Diabetes Prediction using Bayesian Logistic Regression

Final Project Proposal for Bayesian Method for Data Science (DATS 6450)
Project Group 11: Li Du, Xiaochi Li, Yupeng Yang
Data Science, George Washington University

1. The problem definition and motivation

Diabetes is a metabolic disorder disease which has the symptom of high blood sugar level for a prolonged period. If left untreated, diabetes can cause many complications that may even lead to death.[1] The Centers for Disease Control and Prevention(CDC) found that 30.3 million people have diabetes in U.S. which is 9.4% of the U.S. population and diabetes was the seventh leading cause of death in U.S. in 2015. [2]

One way to reduce the healthy risk of having diabetes is the early detection of potential diabetes patients and give necessary treatments in advance.

To better understand the weights of certain variables and diagnostically predict whether a patient has diabetes based on these diagnostic measurements, our team will build a Bayesian logistic regression model based on Pima Indians Diabetes Database originally provided by National Institute of Diabetes and Digestive and Kidney Disease.

2. The proposed method, language and package you will need for the implementation

The project will be mainly based on R and will use JAGS and other necessary Bayesian packages in R.

The method is to use Markov Chain Monte Carlo (MCMC) method to find the posterior distributions of the predictors(i.e. observed diagnostic data)' weights in the Bayesian logistic regression model as well as plausible interpretation of these outcomes. And we will also evaluate how well the model fits the data and compare the prediction of Bayesian logistic regression and traditional logistic regression.

3. The link to the data

https://www.kaggle.com/uciml/pima-indians-diabetes-database

4. The responsibility of each team member

To learn the full process of Bayesian analysis on real world data, all team members will join every part of the project. But each one will have a leading role in different parts of the project.

Yupeng Yang will clarify the mathematical model of the problem using Bayes' theorem.

Xiaochi Li will implement the model in R language.

Li Du will interpret the result of the model, compare it with traditional method and write the summary.

---

[1] Diabetes mellitus Wikipedia https://en.wikipedia.org/wiki/Diabetes_mellitus

[2] National Diabetes Statistics Report, 2017
https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf