

Classification

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

- 1 Motivation
- 2 Risk, Probability, and Classification
- 3 Bayes Classifier
- 4 Logit
 - MLE

Classification: Motivation

- ▶ Many predictive questions are about classification
 - ▶ Email should go to the spam folder or not
 - ▶ A household is below the poverty line
 - ▶ Accept someone to a graduate program or no
- ▶ Aim is to classify y based on X 's

Classification: Motivation

- ▶ Main difference is that y represents membership in a category: $y \in \{1, 2, \dots, n\}$
 - ▶ Qualitative (e.g., spam, personal, social)
 - ▶ Not necessarily ordered

*The prediction question is, given a new X ,
what is our best guess at the response category \hat{y}*

Agenda

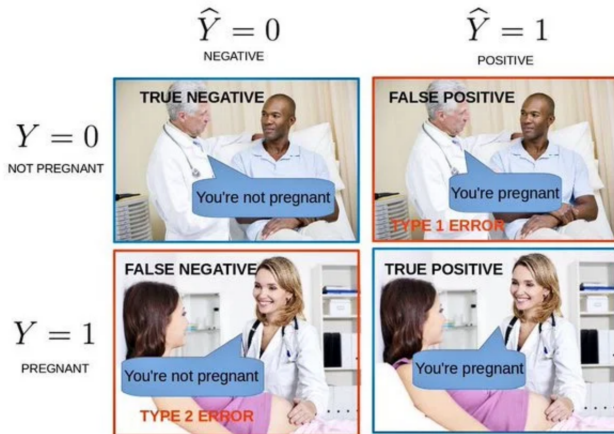
- 1 Motivation
- 2 Risk, Probability, and Classification
- 3 Bayes Classifier
- 4 Logit
 - MLE

Risk, Probability, and Classification

- ▶ Two states of nature $Y \rightarrow i \in \{0, 1\}$
- ▶ Two actions $(\hat{Y}) \rightarrow j \in \{0, 1\}$

		\hat{Y}	
		0	1
Y	0	True Negative	False Positive
	1	False Negative	True Positive

Risk, Probability, and Classification



Source: <https://dzone.com/articles/understanding-the-confusion-matrix>

Risk, Probability, and Classification

- ▶ Two actions $\hat{Y} \rightarrow j \in \{0, 1\}$
- ▶ Two states of nature $Y \rightarrow i \in \{0, 1\}$
- ▶ Probabilities
 - ▶ $p = Pr(Y = 1|X)$
 - ▶ $1 - p = Pr(Y = 0|X)$

Risk, Probability, and Classification

- ▶ Actions have costs associated to them
- ▶ Loss: $L(i, j)$, penalizes being in bin i, j
 - ▶ We define $L(i, j)$

$$L(i, j) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases} \quad (1)$$

Risk, Probability, and Classification

- Risk: expected loss of taking action j

$$E[L(i, j)] = \sum_i p_i L(i, j) \tag{2}$$
$$R(j) = (1 - p)L(0, j) + pL(1, j)$$

- The objective is to minimize the risk

Agenda

- 1 Motivation
- 2 Risk, Probability, and Classification
- 3 Bayes Classifier
- 4 Logit
 - MLE

Bayes classifier

$$R(1) < R(0) \quad (3)$$

Bayes classifier

- ▶ Under a 0-1 penalty the problem boils down to finding

$$p = Pr(Y = 1|X) \quad (4)$$

- ▶ We then predict 1 if $p > 0.5$ and 0 otherwise (Bayes classifier)
- ▶ Many ways of finding this probability in binary cases

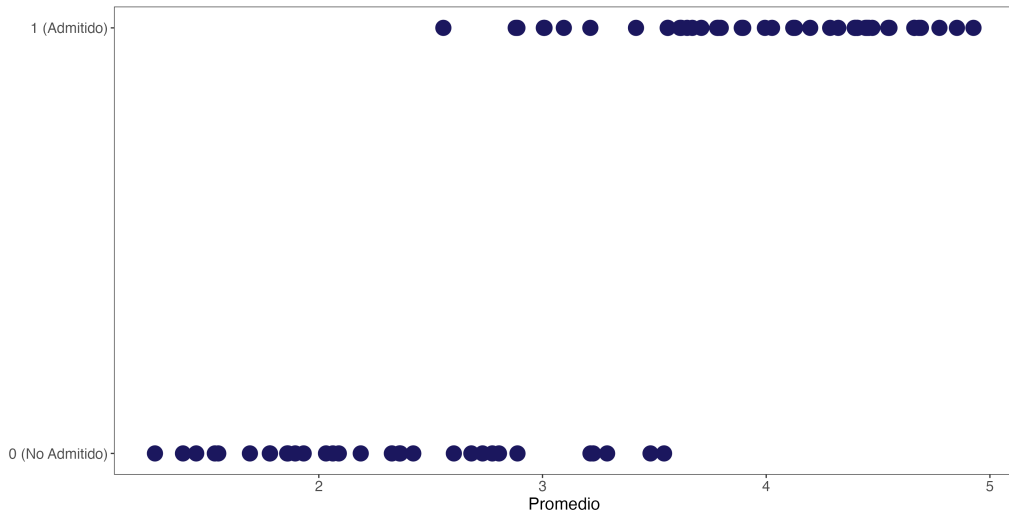
Agenda

- 1 Motivation
- 2 Risk, Probability, and Classification
- 3 Bayes Classifier
- 4 Logit
 - MLE

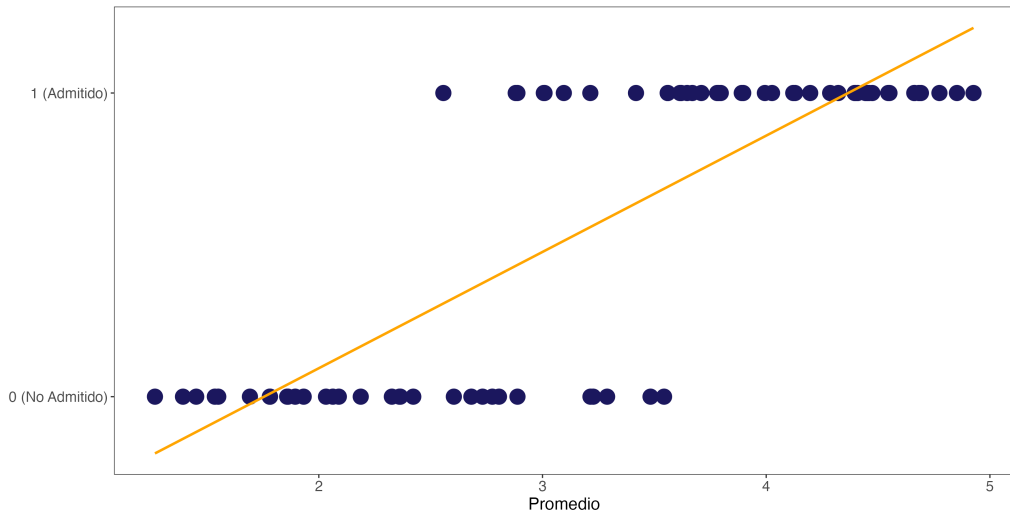
Setup

- ▶ Y is a binary random variable $\{0, 1\}$
- ▶ X is a vector of K predictors
- ▶ $p = Pr(Y = 1|X)$

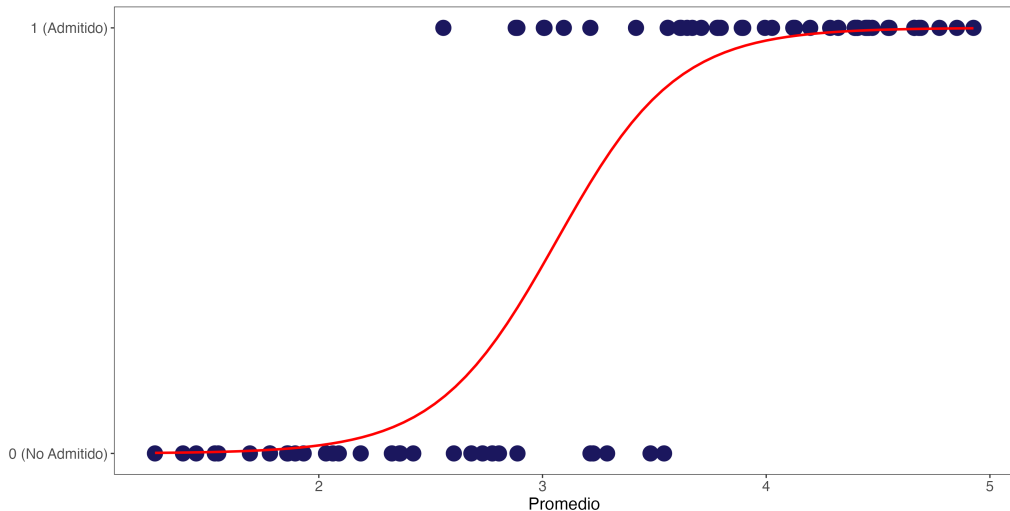
Logit



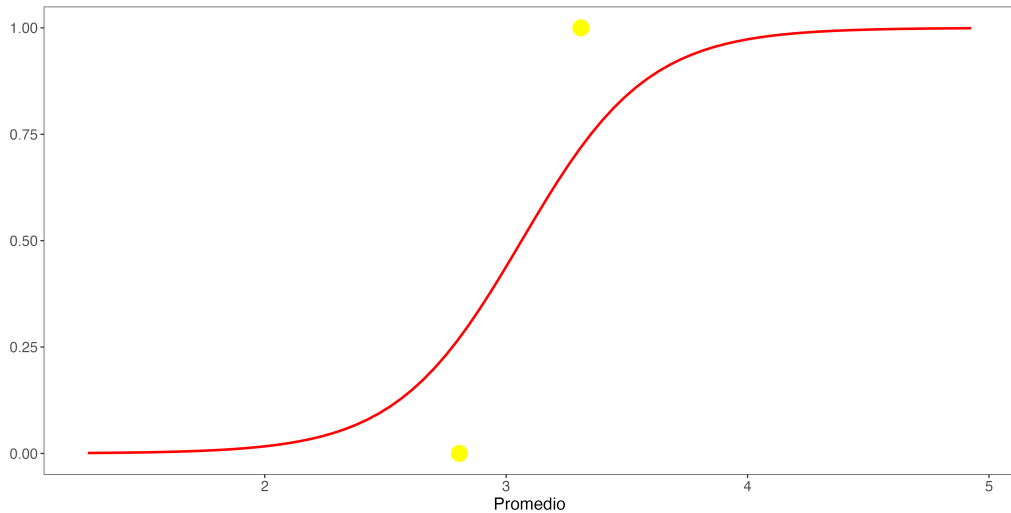
Logit



Logit



Logit



Logit

► Logit

$$\begin{aligned} p &= \frac{e^{X\beta}}{1 + e^{X\beta}} \\ &= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \end{aligned} \tag{5}$$

Logit

► Logit

$$\begin{aligned} p &= \frac{e^{X\beta}}{1 + e^{X\beta}} \\ &= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \end{aligned} \quad (5)$$

► Odds ratio

$$\begin{aligned} \ln \left(\frac{p}{1-p} \right) &= X\beta \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \end{aligned} \quad (6)$$

Agenda

- 1 Motivation
- 2 Risk, Probability, and Classification
- 3 Bayes Classifier
- 4 Logit
 - MLE

Aside: Maximum Likelihood Estimation

- ▶ Developed by Ronald A. Fisher (1890-1962)
- ▶ “If Fisher had lived in the era of “apps,” maximum likelihood estimation might have made him a billionaire” (Efron and Tibshiriani, 2016)
- ▶ Why? MLE gives “automatically”
 - ▶ Consistent
 - ▶ Asymptotically normal
 - ▶ Asymptotically efficient

Aside: Maximum Likelihood Estimation

$$Pr(Y = y|X) = f(y; \theta) \quad (7)$$

- ▶ $f()$ known
- ▶ θ unknown
- ▶ Example:

$$Y|X \sim \text{Poisson}(\lambda) \quad (8)$$

$$f(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (9)$$

Aside: Maximum Likelihood Estimation

► $Y_1, \dots, Y_n \sim_{iid} f(Y; \theta)$

$$Pr(Y_i = y_i | X_i) = f(y_i; \theta) \quad (10)$$

► Likelihood

$$L(\theta; y_i) = f(y_i; \theta) \quad (11)$$

Aside: Maximum Likelihood Estimation

- ▶ For a random sample $y_1, \dots, y_n \sim_{iid} f(y_i; \theta)$
- ▶ The likelihood function is

$$\begin{aligned} L(\theta|y_1, \dots, y_n) &= \prod_{i=1}^n L(\theta; y_i) \\ &= \prod_{i=1}^n f(x_i; \theta) \end{aligned} \tag{12}$$

- ▶ A maximum likelihood estimator of the parameter θ :

$$\hat{\theta}^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta, x) \tag{13}$$

Aside: Maximum Likelihood Estimation

- Note that maximizing (12) is the same as maximizing

$$l(\theta; y_1, \dots, y_n) = \ln L(\theta; y_1, \dots, y_n) = \sum_{i=1}^n l(\theta; y_i) \quad (14)$$

- Advantages of (14)
 - Contribution of observation i : $l_i(x|\theta) = \ln f(y_i; \theta)$
 - Eq. (12) is prone to underflow.

MLE Logit

- Imagine that we have a sample of iid observations $(y_i, x_i); i = 1, \dots, n$, where $y_i \in \{0, 1\}$
- Under logit we have

$$p = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \quad (15)$$

- Then the likelihood

$$L(\theta; y_1, \dots, y_n) = \prod_{y_i=1} p_i \prod_{y_i \neq 1} (1 - p_i) \quad (16)$$

$$= \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (17)$$

$$= \prod_{i=1}^n \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) \quad (18)$$

MLE Logit

- The log likelihood is then

$$l(\theta; y_1, \dots, y_n) = \sum_{i=1}^n \log \left(\frac{p_i}{1 - p_i} \right)^{y_i} + \sum_{i=1}^n \log(1 - p_i) \quad (19)$$

- FOC

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i}{p_i(1 - p_i)} \frac{\partial p_i}{\partial \beta_j} - \sum_{i=1}^n \frac{1}{(1 - p_i)} \frac{\partial p_i}{\partial \beta_j} \quad (20)$$

$$= \sum_{i=1}^n \frac{y_i - p_i}{p_i(1 - p_i)} \frac{\partial p_i}{\partial \beta_j} \quad (21)$$

- Note:

- This is a system of K non linear equations with K unknown parameters.
- We cannot explicitly solve for $\hat{\beta}$
- It's important to check SOC

Summary

- ▶ We observe (y_i, X_i) $i = 1, \dots, n$
- ▶ Logit

$$p_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \quad (22)$$

- ▶ Prediction

$$\hat{p}_i = \frac{e^{X_i \hat{\beta}}}{1 + e^{X_i \hat{\beta}}} \quad (23)$$

- ▶ Classification

$$\hat{Y}_i = 1[\hat{p}_i > 0.5] \quad (24)$$

Example



photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>