Danmarks
Tekniske
Universitet

DTU

42186 Model Based Machine Learning

Student No.	Name	Contribution
s217084	Aryamaan Saha	Data Analysis and Visualization, Hierarchical Logistic Regression Models (by Age, Gender and multi-level), SVI
s216911	Ananthalakshmi Nandakumar	Data Visualization, Bayesian Logistic Regression
s213605	Ananthu Sangeetha Vinod	Data Collection, Gaussian Process Classifier, MCMC Inference

May 27, 2022

Heart Disease Prediction

1 Introduction

Heart disease is considered as the most common cause of death worldwide. The dataset we use consists of information about the medical conditions of 918 individuals. In this project, we explore different probablistic models, namely bayesian logistic regression, hierarchical models and Gaussian process classification (GPC) to predict the possibility of heart disease in an individual based on features like age, sex, resting BP, cholesterol etc.

2 Data Analysis

Prior to building our classification models, we performed some exploratory data analysis.The data did not contain any missing values, or any prediction class imbalance, and was fairly clean to begin with. We checked if a certain gender or age group is more susceptible to heart disease, which further provided a basis and incentive for hierarchical models. We performed data visualization with the help of pairplots between numerical clinical features (like cholestrol, BP), noticing that men tended to have higher values of these numerical features. We also plotted the correlation matrix to observe the correlations between the various numerical features. As a part of data preprocessing, we also one-hot encoded the categorical features in the dataset and scaled the numeric features before feeding it into the model.

3 Modeling

The predominant goal of the project is to explore several classification models for the accurate prediction of heart disease using clinical parameters. We started off with a simple bayesian binary classification model, and followed it up with hierarchical modeling and a Gaussian Process classifier. Our train-test data split is 75-25. The performance of the various classification models are compared using the accuracy metric.

3.1 Bayesian Logistic Regression Model

In a simple bayesian classification model, the weights($\boldsymbol{\beta}$) and biases(α) are identical for all observations. The joint distribution is given by: $p(y, \alpha, \beta, X) = p(\alpha) * p(\beta) * \prod_{n=1}^{N} p(y_n|\alpha, \beta, x_n)$ The accuracy obtained is 87.4%.

3.2 Hierarchical Logistic Regression Models

We noticed a difference in susceptibility towards heart disease between certain groups of the individuals(observations). Men, and as expected, older people were more vulnerable towards heart disease. Using this information, we divided our individuals into distinct groups based on age and gender and performed hierarchical modeling, assigning different weights and biases to different groups, but tied together by a shared prior. However, despite running several hierarchical models, each with a few unique variations, the test accuracy only hovered around 85-87%, whose performance was equivalent to the simple bayesian classification model. Here we implemented a hierarchical model by age, a hierarchical model by age and a multi-level hierarchical model by both age and gender.

3.3 Gaussian Process Classification(GPC)

Gaussian processes are bayesian non-parametric models that are fully specified by a mean function and a covariance function. GPs can be used for classification. The GPC is based on predicting the target values y with respect to the features by **Y** $\sim$ **Categorical(Softmax(F))** where F denotes features and Y is the target Variable. It uses a radial basis function kernel(RBF) as the kernel parameter & binary class as the likelihood parameter of Gaussian Process. We trained the model for 2000 steps and obtained a test accuracy of 92.17%, which was higher than those obtained from all the other models.

4 Inference

We have used two inference algorithms, Stochastic Variational Inference(SVI) and Markov Chain Monte Carlo algorithm(MCMC) in our project. Inference is applied obtain the posterior distributions of the latent variables, which can be utilized to make predictions. In the bayesian logistic regression model we used the AutoMultivariateNormal function as a guide, and in the hierarchical models we used AutoDiagonalNormal. Our loss is EBLO, and our optimization algorithm is ClippedAdam. For MCMC inference, we used the NUTS kernel with 1 chain and 100 warmup steps. The accuracies obtained from both the inference algorithms are similar. The MCMC takes a lot more time to complete than the SVI algorithm.

5 Conclusion

In this project, we performed data analysis and implemented various probabilistic classification models for the prediction of heart disease in an individual. We found that the Gaussian Process classifier achieved the highest accuracy and outperforms all the bayesian classification and hierarchical logistic regression models.

6 References

Heart Disease Prediction Dataset