 23:59 (addl. 10% penalty, total penalty = 40%)

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 the colab[.ipynb], python[.py] and report[.pdf] files here : Link
5. Plagiarism will not be tolerated

Question 1. [30 marks][In-Lab Submission] 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. **[10 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 ***Hold-out cross validation***, ***5-fold cross-validation and repeated-5-fold 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 **[10 marks].**

4. Use **L1** and **L2** as two different *criterion* for split and plot the decision boundary you obtain. Which loss works better?Explain why one of them performed better on the given dataset. **[5 marks]**

Question 2. [70 marks]

Dataset Provided: Iris-Dataset

- This problem consists of two parts: Classification and Regression using Decision Trees.

Classification [50 marks]

For this task we will use **only Petal length and Petal width** attributes from the Iris dataset. Split the dataset into **training and test in the ratio 80:20**.

1. Train a Decision Tree Classifier (max depth=2) on the pre-processed dataset. Plot the decision boundaries of the tree as well as indicate the depth at which each split was made.**[15 marks]**
2. Remove the widest Iris-Versicolor from the iris training set (the one with petals 4.8 cm long and 1.8 cm wide) and train a new Decision Tree. Plot the Decision boundary for this case.**[10 marks]**
3. Train a Decision Tree Classifier with (max-depth = None) on the pre-processed dataset. Plot the Decision boundary for the same. Compare and analyse the results with those in part 1.**[5 marks]**
4. Create a random dataset having 2 attributes(X1 and X2), and 2 classes (y=0 and y=1) .X1,X2 are randomly sampled from the range (0,5). y=0 when X1<2.5, and y=1 when X1>2.5. The dataset should have 100 data points for both the classes. Train a decision tree for such a dataset(max-depth=2). Plot the obtained decision boundaries.Now, rotate the datapoints by 45 degrees in clockwise direction about the origin (X1=0,X2=0). Train another decision tree classifier using sklearn. Compare the plots obtained in both the above methods.**[15 marks]**
5. Put forward your observations about the behaviour of Decision Tree Classifier as seen in tasks 2,3 and 4. **[5 marks]**

Regression [20 marks]

This task involves working on the dataset provided in this link. Perform the following tasks:

1. Train two decision tree models, one with *max_depth* = **2** and another with *max_depth* = **3**. Plot the regression predictions at each depth for each max_depth(for e.g., at depths 0,1 for max_depth = 2) using a line plot. Next, make a scatter plot of the data points on the same plots to give a detailed analysis of what you see in the report along with the plots. **[10 marks]**
2. Plot the data points and make a line graph to show the decision tree fits on the dataset in two

cases: *min_samples_leaf* = **0** and *min_samples_leaf* = **10.** Analyze the two plots and explain your findings in the report. **[10 marks]**

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

- For problem 1: Energy-Efficiency dataset
- For problem 2: Iris-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 a*ny 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.