Take-home exam

CSCE 320

Vidhur Potluri

UIN: 626007235

Cookie

Data science question:

I propose to conduct a study that uses heart rate (pulse) and blood pressure data at a particular time to predict the emotion of a person at that point in time.

Over the last couple of decades, we have seen a significant increase in mental health awareness around the world. However, a majority of the people affected by mental health issues are not inclined to talk to friends, family, or professionals about those issues. I believe a solution to this is using technology to predict the mental health of an individual. Applications with access to this prediction could then make movies, music, books, and other suggestions that might improve the mental state of the individual.

We know that physical health is a major factor that affects mental health. Since most human beings have fluctuating emotions, and emotions are not constant, we can't diagnose the mental state of a person over a long period of time. But professionals would be able to identify a pattern. This, however, is out of the scope of this study. We will be focusing on the emotion at a particular point in time. Each person's body is affected differently by changes in emotions. The data used to train our prediction model will be collected from all 1 million users that use the health-o-meter and have given access to their heart rate and blood pressure data. This study will use blood pressure and heart rate data collecting during the minute that a user takes a picture. The picture could be obtained from the health-o-meter or from facebook.

The prediction of a user's emotion will be made under the assumption that they are a new user with no previous data about their blood pressure and heart rate levels being available. The user

for whom the prediction is being made does not necessarily have to use a health-o-meter. Access to their blood pressure (mm Hg) and heart rate (bpm) data will be enough for us to make our prediction.

The emotion detection module of the OpenCV and Keras toolkits would return the emotion a person is feeling by reading their face. The module classifies all emotions into the basic emotions - 'angry, 'sad', 'happy', 'neutral', 'surprise', 'disgust', 'fear'. The emotions detected would be used to validate our data. Predicting the emotion using regression won't be possible since our dependent variables are not continuous. We will be using an artificial neural network for the prediction analysis.

Preprocessing and exploration of the data: We will use heart rate, blood pressure, and emotional state data of all users from the time they start using the health-o-meter. Before, we begin with the test-train-split, we need to clean our data so that it includes only data relevant to our question. We begin by creating a dataframe by importing the dataset from a csv file (Data matrix) using the Pandas module in Python. This helps us store both continuous variables as well as categorical variables. Cleaning of data and adding columns to a dataframe would require the NumPy module. The following steps prepare the data

1) Identifying and handling missing values: data entries with empty blood pressure, heart rate, and emotions will need to be removed.
2) Deleting outliers: Data entries with malfunctioning health-o-meters need to be removed. These entries can be identified by humanly unachievable heart rates, blood pressure levels, etc.
3) Any letters in pulse or blood pressure data will need to be removed (Using regex).
4) The pulse and blood pressure levels from the minute that the picture was taken need to be averaged and stored as one entry only.
5) Create binary variables for each emotion ('angry, 'sad', 'happy', 'neutral', 'surprise', 'disgust', 'fear'). The values of these variables for each entry depend upon the emotion

detected on the user's face. If a user is happy, the value for that variable for that particular user is 1. It is 0 otherwise.

6) Get rid of the columns for variables that we are not using (name, cholesterol, etc). We will also delete the emotion column since we have the binary variables present.

Data mining and analytics approaches: We will be using predictive analytics for this dataset because we need to assign certain weights or scores to each of the variables that we have in the dataset to calculate the probability of a user having a particular emotion. The model used in this study is the classification model. The technique of predictive analytics that we will use is neural networks (machine learning).

We can't use decision trees since our independent variables are not binary variables or ranges. No clear decisions are made to determine the value of our independent variable. Linear regression cannot be used either since we are dealing with binary independent variables. The only type of regression that could be used is a logistic multiple regression. The artificial neural network uses a network of logistic regression to predict the values of binary variables. It will help us make a better prediction of the emotional state of a person. Since we may not be dealing with a linear relationship, it is best to use a neural network. The advantages of using a neural network:

1) The use of an activation function helps the network learn any complex relationship between the input (pulse, blood pressure) and the output (state of emotion) by introducing nonlinear properties to the network.
2) Works better with larger datasets like we have.

Disadvantages of using a neural network:

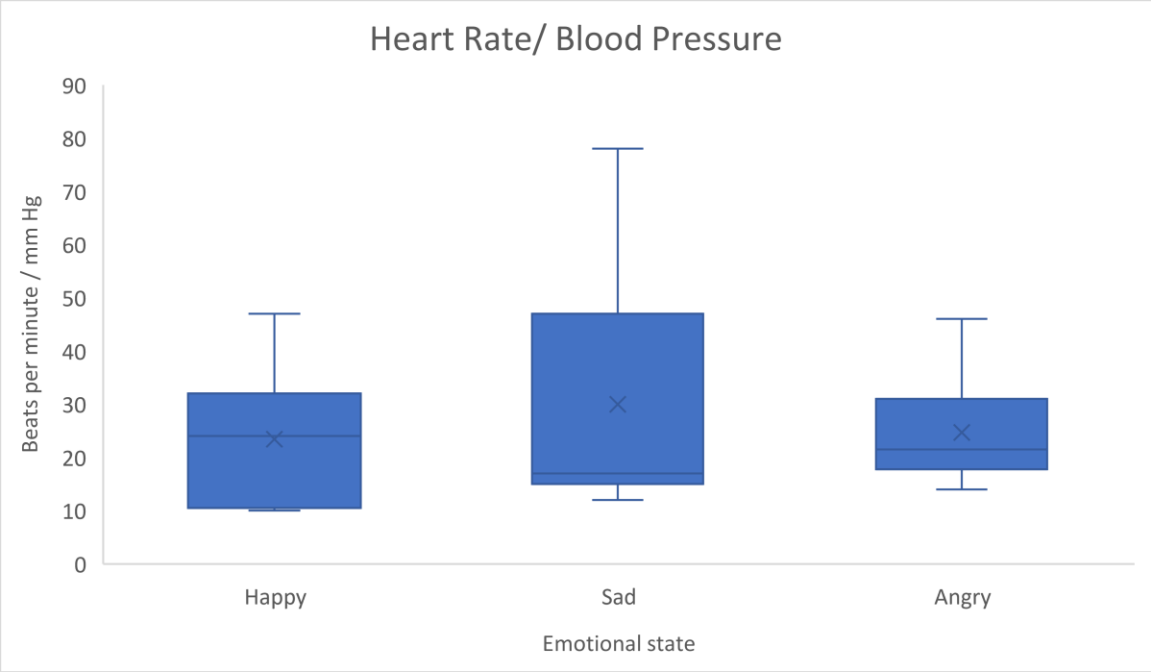
1) We cannot know much about how much each independent variable is influencing the dependent variable.

The train_test_split() function from the sklearn model will be used to generate X_train, X_test, Y_train, Y_test sets. The inputs for this function are:

1) X (independent variables): DataFrame containing the columns blood pressure and pulse.
2) Y (dependent variable): DataFrame containing the binary emotion variables.
3) Test_size: 0.15 (Since we have a large dataset, 15% of the dataset will be enough to validate)

The neural network is generated using the X_train and Y_Train as inputs and X_test and Y_test as validation data. The activation we will use is the sigmoid function since our dependent variable is binary. The output of the predict function of the model would then give us an array of the probabilities of a user experiencing each different emotion given the blood pressure and pulse. The emotion with the highest probability is then considered to be the current emotion of the user.

Data Visualization: Creating a viz from this particular neural network model isn't straight-forward. Visualizing the neural network doesn't give us a readable viz. It would consist mainly of layers, and network connections that might be confusing. We can't create a regression plot since we have two independent variables. So, visualizing our findings from the model isn't possible. Instead, we will create a viz that gives us an idea of how the data for blood pressure and pulse is distributed across different emotions that the users experience. We will be using the initially prepared and cleaned DataFrame for this viz. This DataFrame does not include the binary variables for emotions. We will represent this distribution of the preliminary data using 2 vizs – 1 for blood pressure and 1 for heart rate. Each viz will contain boxplots for each of the emotions. Although we don't know the scales or ranges for bpm or blood pressure, the viz's will look like this:



The pros of using such a visualization include being able to clearly see the distribution of the dependent variables for each emotion and the differences between these values for different emotions. A person looking at these vizs could make a general prediction about the emotion a person is experiencing based on heart rate and blood pressure. The two vizs don't clearly show how both blood pressure and heart rate combined are distributed across the two emotions.

Data validation: The data can be validated using the training and testing data generated through the test train split function. We will not be using cross validation since it is not appropriate and the epochs option in the neural network allows us to run multiple iterations over the model to validate our data. The neural network model allows us to validate the weights assigned to the dependent variables using X_test and Y_test sets as validation data. The value of epochs (iterations over the entire data) could be changed to ensure a more accurate prediction model. The model will be evaluated based on the accuracy. These two relationships help validate the data:

1) Measuring accuracy vs epochs: Accuracy of the validation data vs the accuracy of the training data. Usually an increase in the number of epochs leads to decrease in the gaps between the accuracy of the two datasets.
2) Measuring loss vs epochs: Validation loss vs training loss. A decreasing function when measured against epochs.

We could make graphs for these two measures. For each epoch, if the validation accuracy is more than the training accuracy, and the validation loss is less than the training loss, we know that the model is not overfitting. A test size of 15% will be enough to test such a large dataset. The model works perfectly if it predicts the correct emotion given a particular blood pressure and pulse of a person. The inclusion of many intricate modules like OpenCV for emotion detection and neural networks may cause some obstacles to the course. Not everyone expresses emotions the same way. This study would generalize emotions. In addition, blood pressure and pulse are affected by certain other factors such as cholesterol, etc, and could be different for different users. The pulse of a user may go up when they are exercising without a change in emotion. All of these factors may lead to inconsistencies. I would ultimately characterize this project as ambitious but still doable.

Ethical Considerations: Our study will analyze emotions from pictures of the users that have been provided by the users themselves. Certain users might not approve of this analysis. They may not want us and others to know about how they are feeling. This is the primary concern to consider before embarking on this study. The data from the study could be used to make shopping, movie, music and other suggestions. While some of these suggestions may improve a person's mental health, some could possibly worsen it. People and organizations with knowledge of patterns in mental health issues may have solutions for those affected by these issues. There could be major problems if these solutions are wrong. Access to people's photos may also be considered an invasion of privacy.

Along with the ethical considerations related to processing data on user's emotions, processing data on blood pressure and heart rate could have certain detrimental effects on society.

Pharmaceutical companies, medical and life insurance companies could pester users that they deem mentally and physically vulnerable. The best possible way to mitigate some of these concerns is to ask users if their data could be used in this study, and if they consent to the data being used by applications, companies, etc. to make suggestions. They could also choose to opt in to crowd sourcing instead of providing pictures – The users themselves would provide information about their emotions at a point in time which would be used to train our model. While there are plenty of concerns that may hinder us from pursuing an answer to this data science question, if it is conducted thoroughly and with the proper consent from users, the prediction model could improve mental health for millions.