My Data Science Challenging Problems # 51: How to identify predictive features from wearable device streaming data

Drafted by Haoda Fu

XXX

XXXX

xxx

November 9, 2018

Abstract

If you happen to have a solution of this question, please contact Haoda Fu at fu_haoda@lilly.com or fuhaoda@gmail.com or commit your solution to `https://github.com/fuhaoda/DSCP51_StreamingDataFeatureSelections.git` .

Key words and Phrases: Feature selection; Functional PCA.
Short title: ***

1 Introduction

Digital health attracts increasing attention these days. In a digital health study, patients wear various devices with different types of measurements for a period of time. We also have an outcome measure Y. One research question is what are the derived features from these sensor data to best predict the outcome of Y. We start from a basic form this research challenge, then extend to other situations.

2 Basic Form

Each patient i, this subject has a baseline measures denoted as $Z_i = \{Z_{i1}, \cdots, Z_{ip}\}$, at this time t, we also measure a vector of measurements from various wearable devices,

$$X_i(t) = \begin{bmatrix} X_{i1}(t) \\ X_{i2}(t) \\ \vdots \\ X_{iq}(t) \end{bmatrix},$$

The measurement frequency can be intensive such as 200Hz. For simplicity of our basic cases, we assume that we have 1 measurement per minute. Therefore, we have 1440 measurements per day. The study is about 3 months. So we approximately have 129,600 measurements for each subjects, denoted it as, $\mathbf{X}_i$ which is a $q \times 129,600$ matrix where q is about equal to 10, and use $\mathbf{X}$ as a generic version of $\mathbf{X}_i$. Different rows of this matrix can be correlated.

For each subject, we also have one outcome measurement, say $Y_i \in \mathbb{R}$ which is a binary or continuous measurement.

These studies often have about 50 to 200 subjects.

Our research question is how can we learn the relevant features from $\mathbf{X}_i$ so that it can be associated to Y_i.

There are two existing methods that we used frequently.

- The first approach relays on human generalized features. We summarize each device data into different features, such as mean, maximal values, ranges, area under the curves etc... then we use these features to fit model. This approach relays on experience and we may miss important features.

- The second approach is based on ideas of PCA regression. We first conduct PCA or functional PCA on patients sensor data. Similar ideas also include using auto encoder and auto decoder to learn key features first, then use the features to predict response. However, learning the feature is unsupervised learning. Those features are often good representation of $\mathbf{X}_i$, but they may not be the ideal features to predict Y.

We want to automatically search features in $\mathbf{X}$, but we also need to come out certain regularization to balance the variance and bias trade off. Otherwise, it is easily to generate features which will over fit the data. Along this line, a few ideas,

- the features should be in low frequency domain.

- the features should be in low order moments combinations (e.g. up to the second order moments), such as mean, standard deviation (SD), mean/SD, slop.

- the features should be highly representative to the original $\mathbf{X}$, for example if can be a balance to be representative and predictive, e.g. $\mathbb{E}_n[\mathbf{X} - g\{f(\mathbf{X})\}]^2 + \lambda\mathbb{E}_n[Y - m\{f(\mathbf{X})\}]^2$, where $f(\mathbf{X})$ compressed $\mathbf{X}$ into features.

We the data are large, deep learning approach could be useful here. We can either treat the $\mathbf{X}$ as image using CNN or we can use RNN, such as LSTM algorithms.

3 Further Extensions and Challenges

A few extensions as below,

1. We can extend Y into a vector as $Y(t)$. So that we past information of $X(t^-)$ can be used to predict $Y(t^+)$.
2. Different devices can have different sampling frequency.
3. Missing data issues.
4. Y can be time to events or recurrent events.

4 Framework

Appendix: Proof of Equation ()

.

References