Spring 2022 EE 658/758 Machine Learning in Engineering Dr. Jololian

Assignment #3: Neural Network

Due Date: Monday, March 14th, 2022

The dataset used for this assignment is known as the healthcare-dataset-stroke-data. It consists of 5,110 records of individuals with 12 attributes about each one. Some individuals in the data experienced a stroke, but most individuals in the data did not. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

We want to use this dataset to be able to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient. The first 11 attributes will be used as the features and the 12th attribute (stroke) is the label associated with the individual.

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if not _status

Please answer the following questions:

1) Data Exploration:
   a. For each categorical feature in the data:
      i. find all the unique values.
      ii. find the number of times each value is used.
      iii. plot a bar chart showing the frequency of each value.

   b. For each continuous numerical value, show a histogram of the values.

   c. For each feature in the data, find out the number of missing values.

2) Data Preparation

   a. For each feature that has missing values or unusable values, implement a strategy to handle these values.

   b. Split the data into training and testing subsets.

   c. Normalize the data

3) Model Training

   a. Use the Perceptron class in the Scikit library to create a model that can predict if an individual is likely to have a stroke. Show the confusion matrix and the classification results when you use the trained network on the test dataset.

   b. Use the two-layer neural network developed in class (included in the class notes) to build a predictive neural network model. Experiment with different number of neurons in the hidden layer and show the best obtained results. Show the confusion matrix and the classification results when you use the trained network on the test dataset.

   c. Use Logistic Regression from the Scikit library to create a model. Show the confusion matrix and the classification results when you use the trained network on the test dataset.

   d. Create new data and use it to make predictions with each of the three models above.

Note:

It is a common practice to split the data into training and testing subsets. However, this can produce misleading results when there is a class imbalance in the dataset. Instead, we need to ensure that we have the same ration of both classes in the training and testing data subsets. This can be achieved by using the **stratify=y** option in the train_test_split function, as shown below:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=47, ***stratify=y****)*

Useful commands:

- Drop a column:

  df.drop(*col_name*, axis=1, inplace=True)
  (where 'df' is a DataFrame)

- Replace categorical variables:

```
pd.get_dummies(df, columns=[col1, col2, …], drop_first=True)
```

- Exclude columns from a DataFrame:

```
df = df[~df[col_name].str.match('string')]
```