Dane Zieman CS 412 HW5

Introduction: For this assignment, a data set consisting of a survey given to various young people. In total, there are 1010 respondents and 150 questions answered. For this assignment, the goal was to predict the output of one question, the one regarding their empathy level, using the other 149 responses. For this assignment, an individual is classified to be very empathetic if they answered 4 or 5, and they are classified as not very empathetic if they answered 1, 2 or 3. For the purpose of this assignment, this is a binary classification task where very empathetic is considered the positive class and not very empathetic is considered the negative class

Preprocessing and Model Selection: The first thing to be done was to preprocess the data. All examples with no empathy label were removed and all NaN values were replaced with the column's average value. This data contained some categorical features, these features were converted to numerical features. This was accomplished by giving each categorical feature a value from 1 to n where n is the number of categorical answer types for that question. This was chosen because the answers have a natural ordering. For example for the drinking question setting "never drink" as 1, "social drinker" as 2 and "drink a lot" as 3 is valid, as increasing drinking corresponds to increased value. Two models were chosen for this data: a random forest and a neural network. The random forest was chosen because it would be able to pick important features out of the data and disregard irrelevant ones well. The neural network was chosen mainly for personal interest in learning more about using neural networks, and to give a comparison to the random forest. The neural network wasn't necessarily ideal, as there is not a large amount of data (which neural networks need), however the learning experience and comparison was valuable. For software, sklearn's RandomForestClassifier was used for the random forest, and for the neural network TensorFlow's Keras Sequential model, these models were chosen because they are generally well-regarded implementations of these classifiers, and they interface very well with the format the data is presented in.

Experiment Design and Results: The data was split 75-15-10 train-validation-test, and the majority classifier was used as a baseline to compare against the models. The goal of this experiment was to create a model that outperformed the majority classifier on the test data by 10% absolute. The number/type of layers, as well as the number/depth of trees in the forest were decided by iteration. Additionally, class weights were added as the negative class is under-represented by 2:1, and without weights the models often label everything as the majority class. On the test data, the majority classifier had 63% accuracy, the neural network had 72% accuracy, and the random forest had 73% accuracy. Since the negative class is underrepresented, the most interesting points in the validation data are the negative examples. One negative example the random forest got correct was example #76, the decision tree weighs their answer to the "judgment calls" and "weight" heavily in classification, with lower judgment calls and higher weight leading to more likely negative class, #76 has both of these quantities. On the other hand, the model missed example #562 as a negative, because he answered a high number of the judgment calls question and had a low weight, and while their answers in other important questions were low, they weren't low enough to offset those two. Having more training data would be the best way to remedy this, withstanding that, perhaps changing some of the 1-5 features to 0-1 "high/low" features would help mitigate their large impact for small changes

References:

TensorFlow Keras: https://www.tensorflow.org/guide/keras
Sklearn RandomForestClassifier:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Link to extra credit Jupyter Notebook github

https://github.com/dane8373/EmapthyClassifier