Machine Learning CS 550

Programming Assignment 1: Regression Analysis

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

Note: Up to 15 marks bonus for interesting features and achieving high accuracy.

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: Aug 21, 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 template containing instructions is being posted along with the assignment. Submit your solution in a Jupyter Notebook only, add cells at appropriate locations & write down your solution. Try to give proper justification of your answer if needed. We may ask you for a demo.

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 1: New York City Taxi Fare Prediction [80 marks]

Can you predict a rider's taxi fare?

Dataset: New York City Taxi Fare Prediction

(Kaggle : https://www.kaggle.com/competitions/new-york-city-taxi-fare-prediction/code)

Here, the primary goal is to predict a travel fair with the help of features provided like, (Pickup Location, Drop Off Location, Pickup time etc.). The training file contains 1 million instances to train from.

Hint: If your laptop/Colab doesn't handle the large dataset, we have also provided a subset of the data here. Feel free to use it for code development.

Before you start coding, think of the approach and which features, the taxi fare might be dependent on. If you need to create new features, add them from the beginning and redo each step.

Hint: Which distance measure would be appropriate? Does fare depend on time of the day and season?

A. Data Cleaning and Visualization: (15 marks)

In data cleaning, check for suspicious or null values in each feature of the dataset & handle them properly. Next, visualize various aspects of the dataset features, like distribution, pair-wise correlation etc.

- *Hint: You can use box-plots,* histograms, *Violin Plots, Pair-Plots etc. available in the Seaborn Library: https://seaborn.pydata.org/ or any other plotting library*

Don't forget to write the outcome or insight of the visualization (what you can conclude from the visualization). You can experiment with visualization of your own choice and write down why you use them and what you conclude from the visualization.

Note: *In case you create new features, make sure to clean and visualize them as well.*

B. Data Scaling: (5 marks)

Make appropriate scaling & standardization functions.

Note: In case you create new features, scale them appropriately. This is done to ensure that the matrix operations don't become unstable.

C. Building a Pipeline: (10 marks)

Now, build a scikit-learn pipeline, which will take the raw data as input, clean it, pre-process it, standardize-scale it and the output should be a numpy/pandas dataset ready to be feed into the Regression Model.

D. Use of Validation Set and Cross Validation Approach: (20 marks)

Apply k-fold cross validation set algorithm to evaluate your model and choose hyper-parameters. Writing an algorithm from scratch is mandatory. Using sklearn or any library functions will not fetch you any marks. You can use numpy.

E. Linear Regression: (30 marks)

   Train various regression models on the training data. You can use the built-in Regressor class from sklearn package, but if you want to write your own regression model from scratch, you are also welcome to do so (it is recommended to write algorithms from scratch as enough math and insight has been taught in the class). It is recommended to use the following types of regressors:

   - Matrix Based: Write the equation finally obtained
   - Optimization Based: Write the equation finally obtained
   - Non-parametric

   Which of the above are most suited for this problem? Why?

   Hint: Study the convergence time, accuracy etc.

Part 2: Life Expectancy (WHO) [70 marks]

Statistical Analysis on factors influencing Life Expectancy

Dataset

The data was collected from WHO and the United Nations website. The objective of this practice is to

A. Feature Selection: Do various predicting factors which have been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting life expectancy?
B. Should a country having a lower life expectancy value (<65) increase its healthcare expenditure in order to improve its average lifespan?
C. How do Infant and Adult mortality rates affect life expectancy?
D. Does Life Expectancy have positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.
E. What is the impact of schooling on the lifespan of humans?
F. Does Life Expectancy have a positive or negative relationship with drinking alcohol?
G. Do densely populated countries tend to have lower life expectancy?
H. What is the impact of Immunization coverage on life Expectancy?

(Kaggle:https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv)

The tasks in this assignment are,

1. Data Cleaning and Visualization: (10 marks)

   Similar to previous part, but additionally use maps to explain the data and the results. https://plotly.com/python/choropleth-maps/

2. Distribution analysis (15 marks)
   Perform Kolmogorov–Smirnov test to find which of numerical attributes are close to normal (significance 5%).

Now identify 3 metrics (columns) which don't seem to be normally distributed. Perform the KS test again (significance 5%) and convince yourself that it actually works! For these 3 metrics, can you identify which known distribution it closely matches?

3. Data Scaling: (5 marks)

4. Building a Pipeline: (5 marks)

5. Use of Validation Set and Cross Validation Approach: (5 marks)

6. Feature Selection: (30 marks).

    a. The lasso method regularizes model parameters by shrinking the regression coefficients, reducing some of them to zero. The feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model. Apply lasso on the dataset and remove unnecessary features.
    b. Use OLS (statsmodel.api) and investigate the p-values, performing forward/backward/mixed selection as explained in the class to arrive at a good model.
    c. Use the feature selection in Scikit learn library (implements a backward selection algorithm)

    https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html