ORIE 4741 Project Proposal

Kathy JaYoung Byun, Raye Liu, Jody Zhu

October 4, 2020

1 Question

The question we are asking in this project is how to make valid population segmentation of a certain group of patients given the various information and individual features of this group. In other words, we are aiming to come up with valid clustering of the patients, and within each of the patient clusters, the homogeneousness should be maximized.

We will be doing segmentation and validating the approaches by trying them on subgroups of the patients with certain diseases. We will test the inside clusters by examining the similarity of patients in each cluster, for instance, calculating the variance of the length of ICU-stay time and the number of hospitalizations of patients in each cluster, whereas smaller variance indicates higher homogeneousness.

Methods we intend to use include: (1) Low rank model: k-means & PCA (2) Tree models: decision tree & random forest (3) Maximal Coding Rate Reduction (MCR^2).

2 Dataset

The dataset we will by using is the MIMIC-III, third edition of Medical Information Mart for Intensive Care. Mimic-III is a deidentified clinical database that consists of health data of 3,423 distinct critical care hospital admissions from 38,597 distinct adult patients at the Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. Information such as gender, race, diagnosis, ICD9 codes, prescriptions, ICU stays, procedures, vital signs, mortality, laboratory measurements, and unstructured textual data from various healthcare provider notes and analyses, are included.

We think the data set will allow us to (begin to) answer the question we raised above due to the following reasons: first of all, it is a medical health care dataset with high credibility, and it has been used in industrial research, quality improvement initiatives, and higher education coursework. It has a relatively large number of observations, and thus can be seen as somehow representative and indicative for a diverse population of patients in a health care system. It also includes multiple features in different fields that are potentially helpful in both supervised and unsupervised learning, thus giving us the chance to enhance our approaches of training.

Link to the data: https://mimic.physionet.org/

3 Problem Significance

Health data is very complicated, and valid segmentation of the patients can provide intelligible insight for redistributing limited resources and improving the overall healthcare system's performance. For instance, more consistent care can be delivered to patients with similar needs or conditions. Patients can also be connected to the appropriate service providers: for example, many people who go to the ICU for acute stomach pain should be allocated to consult their primary care physicians instead of consuming emergency care resources. A greater number of early diagnoses can be made for individuals without known hereditary predispositions as well. On the hospital management side, well-segmented clusters can be used to help anticipate the high cost and long stay patients.