PS5841

Data Science in Finance & Insurance

Logistic Regression

Yubo Wang

Spring 2022

Data Generating Scheme

Independent RV

$$Y_i = Ber(\pi_i),$$

$$E(Y_i) = \pi_i$$

$$Var(Y_i) = \pi_i(1 - \pi_i)$$

Model Probabilities: Directly

- Under GLM

$$g[E(Y_i)] = g(\pi_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

- Model π as a function of features

$$\pi(\boldsymbol{x}) = g^{-1}(\boldsymbol{x}^T \boldsymbol{\beta})$$

such that

$$\pi(\boldsymbol{x}) \in [0,1]$$

$$\lim_{\boldsymbol{x}^T \boldsymbol{\beta} \to -\infty} \pi(\boldsymbol{x}) = 0$$

$$\lim_{\boldsymbol{x}^T \boldsymbol{\beta} \to \infty} \pi(\boldsymbol{x}) = 1$$

Model Probabilities: Tolerance Distribution

- Any continuous probability distribution defined over the real line

$$\pi = \int_{-\infty}^{t} f(s)ds$$

What if …

- we use this tolerance distribution

$$f(s) = \frac{\exp(\beta_0 + s)}{[1 + \exp(\beta_0 + s)]^2}$$

- we get

$$\pi = \int_{-\infty}^{t} \frac{\exp(\beta_0 + s)}{[1 + \exp(\beta_0 + s)]^2} ds$$

$$= -\frac{1}{1 + \exp(\beta_0 + s)}\bigg|_{-\infty}^{t} = \frac{\exp(\beta_0 + t)}{1 + \exp(\beta_0 + t)}$$

GLM with logit link

$$g[E(Y_i)] = g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + t = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

$$t = \beta_1 x_1 + \cdots + \beta_p x_p$$

- Useful results

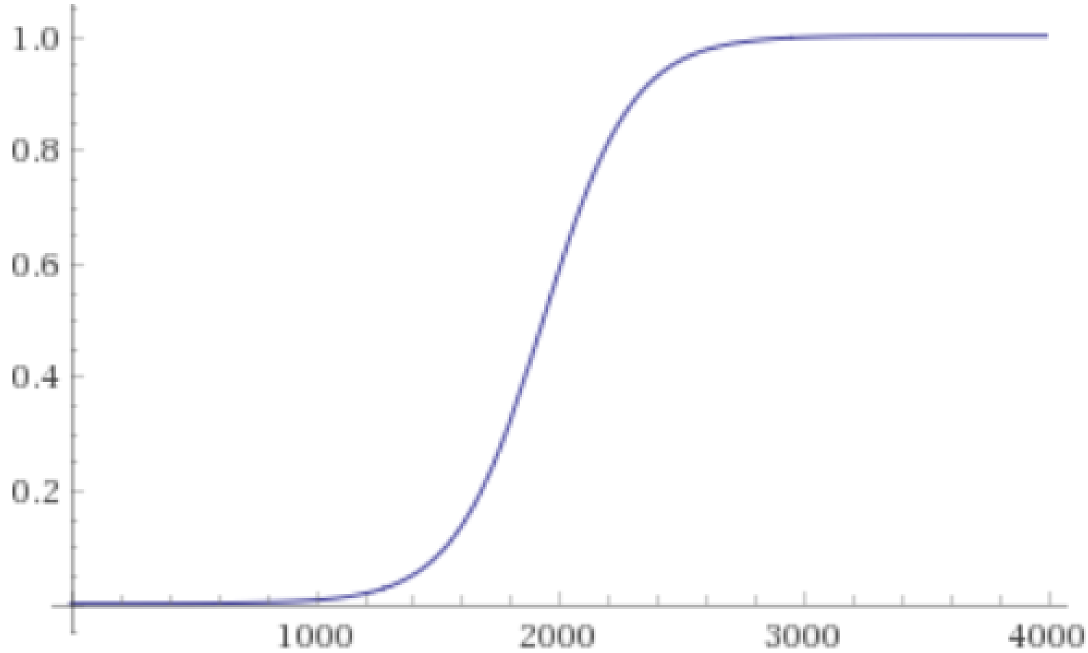
log-odds, logit
$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

$$\ln(1-\pi_i) = -\ln\left(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)$$

$$\pi_i = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp\,(\boldsymbol{x}_i^T\boldsymbol{\beta})}$$

Example: Estimated Probabilities

$$\pi(x) = \frac{\exp(-10.6513 + 0.0055x)}{1 + \exp(-10.6513 + 0.0055x)}$$



Maximum Likelihood Estimation

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left[\sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right]$$

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)$$
$$= \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)$$

$$\frac{\partial l_i}{\partial \beta_j} = (y_i - \pi_i)x_{ij}, \qquad x_{i0} = 1$$

Prediction & Classification

$$\hat{y}_i = E(Y_i) = \hat{\pi}_i = \frac{\exp\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)}{1 + \exp\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)}$$

- Classify $\boldsymbol{x}_i$ to class 1 if $\hat{\pi}_i > \pi^*$, otherwise to class 0

Linear Decision Boundary

- $\mathrm{logit}(\pi)$ is an increasing function in π,
  $\pi > \pi^* \Leftrightarrow \mathrm{logit}(\pi) = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} > \mathrm{logit}(\pi^*)$

- Decision boundary for responses with 2D features ?
  $$Y,\ X = (X_1, X_2)$$

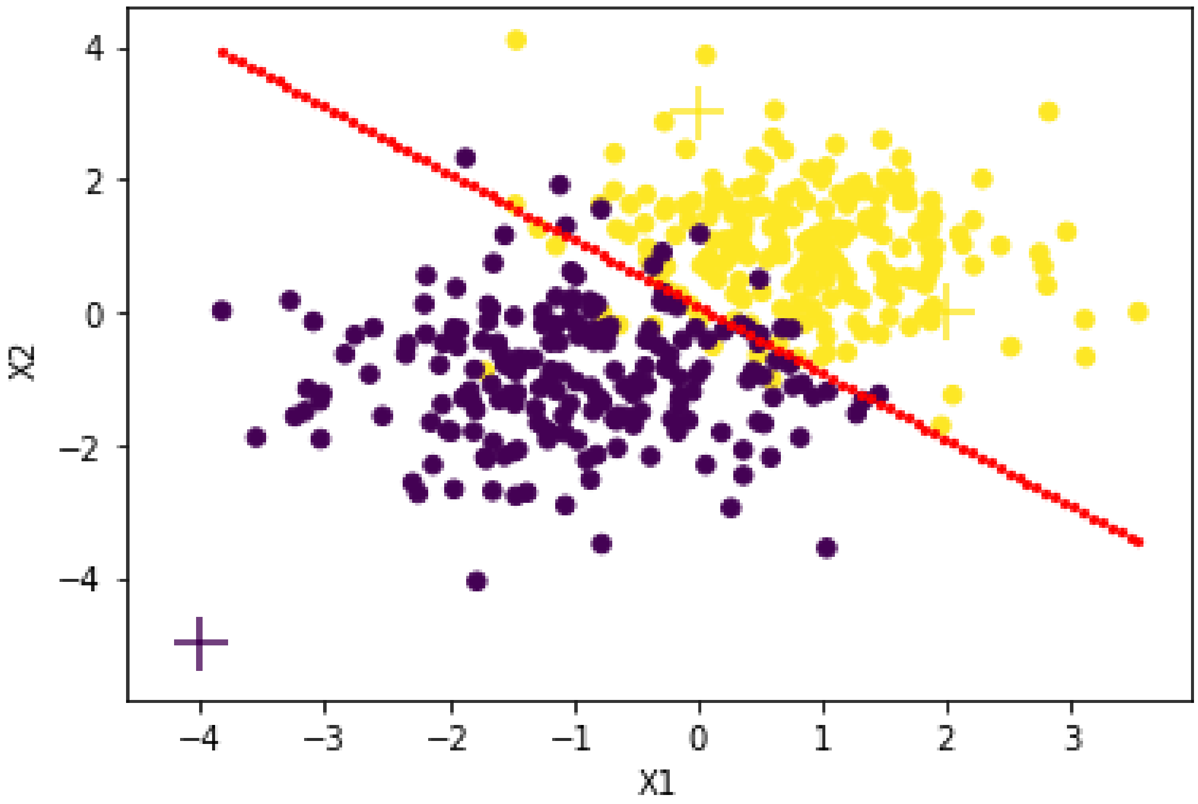
Example: 2D Features

- Let π^* be the classification threshold
  - e.g. $\pi^* = 0.5$

$$X_2 = \frac{\operatorname{logit}(\pi^*) - \beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} X_1$$

- Classification when $\beta_2 > 0$?

Decision Boundary



That was