EC 503 Project Summary

Qijun Liu (Mandy)
liuq@bu.edu

Ye Lin (Clara)
yelin@bu.edu

Qianqian Ma
maqq@bu.edu

Abstract

In this project, Cox Proportional Hazard Model (CoxPH), Logistic Regression Model (LR), and Multi-task Logistic Regression (MTLR) are proposed to predict the occurrence of a time-event.We aim to clarify the difference and relationship among these algorithms, with the validation of data analysis coming from at least two published datasets.

1. Introduction

Survival analysis/time-to-event models widely emerge in recent medical research as a way to model a patient's survival. Recently, survival analysis has gained a lot of traction in some other fields, as it's extremely useful to help predict whether a certain event will happen in given period of time. For example, it could help the companies to predict when a customer will buy a product. In this project, the team wants to incorporate some of the popular survival analysis methods, in order to tackle the employment prediction problem.

In this project, the team aim to implement and compare three methods; two of them are popular survival analysis: Cox Proportional Hazard Model and Multi-task Logistic Regression Model; One of them is traditional Logistic Regression model.

In section 2, we briefly review the literature that inspired the team to start this project. In section 3, we give details on the two new models (CoxPH and MRLR). Since we assume the readers have a basic understanding of Logistic Regression, we will skip the discussion of LR in this summary. In section 4, we outline our objective plan and fall-back plan for the project. In the last section, we discuss our division of labor throughout the project.

2. Literature Review

Cox Proportional Hazard Model (CoxPH) is a time-to-event model which has been widely applied in the biomedical research area to predict a patients survival[1, 2, 3, 4]. More recently, this traditional model can be found in many other fields, such as advertisement, economics and telecommunication, etc. In January 2018, a paper developed the CoxPH model and Multi-task Logistic Regression (MTLR) into the Neural Multi-Task Logistic regression model (NMTLR) [5], which inspires the team to explore aforementioned methods to establish unemployment prediction models.

3. Problem Formulation and Solution Approaches

Survival analysis is generally defined as a set of methods for analyzing data in order to estimate the time until a certain event of interest occurs. In the medical research field, the event can be death, the occurrence of a disease, etc. The response variable, also known as the survival time, is the time to event, which can be measured in days, weeks, etc. The survival time is recorded from the beginning of a time series until the occurrence of the event of interest or until the subject exits the analysis (when it reaches the end of the time series).

When we describe a survival time, we can use:

- Probability distribution function

$$F(t) = \int_{-\infty}^{t} f(s)ds \tag{1}$$

- Survival function

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2}$$

- Hazard function

$$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

To compare two hazard functions, we can use the Hazard Ratio (HR):

$$HR(t) = \frac{h_2(t)}{h_1(t)} \tag{4}$$

Especially, in Cox Proportional Hazard regression, it assumes that the hazard ratio does not very with time, i.e. $HR(t) = HR$.

3.1. Cox Proportional Hazard Regression

Cox Regression model is given by

$$\begin{aligned} & h_i(t|X_{1i}, \ldots, X_{pi}) \\ & = h_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = h_0(t)\eta(X_i) \end{aligned} \tag{5}$$

where $h_i(t|X_{1i}, \ldots, X_{pi})$ is the hazard function for the i th person at time t. $h_0(t)$ is the baseline hazard function. $\eta(X_i)$ is the risk function such that $\eta(X_i) = \exp(\sum_{j=1}^{p} x_j^j \beta_j)$ with β_js being the coefficients to determine. The CoxPH model allows us to predict the survival and hazard functions of a subject based on its feature vector, however, it exhibits the following limitations:

- It assumes that the hazard function is powered by a linear combination of features of a subject.
- It relies on the proportional hazard assumption, which specifies that the hazard function of two individuals has to be constant overtime.
- The computation is not efficient.
- The baseline function remains unspecific makes the model ill-suited for actual problems.

3.2. Multi-Task Logistic Regression

To overcome these limitations, Chun-Nam Yu et al.[6] introduced the Multi-task Logistic Regression model that can calculate the survival function without any of the aforementioned assumptions. It can be seen as a series of logistic regression models built on different time intervals, so that the problem reduces as to estimate the probability that the event of interest happened within each interval.

In this model, except dividing the time axis into J time intervals and building a logistic regression model on each interval a_j, the authors[6] also proposed new definitions for the density and survival functions.

3.3. Logistic Regression

Compared to CoxPH model, logistic regression aims to estimate the odds ratio, which is different from the hazard ratio that CoxPH model looks at. Additionally, using CoxPH model allows us not to worry about the distribution of residuals.

4. Proposed Work

The ideal objective of this project is to achieve all the methods mentioned in the Introduction Section. Furthermore, the team would like to explore and compare other regression models with the three methods mentioned above. Based on our previous study and research experiences, we have the confidence to complete the CoxPH model, Logistic Regression Model, and Multi-task Logistic Regression model.

4.1. Dataset

Currently, the team has collected datasets as follows:

- `https://docs.google.com/file/d/0BwogTI8d6EEiM2stZzdFNXdxM00/edit`
- `http://data.princeton.edu/wws509/datasets/#phd`
- `https://github.com/rfcooper/whas/blob/master/whas500.csv`
- `https://github.com/rfcooper/whas/blob/master/whas500.csv`

The first data set is to explore the survival of job hunters, i.e., recording the job hunting for different hunters within 28 time periods. If a job hunter gets a job in a certain time period, he/she will be labeled as ”1” (corresponds to Death in the biomedical area) in this time frame, and ”0” in the previous time frames. Then it will stop tracking. While if a job hunter cannot find a job in the entire time periods, he/she will be labeled as ”0” (corresponds to survive in the biomedical area) in the 1st-27th time periods, and ”1” in the 28th time period. Pre-possessed of the second dataset is required, so that the graduated year is recorded as the span of 14 years (14 variables) rather than one variable. The last two links are about Worcester Heart Attack Study and Veterans Administration Lung Cancer, respectively.

It’s worth noting that we focus on exploring the algorithm, thus it is possible that we would not try on all datasets.

4.2. Tools

Based on the collected data, the team is going to apply the following tools to solve problems: Matlab, R, Python, SPSS.

4.3. Performance Evaluation

To better evaluate the performances of different models, AUC (Area Under Curve) is selected as the main evaluation index. Mostly, AUC refers to AUROC (Area Under the Receiver Operating Characteristic Curve), of which interpretation can be as follows: a value of 0.5 denotes a random model, a value of 1.0 denotes a perfect model, while a value of 0.0 denotes a wrong model.

5. Division of Labor

- Mandy Liu: Multi-task Logistic Regression, data acquisition, testing, report
- Clara Lin: Cox Regression, data acquisition, testing, report
- Qianqian Ma: Logistic Regression, data acquisition, testing, report

References

[1] Harrell, F.E., Lee, K.L. and Mark, D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361-387,1996.

[2] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J. and Kattan. M.W. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128,2010.

[3] Heagerty, P.J., Lumley, T. and Pepe, M.S. Timedependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337-344,2000.

[4] Kamarudin, A.N., Cox, T. and Kolamunnage-Dona, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17(1):53,2017.

[5] Fotso, S. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework.2018.

[6] Chun-NamYu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors.*In Advances in Neural Information Processing Systems* , 1845-1853, 2011.