CSE4/510: Applied Deep Learning
Summer 2020

Instructor: Alina Vereshchaka

Project 1 -- Data Analysis

Due Date: July 14, Tue, 11:59pm

Description

Our first project is focused on preprocessing, analysing and visualizing real-world datasets. Applying the basic statistical methods and extracting valuable summary about it. During the third part of the project you are expected to employ multiple datasets and extract insights from them.

Dataset

Requirements to the datasets:

- Represent the real-world data
- Contain at least 50k entries

Possible resources includes:

- Open Data Buffalo - https://data.buffalony.gov/
- Google Dataset - https://datasetsearch.research.google.com/
- US Government's Data https://www.data.gov/

Tasks

Part I: Perform data analysis of the dataset [20 points]

1. How many entries and variables does the data set comprise?
2. What types of data is included?
3. Are there any data missing?
4. Provide the main statistics about the entries of the dataset (mean, std, etc.)

5. Visualize the data (min 3 graphs), e.g. correlation between different variables. Are there any interesting patterns?

Part II: Apply ML analysis [40 points]

1. Choose the features and targets in the dataset.
2. Preprocess the dataset for training (e.g. cleaning and filling the missing variables, split between training/testing/validation).
3. Apply ML algorithms (min 3 algorithms) to model the target variable. This can be either classification or regression task. You can use any of the libraries with inbuilt ML functions.
4. Provide the comparison of the results of different ML models you have used. This can be in the form of graph representation and your reasoning about the results.

Part III: Employ multiple datasets and extract insights [40 points]

1. Choose any related dataset to your current one. Combine the two into one dataset.
2. Choose the correlated variables.
3. Perform statistical analysis on finding the correlation between selected features from both datasets. Examples:
    a. Find the correlation between the crime and the number of schools in the area.
    b. Find the correlation between the traffic and the population in the area
4. Analyse the results and any interesting patterns.

Submit the Project

- Submit at **UBLearns > Assignments**
- The code of your implementations should be written in Python. You can submit multiple files, but they all need to have a clear name.
- All project files should be packed in a ZIP file named **YOUR_UBIT_project1.zip** (e.g. **avereshc_project1.zip**).
- Your Jupyter notebook should be saved with the results. If you are submitting python scripts, after extracting the ZIP file and executing command python main.py in the first level directory, all the generated results and plots you used in

your report should appear printed out in a clear manner.
- In your report include the answers to questions for each part. You can complete the report in a separate pdf file or in Jupyter notebook along with your code.
- Include all the references that have been used to complete the project.

Late Days Policy

Up to 5 free late days can be used throughout the course. They can be applied towards any project. No need to inform the instructor, late submission will be tracked at UBlearns.

Important Information

This project should be done individually. The standing policy of the Department is that all students involved in an academic integrity violation (e.g. plagiarism in any way, shape, or form) will receive an F grade for the course. Refer to the Academic Integrity website for more information.

Important Dates

July 14, Tuesday, 11:59pm - Project 1 is Due