Logistic Regression and Regularization

Applied Machine Learning for Educational Data Science

true

10/26/2021

Contents

[Updated: Thu, Oct 28, 2021 - 12:50:54]

Overview of the Logistic Regression

Logistic regression is a type of model that can be used to predict a binary outcome variable. Linear regression and logistic regression are indeed members of the same family of models called *generalized linear models*. While linear regression can also technically be used to predict a binary outcome, the bounded nature of a binary outcome, [0,1], makes the linear regression solution suboptimal. Logistic regression is a more appropriate model and takes the bounded nature of the binary outcome into account when making predictions.

The binary outcomes can be coded in a variety of ways in the data such as 0 vs 1, True vs False, Yes vs. No, Success vs. Failure. The rest of the notes, it is assumed that the category of interest to predict is represented by 1s in the data.

The notes in this section will first introduce a suboptimal solution to predict a binary outcome by fitting a linear probability model using linear regression and discuss the limitations of this approach. Then, the logistic regression model and its estimation will be demonstrated. Finally, different regularization approaches for the logistic regression will be discussed.

Throughout these notes, we will use the Recidivism dataset from the NIJ competition to discuss different aspects of logistic regression and demonstrations. This data and variables in this data were discussed in detail in Lecture 1a and Lecture 2a. The outcome of interest to predict in this dataset is whether or not an individual will be recidivated in the second year after initial release.

```
recidivism <- read.csv('https://raw.githubusercontent.com/uo-datasci-specialization/c4-ml-fall-2021/mai
                       header=TRUE)

# Outcome variable

table(recidivism$Recidivism_Arrest_Year2)
```

```
    0     1
13544  4567
```

Linear Probability Model

Linear probability model is just fitting a typical regression model to a binary outcome. When the outcome is binary, the predictions from a linear regression model can be considered as probability of outcome being equal to 1,

$$\hat{Y} = P(Y = 1).$$

Sup

Description of Logistic Regression Model

Model Estimation

`glm` function

Building a Prediction Model for Recidivism

Initial Data Preparation

Train/Test Split

Model Fitting with the `caret` package

Regularization in Logistic Regression

Ridge Penalty

Model Fitting with the `caret` package

Variable Importance

Lasso Penalty

Model Fitting with the `caret` package

Variable Importance

Elastic Net

Model Fitting with the `caret` package

Variable Importance

Using the Prediction Model for Future Observations