Assignment 1

CSL7620: Machine Learning
AY 2022-23, Semester – I

Due on: August 28, 2022

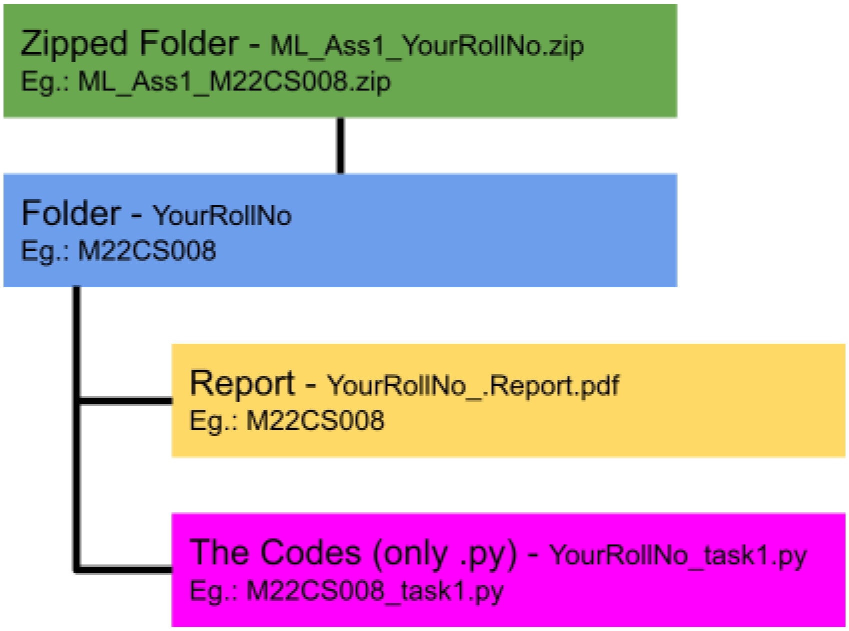
M.M: 100

General Instructions:

1. Clearly, mention the assumptions you have made, if any.
2. Clearly, report any resources you have used while attempting the assignment.
3. Any submission received in another format or after the deadline will not be evaluated.
4. Make sure to add references to the resources that you have used while attempting the assignment.
5. Plagiarism of any kind will not be tolerated and will result in zero marks.

Submission Guidelines:

1. Prepare separate Python code files for each task and name them as YourRollNo_task1.py, YourRollNo_task2.py, YourRollNo_task3.py, and YourRollNo_task4.py, respectively.
2. Also, provide your colab file link in the report. Make sure that the file is sharable.
3. Put both the codes and a report in a folder named <YourRollNo>, create a zip folder named ML_Ass1_YourRollNo.zip, and upload in google-classroom. See attached image to get better clarity.
4. Do not download the .ipynb file, rename it as .py, and upload it. .ipynb files are not exactly in a readable form, and hence uploading it will only result in you receiving 0 marks for the same. You have an option to download .py file in google colab. Use it to get the .py format.
5. Submit a single report depicting the method, results, and observations for all the tasks. There is no need to add theory behind the concepts. Preparing a report is mandatory; failing which will lead to non-evaluation of the assignment.
6. Do not copy paste code or screenshot, etc. in the report. Report should look like a technical document, containing plots, tables etc. whenever necessary.


Zipped Folder - ML_Ass1_YourRollNo.zip
Eg.: ML_Ass1_M22CS008.zip
Folder - YourRollNo
Eg.: M22CS008
Report - YourRollNo_.Report.pdf
Eg.: M22CS008
The Codes (only .py) - YourRollNo_task1.py
Eg.: M22CS008_task1.py

Task 1: Simple Linear Regression [10 marks]

This year Jodhpur district has seen an ample amount of rainfall. Arun is an agriculturist who has been assigned a job to predict the crop yield for the current year given the past data. Help him to find out the dependent variable, independent variable and the relationship coefficients between the two variables. Also, plot the data samples, the regression line and the coefficients of regression. Report MAE and MSE. It is expected that Jodhpur will receive about 560 mm avg. rainfall in 2022, what would be the predicted crop yield for the year? Get past data here.
Hint: Use your regression equation to estimate.
Instruction: Do not use any in-built library. Use Least Squares method.
Bonus: Make .gif for each iteration [2 marks]

Task 2: Multiple Linear Regression [20 marks]

Download the data from here. Build a regression model and evaluate the model performance using R-square. Handle missing data and outliers, appropriately, if any. Also, mention dependent and independent variables. Does your model overfits, give reason for your choice.
Instruction: Do not use any in-built library. Use Gradient Descent method for optimization.

Task 3: Polynomial Regression [20 marks]

Ramya has joined ABC company as a Data Analyst. Her manager has asked her to find a best model that fits the dataset provided here. In order to ensure that the model best fits the dataset, she prepares three models:
Model 1: Linear Regression
Model 2: Polynomial Regression with degree 2
Model 3: Polynomial Regression with degree 3
Help her to convince her manager that why one model describes the dataset better than the other models using metrics. Provide the assumptions, plots and other details as much as possible.
Instruction: You may use the in-built library of your choice for optimization.

Task 4: Half-space Classifier [20 marks]

A half-space classifier is a classifier that classifies a feature vector either as -1 or +1 depending upon the side of a hyperplane it lies. Implement a simple half-space classifier on this dataset using an LP solver. Train your the model on - 60:40, 70:30, 80:20 and 90:10 train-test splits. Evaluate your model using the following evaluation metrics - Accuracy, Precision, Recall and F1-score.
Do not use any in-built function for evaluation rather, design your own functions. You may use any real dataset for your experiment.
Hint: You may use scipy.optimize.linprog

Task 5: Logistic Regression [20 marks]

5.1 Plot a graph for sigmoid function.

5.2 Let ABC be the last three digits of your roll number. If

- ABC % 2 == 0, select experiment 1.
- Else, select experiment 2.

Experiment 1: Download dataset from here and classify them into window glass vs. non-window glass. Use min-max normalization.

Experiment 2: Download dataset from here and build a linear classifier.

5.3 Build a logistic regression model with 60:40 train-test dataset split. Use Gradient Descent optimization method.

5.4 Plot confusion matrix and AUC-ROC curve (You may use the in-built library).

Instruction: You have to write a code for your choice selection also.

References:

1. **https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc**
2. **https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html**

Task 6: Report [10 marks]