Machine Learning CS 550

Programming Assignment 2A: Basic ML

Total Marks: 150 (obtained marks will be divided by 10 for grading)

Note: This is an assignment to help those students who are new to ML and are facing difficulty in solving Assignment 1 or 2. You can submit this instead of Assignment with a comment. Note that maximum marks obtained can only be 120.

Instructions: This assignment must be done by individually. Each student must understand the code and techniques used and be able to answer each question. It is fine to use web-resources and help, provided you reference them and understand the code. **Please don't copy code from the internet resources blindly.**

Due Date: Sep 5, 2022 (End of day): **Hard Deadline**

Late Submissions: There are **ZERO late days** allowed for this assignment. We won't consider late submissions. Submit whatever you have completed by the deadline.
4 out of 7 of your best assignments will be considered in the final grade. That can be used to handle emergencies, sickness etc., so in general no exceptions will be made.

Submission: A Jupyter notebook.

Good solutions will be posted on the Course Git-hub.
Write all your references used in completing the assignment at the beginning of your Notebook in the space provided.
If you change the values in the cell manually to generate outputs and try different combinations, we encourage you to show that work in your notebook with appropriate comments. However, irrelevant details should be put at the end as supplementary work.

Note: Colab or link files will not be accepted. Your submission must be one single jupyter notebook (.ipynb) file and the name of the file must be, FirstName_RollNumber.ipynb (Like, for Anirban Haldar, 12110240, the file name should be, Anirban_12110240.ipynb)

Part A: Numpy [40 marks]

1. [5 marks] Create a numpy array A with first 10 multiples of 5, i.e., [5,10,15…50]
2. [5 marks] Compute the average and standard deviation of A.
3. [5 marks] Using the numpy.random.normal()function create a 10x10 matrix B, where each entry is sampled from a normal distribution of mean=0 and sigma=1. Count how many entries of the matrix are positive? Repeat this a few times. Does the count of +ve entries keep changing? Why?
4. [5 marks] Print all the diagonal entries of B in previous question. What is their sum and what is their product?
5. [5 marks] Multiply A with B and report the answer.
6. [15 marks] Write your own function in Python to compute the dot product of two numpy arrays. Please also check if the operation is valid or not and print an error message if it is not defined.

Part B: Regression [60 marks]

Use pandas to load the diamonds dataset: https://github.com/gagan-iitb/CS550/blob/main/Prog_Assignments/diamonds.csv
Beginning with the notebook provided here (taken from Kaggle), answer the following questions. You may have to edit the code or write new code.
https://github.com/gagan-iitb/CS550/blob/main/Prog_Assignments/diamond-price-prediction.ipynb

a. [5] What is the use of describe () function in pandas? Interpret the output. Why are some attributes missing in the output?
b. [5] Create a pair-plot between the following attributes/columns: Carat, Price, X
c. [5] Explain how regression lines can be used to identify outliers? Write a justification for the valid range of Y and Z.
d. [5] Create a catplot in seaborn to visualize the price distribution per category for 'Cut', 'Color' and 'Clarity'.
e. [5] Use an encoder to encode the categorical attributes. Sklearn provides 3 encoders: Label, One-hot Encoder and Ordinal Encoder. Which of these are appropriate choice for each of the categorical attribute? Please justify.
f. [5] Please write how to interpret the correlation matrix and its heatmap?
g. [10] Divide the data into the form: Aw=y as taught in the class. What will be the dimension of A^TA? Solve the normal equations to obtain optimal parameters: $\boxed{\widehat{\boldsymbol{w}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}}$ Please write the equation and its interpretation.
h. [5] Calculate the MSE for this solution.
i. [15] Train the KNN regressor for various values of K and find the best value using cross-validation.

Part C: Classification [50 marks]

You started a Car Selling business and you are giving recommendations to people for buying cars. We have provided a dataset to help you start your business.
Dataset: https://github.com/sarang-iitb/Car_Condition_evaluation_dataset

Now, your job is to train a robust model and we will test how good you are on the test set.

- [20 marks] **Data Preparation:** Perform necessary transformations on the input dataset to prepare it for ML model training.
- [30 marks] **Decision Tree Model:** Train a decision tree and provide us your outputs for the test set.