Alexandria National University
Faculty of Computer and Data Science

Introduction to Machine Learning
Assigned: Thursday, March 7, 2024
Due: Monday, March 18,2024

Assignment 1

The objectives of this assignment are as follows:

1. Understand the concept of a classification task.
2. Understand the concept of a regression task.
3. Comprehend the process of training, validation, and testing data split.
4. Learn how to modify model parameters.
5. Gain proficiency in utilizing machine learning frameworks such as NumPy, Pandas, and Scikit-Learn.

Problem Statement 1 (Classification)

Given the MAGIC gamma telescope dataset. This dataset is generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The dataset consists of two classes: gammas (signal) and hadrons (background). There are 12332 gamma events and 6688 hadron events.

You are required to do the following:

1. the dataset is class imbalanced. To balance the dataset, randomly put aside the extra readings for the gamma “g” class to make both classes equal in size.
2. Split your dataset randomly so that the training set would form 70% of the validation set 15% and 15% for the testing set (Don’t use it while tunning the model parameters).
3. Apply K-NN Classifier to the data.
4. Apply different k values to get the best results.
5. Report all of your trained model accuracy, precision, recall and f-score as well as confusion matrix.
6. Add your comments on the results and compare between the models.

Problem Statement 2 (Regression)

Given California Houses prices data. This data contains information from the 1990 California census., it does provide an accessible introductory dataset the basics of regression models.

The data pertains to the houses found in each California district and some summary stats about them based on the 1990 census data. The columns are as follows; their names are self-explanatory:

- Median House Value: Median house value for households within a block (measured in US Dollars) [$]
- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
- Median Age: Median age of a house within a block; a lower number is a newer building [years]
- Total Rooms: Total number of rooms within a block
- Total Bedrooms: Total number of bedrooms within a block
- Population: Total number of people residing within a block
- Households: Total number of households, a group of people residing within a home unit, for a block

Dr. Marwan El Torki
Dr. Mohamed Khamis
Eng. Abd El Rahman Sakr
Eng. Youssef Nawar
Eng. Engy Ibrahim

- Latitude: A measure of how far north a house is; a higher value is farther north [°]
- Longitude: A measure of how far west a house is; a higher value is farther west [°]
- Distance to coast: Distance to the nearest coast point [m]
- Distance to Los Angeles: Distance to the center of Los Angeles [m]
- Distance to San Diego: Distance to the center of San Diego [m]
- Distance to San Jose: Distance to the center of San Jose [m]
- Distance to San Francisco: Distance to the center of San Francisco [m]

You are required to do the following:

Dataset:
https://drive.google.com/drive/folders/1VZM_IzvoZUKWrdUIhEadv4X9kflQW0gA?usp=sharing

1. Split your dataset randomly so that the training set would form 70% of the validation set 15% and 15% for the testing set (Don't use it while tunning the model parameters).
2. Apply linear, lasso and ridge regression to the data to predict the median house value.
3. Report Mean Square Error and Mean Absolute Errors for all your models.
4. Add your comments on the results and compare between the models.

Grading Scheme

1. Data Splitting 20%
2. Classification Problem 40%
    - Using the model correctly 10%
    - Trying Different K Values 10%
    - Report all the required performance measures and reasonable comments 10%
    - Student understanding of the algorithm 10%
3. Regression Problem 40%
    - Using the model correctly 10%
    - Trying all the variants of linear regression 10%
    - Report all the required performance measures and reasonable comments 10%
    - Student understanding of the used algorithm 10%

Final Notes

1. You should work in groups of three.
2. You should deliver a python notebook attached with the comments in markup cells, you should export it as pdf.
3. We will need both the pdf and the notebook in zipped file.
4. You should deliver with a naming scheme id _assigment.zip.
5. Delivery will be ignored if you didn't follow the naming scheme provided in 4, any one of the team ids can be used.

Dr. Marwan El Torki
Dr. Mohamed Khamis
Eng. Abd El Rahman Sakr
Eng. Youssef Nawar
Eng. Engy Ibrahim