MSBD5005x (Big Data Technology Capstone Project)
Data Mining Project: COVID-19 Infection

Pre-requisite: Note that this project requires learners to know how to write some simple Python programs and to execute Python programs in a command line.

1. Introduction

You are given two real datasets which contain the information about people who might have a COVID-19 disease. In the first real dataset, each person is associated with 12 attributes and 1 additional Boolean attribute called “infected” indicating whether the person has been infected with COVID-19. The second real dataset is the same as the first dataset but the second one contains only the first 12 attributes but no attribute “infected”. The objective of this project is to predict whether each person in the second dataset has been infected with COVID-19.

In this project, you are required to design 5 models (e.g., neural network) where each model could be constructed based on the first dataset and could be used to do the prediction on the second dataset. Finally, you are required to submit a final report, your program (written in Python) and 5 output files (where each output file corresponds to one output of the 5 models).

2. Milestones

1. **Milestone 1 (Python Program Learning for Classification)**
   i. If you know how to write a Python program for the given task, you could skip this step.
   ii. Otherwise, you are required to do the following.
      1. Learn how to execute a Python program in a command line (which is not in the environment like "jupyter notebook").
         That is, you could type "python program.py" in the command line (e.g., "DOS prompt" in Windows and "Terminal" in Mac) for execution where "program.py" is the Python program name.
         Note that this is one of the standard ways of executing a Python program. You could find a lot of resources in the internet. You could consider finding resources related to some Python environment management system such as Conda, Minicoda and Anaconda.
      2. Download the following 3 PPT files
         - 1-NeuralNetwork-Concept.ppt
         - 2-ClassificationConcept.ppt
         - 3-NeuralNetwork-Keras.ppt
      3. Self-learn these 3 PPT files.
      4. Download the following 2 sampled programs and 2 sampled datasets.
         - Sampled Programs
            i. program-NeuralNetwork.py
            ii. program-Enhanced-NeuralNetwork.py
         - Sampled Datasets
            i. Training-Dataset1-NoOfDim-8-Target-Binary.csv
            ii. New-Dataset1-NoOfDim-8-Target-None.csv
      5. Execute each of the above 2 sampled programs in the command line. That is, type
         python program-NeuralNetwork.py
         in the command line, and type
         python program-Enhanced-NeuralNetwork.py
         in the command line.
      6. Self-learn the 2 sampled programs with the guideline of the 3 PPT

files.

2. **Milestone 2 (Python Program Writing for Classification)**
   i. If you know how to write a Python program for the given task, you could design 5 different models conceptually and write a single Python program including 5 different models for the given objective, which could generate the 5 output files where each output file corresponds to the output of one model.
   ii. Otherwise, you are required to do the following.
      1. Based on the 3 PPT files, you could design 5 different models conceptually.

         Note that a neural network with one set of parameter values is regarded as a model, and another neural network with another set of parameter values is regarded as another model. Thus, you could have 5 neural networks each with a different set of parameter values in this project since they all are regarded as 5 different models.
      2. Based on the 2 sampled programs, you could write a single Python program which includes all 5 models and generates the 5 output files where each output file corresponds to the output of one model.
   iii. After you finish writing a single Python program, when you execute your Python program, your Python program will generate 5 output files called "predicted1.txt", "predicted2.txt", "predicted3.txt", "predicted4.txt" and "predicted5.txt", where each output file corresponds to the output of a model.
3. **Milestone 3 (Final Report Writing)**
   i. After you finish writing your Python program, you are required to write a final report.

3. Data Specifications

There are 12 attributes in the first dataset and 1 additional Boolean attribute called "Infected".

No.	Attribute Name	Attribute Content
1	Fever	"1", "0" (denoting whether the person has a fever where "1" means that the person has a fever and "0" means that the person does not have a fever)
2	Tiredness	"1", "0" (denoting whether the person is tired)
3	Dry-Cough	"1", "0" (denoting whether the person has a dry cough)
4	Difficulty-in-Breathing	"1", "0" (denoting whether the person has difficulty in breathing)
5	Sore-Throat	"1", "0" (denoting whether the person has a sore throat)
6	Pains	"1", "0" (denoting whether the person has pains)
7	Nasal-Congestion	"1", "0" (denoting whether the person has a nasal congestion)
8	Runny-Nose	"1", "0" (denoting whether the person has a running nose)
9	Diarrhea	"1", "0" (denoting whether the person has a diarrhea)
10	Age	"0-9", "10-19", "20-24", "25-59", "60+"
11	Gender	"Male", "Female", "Trans_Gender"
12	Country	"France", "Spain", "Iran", "China", "Italy", "Republic-of-Korean", "Germany", "UAE", "Other-EUR", "Other"
13	Infected	"1", "0" (denoting whether the person has been infected with COVID-19)

4. File format of Predicted Attribute File

You are required to submit 5 predicted attribute files for the second dataset. The files should be named as “predicted1.txt”, “predicted2.txt”, “predicted3.txt”, “predicted4.txt” and “predicted5.txt” where “predicted1.txt” corresponds to the output of the first model and the other files have a similar meaning. The file format of each file is shown as follows.

<1st row: 1 or 0 where 1 corresponds to that the first person in the second dataset has been infected with COVID-19 and 0 corresponds to that s/he does not>
<2nd row: 1 or 0 where 1 corresponds to that the second person in the second dataset has been infected with COVID-19 and 0 corresponds to that s/he does not >
…

Here is a sample file.

1
0
1
0
0
0
0
0
…

We have an answer file for the predicted attribute file. Among 5 files given by you, we will select the one with the highest accuracy as the final file for marking.

5. Format of Final Report

You are required to submit a final report to describe 5 models you used in this project. In this report, you should include the following.

1. **Personal Information**
   Please include your personal information (e.g., name and email address)
2. **Objective**
   Please include the objective of this project
3. **Datasets**
   Please include the list of datasets you use (due to your possible transformation of our given datasets)
   Please describe the steps of how you do the transformations.
   Note that these transformations must be included in your single Python program.
4. **Models**
   Please describe the list of 5 models you have implemented. In each model,
   a. please state the model type you are using (e.g., neural network)
   b. please state the parameter values you are using
   c. please state what dataset you used that you described in the previous section
5. **Results**
   Please include the following.
   a. For each of the 5 models, please describe the results briefly
   b. Please write a summary for all the results
6. **Conclusion**
   Please include a summary for the report
7. **References**
   If there is any reference, please include it

6. Turn-in

A *single* zipped file (in ".zip" file format) containing the following should be submitted.

1. A Soft Copy of Final Report (in PDF format) called "report.pdf"
2. A Soft Copy of your Python program called "main.py" which could do the following.
   a. It could read the first given dataset called "firstData.txt" and the second given dataset called "secondData.txt". We will use these 2 names and thus, your program could directly read these 2 files.
   b. It could generate the 5 output files, namely "predicted1.txt", "predicted2.txt", "predicted3.txt", "predicted4.txt" and "predicted5.txt", where each output file follows the format specified in Section 4.

   Note that we will execute your program with the command line as "python main.py". Please make sure that your program could be executed with "python main.py" in the command line (which is not in the environment like "jupyter notebook").
3. A Soft Copy of 5 Predicted Attribute Files for the Second Real Dataset (in TEXT file format)
   (Please name your files as "predicted1.txt", "predicted2.txt", "predicted3.txt", "predicted4.txt" and "predicted5.txt" as described before.)
   Note that these 5 files should also need to be included explicitly in the zipped file even though your Python program could generate these 5 output files.

In summary, the file/directory structure of your zipped file is shown as follows.

- report.pdf
- main.py
- predicted1.txt
- predicted2.txt
- predicted3.txt
- predicted4.txt
- predicted5.txt