Predictive Analytics
Regression and Classification

Lecture 7 : Part 1

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2019

Introduction

- logistic regression model is used to model the probability of a binary class of event
- Example: pass/fail, win/lose, alive/dead or healthy/sick
- Suppose you are an analysts in a Bank. You want to help the management to build a model to predict whether a loan applicant will be bad creditior in future.

Motivating Example

- So you look into historical data $\mathcal{D}$ on n existing customers in Bank's book

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \cdots, (y_n, \mathbf{x}_n)\},$$

where

$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ covariates or predictor or features of i^{th} customer in the bank's book.

Objectives

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \cdots, (y_n, \mathbf{x}_n)\},$$

where

$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ covariates or predictor or features of i^{th} customer in the bank's book.

1. Which covariates has impact on y_i ? Statistical Inference

2. For a new loan applicant $\mathbf{x}^0 = \{x_1^0, x_2^0, \cdots, x_p^0\}$ – what is the $\mathbb{P}(y^0 = 1) =$? Prediction

Latent Variable

$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

Equivalently, we can write

$$y_i = \begin{cases} 1 & z_i \geq 0 \\ 0 & z_i < 0 \end{cases}$$

z_i is the unobserved latent score.

Probit Model

We can model z_i as

$$z_i = x_i^T \boldsymbol{\beta} + \epsilon_i$$

What we want to model:

$$\begin{aligned}
\mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \boldsymbol{\beta} + \epsilon_i \geq 0) \\
&= \mathbb{P}(\epsilon_i \geq -x_i^T \boldsymbol{\beta}) \\
&= \mathbb{P}(\epsilon_i < x_i^T \boldsymbol{\beta}) \text{ by symmetry of the distribution}
\end{aligned}$$

1. If assume $\epsilon \sim N(0, 1)$ then it is known as probit model or logistic regression with probit link

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \boldsymbol{\beta}) = \int_{-\infty}^{x_i^T \boldsymbol{\beta}} \phi(\epsilon_i) d\epsilon_i = \Phi(x_i^T \boldsymbol{\beta})$$

Logit Model

We can model z_i as

$$z_i = x_i^T \boldsymbol{\beta} + \epsilon_i$$

What we want to model:

$$\begin{aligned}
\mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \boldsymbol{\beta} + \epsilon_i \geq 0) \\
&= \mathbb{P}(\epsilon_i \geq -x_i^T \boldsymbol{\beta}) \\
&= \mathbb{P}(\epsilon_i < x_i^T \boldsymbol{\beta}) \ \text{ by symmetry of the distribution}
\end{aligned}$$

2 If assume $\epsilon \sim \mathit{Logistic}(0, 1)$ then it is known as Logit model or logistic regression with logit link

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \boldsymbol{\beta}) = \frac{\exp\{x_i^T \boldsymbol{\beta}\}}{1 + \exp\{x_i^T \boldsymbol{\beta}\}}$$

Logistic Regression

- Logistic Regression with logit-link

$$\log\left(\frac{p}{1-p}\right) = x^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + x_p\beta_p$$

- Logistic Regression with probit-link

$$\Phi^{-1}(p) = x^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + x_p\beta_p$$

- How to estimate $\boldsymbol{\beta}$?

In the next video...

- I will discuss the methodology to estimate β...