CS550 2023-2024 W: Assignment 1

February 8, 2024

Due Date - February 25, 2024

General Guidelines

1. The training dataset for all the problems can be found here
2. The submission file must be of .zip format and should contain the relevant .ipynb files and an optional report.pdf. Although .ipynb is an acceptable report as well. Bonus marks will be given for good explanation of results, experimentation and visualizations.
3. Do not plagiarize as it is considered highly unethical in academia. Do cite any websites or articles that you may have used while solving these questions.
4. Even though the assignments are individually graded, we encourage group discussion and so you may mention a few collaborators, **although if too much similarity** is found, the **involved parties shall be penalized/debarred**. **There should not be more than 2 collaborators.**
5. Please feel free to reach out to me over email. You may also hit any of the TAs up for doubts.
6. If anyone does not have a very good laptop, please try google collab. It is a very simple to use and you may also import files from your drive. You may download the notebook as an.ipynb file and submit it on canvas.
7. We shall be sharing a test set in the final week before the deadline. Please submit the predictions.csv in the prescribed format (shall be conveyed later). If anyone does it earlier, you may inform us and we shall prepone the evaluations :)
8. Ranking shall be prepared based on the aforementioned test set evaluations. Top 3 students shall be given some bonus marks!

Body Temprature dataset : Linear Regression [35 marks]

To evaluate the benefit of implementing standardized deployment and acquisition practices in the measurement of elevated body temperature (EBT) with infrared thermographs (IRTs), a study was conducted with more than a thousand subjects. Subject oral temperatures were measured and facial thermal images captured with two evaluated IRTs. Based on the thermal images, temperatures from different locations on the face were extracted based on developed method and are listed in six CSV file as the open database. All data in these files has been de-identified. The 33 features consist of gender, age, ethnicity, ambiant temperature, humidity, distance, and other temperature readings from the thermal images. The dataset is intended to be used in a regression task to predict the oral temperature using the environment information as well as the thermal image readings. (There are 2 target columns)

Link to the the dataset - data

1. Perform Exploratory data analysis and study the features of the dataset. Perform the necessary data cleaning, feature extraction and give a brief note of your understanding of the features. Use various plotting libraries such as matplotlib, seaborn and plotly to draw effective visualizations of the dataset to make it easier for us to check assignments :)

2. Try to model this data using your own linear regressor. Do not use any inbuilt functions or else negative points shall be awarded. Evaluate the performance using various metrics such as R-squared error, RMSE etc. The more the merrier ! Also explain your results and compare it with pseudoinverse based regressor. You may your own various batch size and learning rate. Be prepared as we might randomly pick you up for a demo :)

3. Experiment with the training process and plot the loss with the number of iterations of the gradient-descent based linear regressor. Experiment with the injection of noise in the inputs and the parameters. What did you understand from this exercise ? Suggest a use-case based on your understanding.

Egyptian Hepatitis C Virus : Classification [30 marks]

Consider a dataset containing various demographic, clinical, and laboratory features of Egyptian patients who underwent treatment dosages for Hepatitis C Virus (HCV) over a period of 18 months.

The dataset includes features such as age, gender, BMI (Body Mass Index), symptoms like fever, nausea/vomiting, headache, diarrhea, fatigue & generalized bone ache, jaundice, epigastric pain, as well as laboratory measurements like white blood cell count (WBC), red blood cell count (RBC), hemoglobin (HGB), platelet count (Plat), and levels of liver enzymes (AST, ALT) at different time points. Additionally, RNA measurements at different time points (RNA Base, RNA 4, RNA 12, RNA EOT, RNA EF) are included, which are indicators of the response to treatment. The target variable for classification is the "Baseline histological staging", representing the severity or stage of liver disease at the beginning of the treatment.

Your task is to perform classification on the "Baseline histological staging" column using various classifiers. Feel free to use classifiers such as SVM, KNN, Decision Trees, Ensemble models etc. The more the merrier !

Link to the dataset - data

Mysterious Dataset [35 marks]

In the industry, we often have to deal with unstructured data. Sometimes we have the privilege of structure but the features may or may not be intuitive! In such a scenario, we need to study the distribution of the data and make predictions.

Enough words, now its time to jump to action. Given the dataset here, perform distribution analysis and classify the data into appropriate labels.