CDS503: Machine Learning

Assignment 1: Classification

Problem

Depression is a common but serious mood disorder affecting an estimated of 121 million people worldwide. Emotion can possibly be an indicator of depression. The main objective of this assignment is to recommend the best classification model that can determine if an individual is showing signs of depression given the individual's emotion pattern in the duration of two weeks. We have worked together to collect a data set containing self-report of the emotions you experienced in a period of two weeks. Each instance includes an individual's gender and emotion pattern (based on 8 emotion categories: joy, sadness, anger, disgust, fear, surprise, contempt and neutral) and also a class label: YES (showing signs of depression) and NO (not showing signs of depression).

1) Study the data set carefully and answer the questions below:
   a. Report the class distribution. Is this a balanced or unbalanced data set?
   b. Please select and justify a suitable metric to evaluate the performance of your classification model.
   c. Given the size of the data set, which validation option (e.g., percentage split, k-fold cross validation) do you think is suitable to be used in your machine learning experiments. Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

2) Features extracted from the emotion data are represented in two forms:

   a. **Freq-PHO-Binary**: Total count of each emotion expressed by an individual in a period of 2 weeks. The total count of emotions differs for each individual depending on the number of emotions recorded in a day. An individual who recorded more than one emotion a day would produce a higher frequency number compared to an individual who only recorded one emotion a day.

   b. **Norm-PHO-Binary**: Emotion counts are normalized so that multiple emotions experienced in a day sums up to 1. Each individual recorded a different number of emotions a day. If the individual expressed two emotions (i.e., joy and sadness) on the same day but at different times of the day, the attribute Emotion_Joy is assigned a value of 0.5 and the attribute Emotion_Sadness is assigned the value of 0.5. Regardless of the number of emotions recorded in a day, the sum of all emotions is standardized to 15 (from 18 April 2022 – 2 May 2022).

   First, create a random baseline classifier (sklearn DummyClassifier, set parameter strategy = "uniform"). This baseline classifier will be used as a point of comparison when evaluating all other machine learning algorithms. For more information, check out:

   https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

Then, decide at least 3 classification algorithms you are going to run on the Freq-PHO-Binary and Norm-PHO-Binary respectively. You can tweak the parameters of the selected machine learning algorithms to obtain the best performing model. Report the machine learning experiments you ran and identify the best performing model using Freq-PHO-Binary and also the best performing model using Norm-PHO-Binary (both the best performing models should yield higher performance than your random baseline). Record all the results of your experiments in A1_Experiment_Sheet.xlsx and highlight the row indicating the best performing model.

Which feature representation produces a better model? Explain how you determine the best performing model based on the performance metric you have selected. Can you explain why one feature representation is better than the other?

3) Save and submit your best performing model. Use the pickle library and the following code excerpt to save your best performing models.

```
# Import pickle
import pickle

# Specify the file name to save the model
# Use filename='freq_model.sav' for Freq-PHO-Binary
# Use filename='norm_model.sav' for Norm-PHO-Binary
filename='freq_model.sav'

# Open the file name in write mode. Pass the filename and model.
# Replace modelname with the name of your model
pickle.dump(modelname, open(filename, 'wb'))
```

I will provide a separate test set after the deadline. We will then evaluate the performance of this model on the test set to see how generalizable your best performing model is.

4) Is your overall best performing model able to achieve very promising results (reach at least 0.8 of your selected performance metric)? Provide two suggestions on future strategies that can be used to improve the performance of your best performing model.

Submission Instructions

Assignment 1 is an **individual** assignment. Submit the following items in the assignment dropbox at elearn@usm latest by **11.59 p.m.** on **Sunday**, **5 June 2022**. Please make sure you include your full name and matric number on your report.

a) **Assignment report (maximum 5-page Word document file answering Questions 1, 2 and 4)**
b) **Experiment result spreadsheet (A1_Experiment_Sheet.xlsx)**
c) **Best performing model (.sav file)**
d) **Jupyter notebook containing your python scripts (.ipynb file)**

Grading Guidelines

Q1 (10 marks): Clearly address all the questions and provide rational justifications in setting up the machine learning experiments.

Q2: (20 marks): Experiments are set up correctly. How comprehensive are the machine learning experiments. Describe at least 6 classification experiments (3 using Freq-PHO-Binary and 3 using Norm-PHO-Binary). The best models are identified (performance results should be better than baseline classifier). Discuss how the classification results are judged and compared.

Q3: (10 marks): Based on how many instances in the test set are classified correctly by your best performing model.

Q4 (10 marks): Provide relevant and convincing suggestions.