EEP 596: AI and Health Care || Programming 2

Dr. Karthik Mohan

Univ. of Washington, Seattle

Apr 10, 2022

Submission guidelines

- Please submit a Jupyter/IPython notebook file, report and Kaggle predictions as part of your submission.
- You can start with the template notebook provided and add in your solutions to it.
- The report should be in a pdf format and have plots, correlation matrices and tables added in as mentioned in the assignment. Feel free to use either latex or word for creating it. Include answers to conceptual questions, and your insights as well. Ideally you should NOT use comments in ipynb to answer any conceptual question.
- There is a Kaggle competition as well for this assignment, submit your predictions on the "held out" test data set for a fun peer learning experience!
- You may discuss/brainstorm ideas to solve the assignment with peers - However, your submission should be your own and show your code implementation.

Problem overview

The main objective of this assignment is to use machine leanring techniques to predict stress. Psychological stress is defined as “a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being”.(Lazarus Folkman, 1984) There are 3 levels of stress - relaxed, neutral and stressed, they correspond with 0, 1 and 2 in the dataset.

The data comes from this paper. Wearable measurements of various young people driving in stressing environments, e.g. rush hour, highways, red lights, as well as a relaxation period are used. The objective is to use readings from a Microsoft Band 2 smart watch to predict the stress levels of the user.

Dataset description

RR interval refers to time interval between two heartbeats in milliseconds. Heart rate variability (HRV) is an umbrella term for a range of features derived from the time interval between two heartbeats. HRV can now be measured and calculated upto a good accuracy using advances fitness trackers. The raw data was in timeseries format, it has been converted to form a csv for ease of prediction. For every timestamp multiple features have been extracted and added to the excel sheet. There are also many time and frequency domain features, for a complete understanding of how the frequency domain features are calculated, you can refer to this thesis. The input features used are -

- AVNN - Average of all NN intervals
- SDNN - Standard deviation of all NN intervals
- pNN50 - Percentage of differences between adjacent NN intervals that are greater than 50 ms

Dataset description

- HR - Heartbeats per minute
- RMSSD - Square root of the mean of the squares of differences between adjacent NN intervals
- TP - Total spectral power of all NN intervals up to 0.04 Hz
- time - The starting time of corresponding window in which readings are taken.
- ULF - Ultra Low Frequency – Total spectral power of all NN intervals up to 0.003 Hz
- VLF - Very Low Frequency – Total spectral power of all NN intervals between 0.003 and 0.04 Hz
- LF - Low Frequency – Total spectral power of all NN intervals between 0.04 and 0.15 Hz
- HF - High Frequency – Total spectral power of all NN intervals between 0.15 and 0.4 Hz
- LF/HF - Ratio of low to high-frequency power

Stress Prediction Deliverables

- Do pre-processing(data cleaning and normalization) of the input features, and clearly describe the steps taken in the report. Is there a need to remove few features and why? Discuss change in f1score due any 1 additional pre-processing added. (20 points)
- Briefly describe how much data is missing in the dataset provided. How did you handle the missing data? Discuss at least 2 different methods used over here, and which one gives better increment in metrics and why? (15 points)
- For the classification, you can try logistic regression, random forests, kNN or any ML models that apply. Please share metrics (F1score, precision, recall and accuracy score) from at least 2 models in your report. (25 points)

Stress Prediction Deliverables

1. Pick any model used and answer the following interpretability questions -
   - What are the top 5 features that are predictive of the stress levels. What method was used to arrive at the top 5 features? (10 points)
   - Starting with the top 5 features identified previously, find the top pairs of features most relevant for the prediction. (10 points)
2. What is your best model and why does it stand out as compared to the other models you tried? (10 points)
3. Kaggle submission - Which model gets the best Kaggle accuracy for you? (5+5 points)

Bonus on Kaggle ranking!

Those in the top 10 on the Kaggle leaderboard shall get 5 bonus points.

Reference

1. Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions in Intelligent Transportation Systems 6(2):156-166 (June 2005).
2. Lazarus, R. S., Folkman, S. (1984). Stress, appraisal, and coping. New York: Springer.