Machine Learning Project

Introduction

The project involves developing a prediction model using Machine Learning techniques. Students will have access to as database and the main objective of the project is to build a performant model capable of predicting a target variable based on the features provided in the data.

Objectives

1. Analyze and understand the provided data.
2. Preprocess the data if necessary (cleaning, transformation, etc.).
3. Explore different Machine Learning techniques to build a predictive model.
4. Evaluate and compare the performances of different models.
5. Develop a Python script to load the trained model and make predictions on new data.

Project Description

The files containing the data are available on Moodle: one for the features of the examples (coloured images) and one for the target variable. The data can be donwloaded with the following python script:

```
import numpy as np
import gzip

# Load the images from the compressed CSV file
with gzip.open('x_train.csv.gz', 'rb') as f:
    x_train = np.loadtxt(f, delimiter=',')
x_train = x_train.reshape(-1, 32, 32, 3)

# Load the labels from the compressed CSV file
with gzip.open('y_train.csv.gz', 'rb') as f:
    y_train = np.loadtxt(f, delimiter=',', dtype=int)
```

Students are encouraged to explore and experiment with different Machine Learning techniques, such as supervised methods (regression, classification), unsupervised methods (clustering), or deep learning methods (neural networks).

After training and evaluating multiple models, students will select the most performant final model. They will then develop a Python script that takes a data file in the same format as input and save the model's predictions on that data. The behavior of the script should be similar to this one:

```
import sys
import numpy as np
import gzip

if len(sys.argv) != 3:
    print("Usage: python script.py <image_filename.csv.gz>\
     <label_filename.csv.gz>")
    sys.exit(1)

# Get filenames from command-line arguments
image_file = sys.argv[1]
label_file = sys.argv[2]

# Load the images from the compressed CSV file
with gzip.open(image_file, 'rb') as f:
    x_train = np.loadtxt(f, delimiter=',')
x_train = x_train.reshape(-1, 32, 32, 3)

# TODO: replace with proper processing
y_train = -np.ones(x_train.shape[0])

# Save y_train
with gzip.open(label_file, 'wb') as f:
    np.savetxt(f, y_train, delimiter=',', fmt='%s')
```

Deliverables

> The project can be carried out in pairs.

1. On Moodle: a FileSender link to dowload the mendatory python script (untitled `predict.py`) and any file required to make it work properly. **Deadline: March 30 at 23:59**
2. During **May 2** session: a 5 minutes oral presentation on the nodel, on the selection procedure, and if relevant on the data analysis. The presentation will be followed by 5 minutes of questions.

Evaluation Criteria

1. Understanding of the data and appropriate feature selection.
2. Correct application of data preprocessing techniques if necessary.
3. Thorough exploration of Machine Learning techniques and justification for the choice of the final model.

4. Quality and performance of the final model.
5. Robustness and efficiency of the Python prediction script.