Classification

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Classification: Motivation

- Many predictive questions are about classification
  - Email should go to the spam folder or not
  - A household is bellow 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: $y \in \{1, 2, ..., n\}$
  - Qualitative (e.g., spam, personal, social)
  - Not necessarily ordered

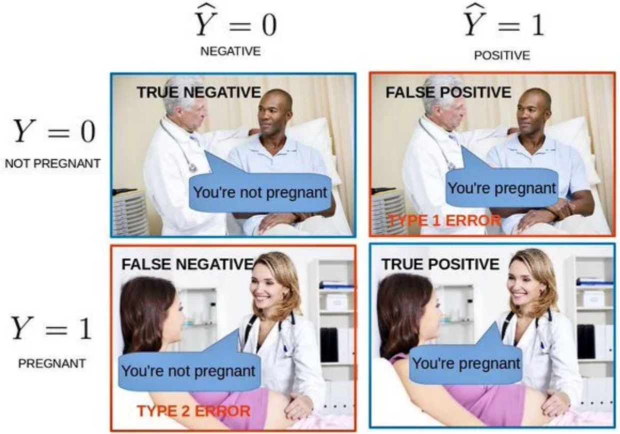
The prediction question is, given a new X,
what is our best guess at the response category $\hat{y}$

Risk, Probability, and Classification

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

		$\hat{Y}$ = 0	$\hat{Y}$ = 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 $\hat{Y} \to j \in \{0, 1\}$
- Two states of nature $Y \to i \in \{0, 1\}$
- 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 $L(i,j)$

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

Risk, Probability, and Classification

- Risk: expected loss of taking action j

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

- The objective is to minimize the risk

Bayes classifier

$$R(1) \quad < R(0) \tag{3}$$

Bayes classifier

$$
\begin{aligned}
R(1) &< R(0) \qquad (3)\\
1-p &< p
\end{aligned}
$$

Bayes classifier

$$
\begin{aligned}
R(1) &< R(0) \qquad (3) \\
1-p &< p \\
p &> \frac{1}{2}
\end{aligned}
$$

Bayes classifier

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

$$p = Pr(Y = 1|X) \tag{4}$$

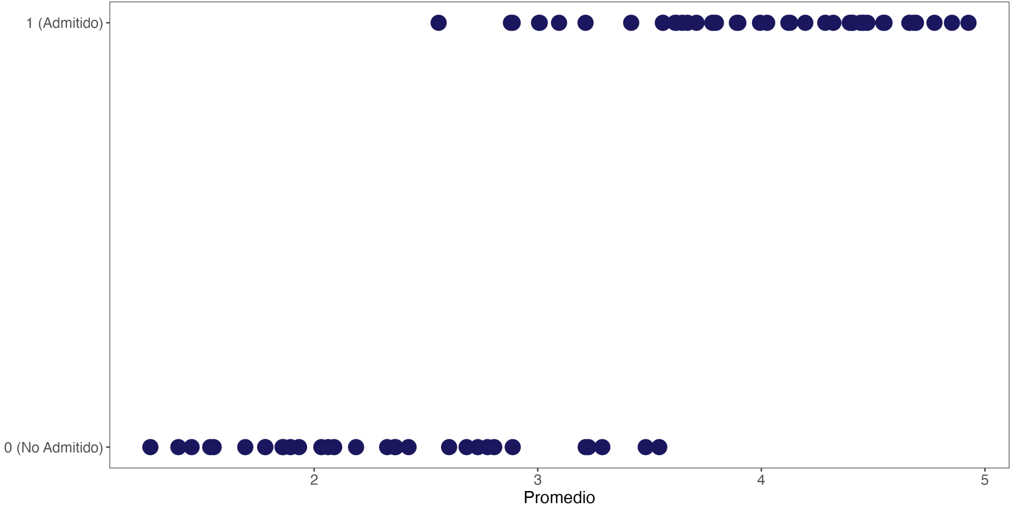
- We then predict 1 if $p > 0.5$ and 0 otherwise (Bayes classifier)
- Many ways of finding this probability in binary cases

Agenda

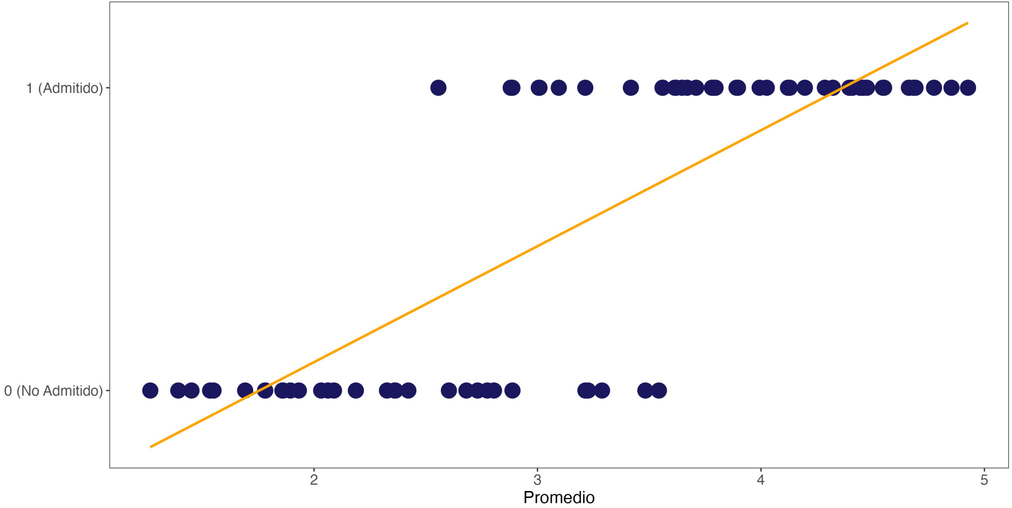
Setup

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

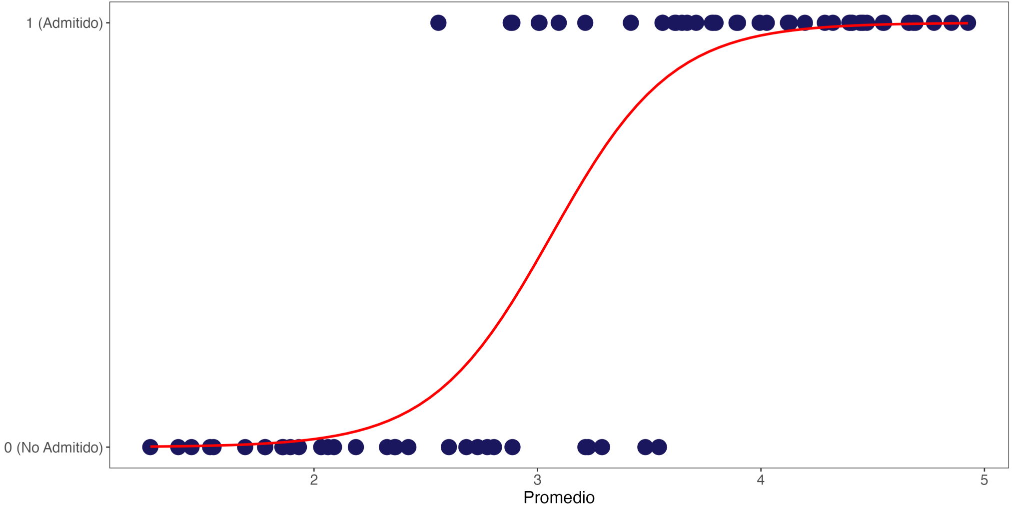
Logit



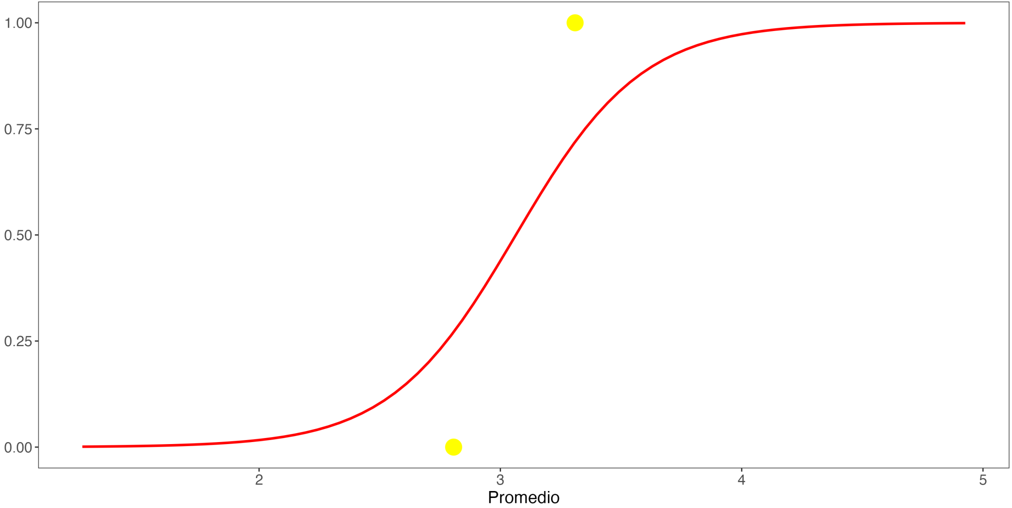
Logit



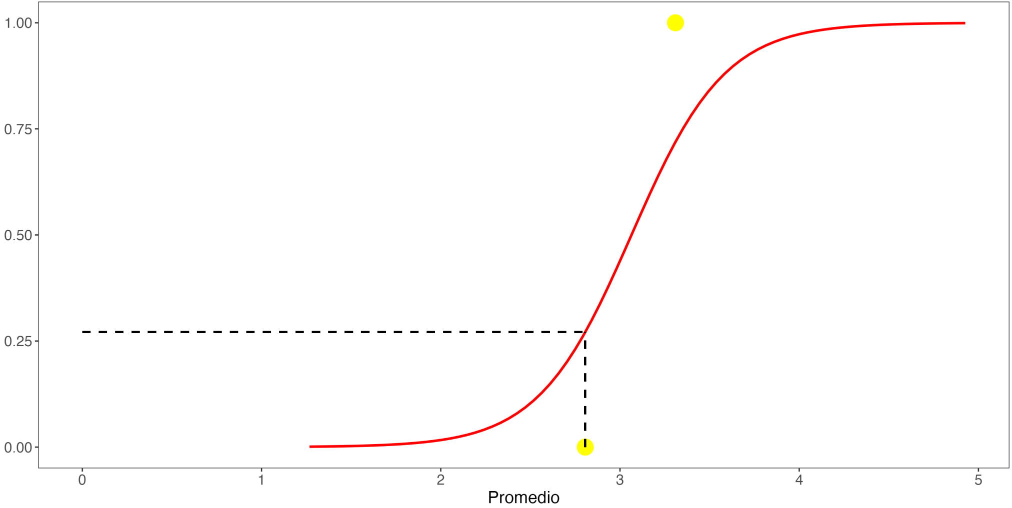
Logit



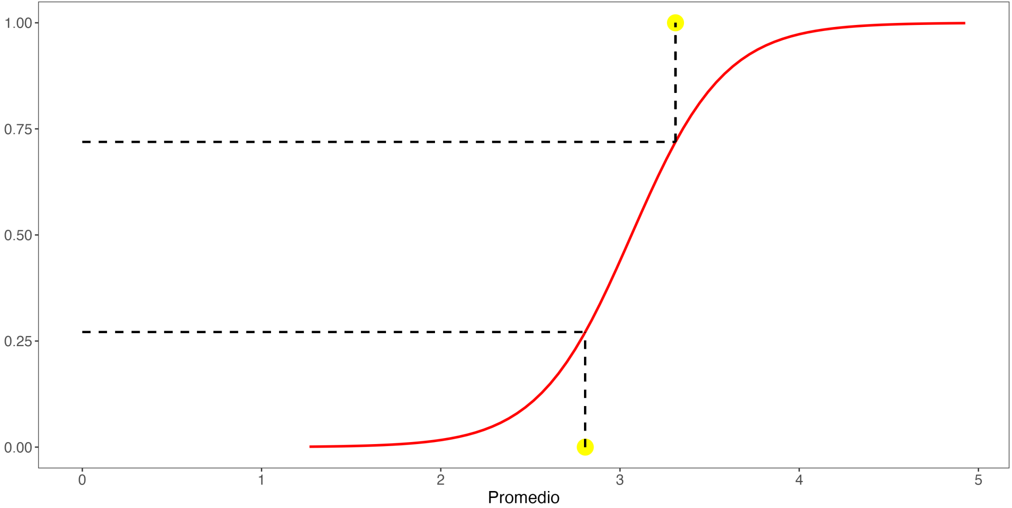
Logit



Logit



Logit



Logit

- Logit

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

Logit

- Logit

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

- Odds ratio

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

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

$$Pr(Y = y|X) = f(y;\theta) \tag{7}$$

- $f()$ known
- θ unknown
- Example:

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

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

Aside: Maximum Likelihood Estimation

- $Y_1, \ldots, Y_n \sim_{iid} f(Y;\theta)$

$$Pr(Y_i = y_i | X_i) = f(y_i;\theta) \tag{10}$$

- Likelihood

$$L(\theta; y_i) = f(y_i;\theta) \tag{11}$$

Aside: Maximum Likelihood Estimation

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

$$L(\theta | y_1, \ldots, y_n) = \Pi_{i=1}^{n} L(\theta; y_i) \quad (12)$$
$$= \Pi_{i=1}^{n} f(x_i; \theta)$$

- A maximum likelihood estimator of the parameter θ:

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

Aside: Maximum Likelihood Estimation

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

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

- 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); $i = 1, \ldots, n$, where $y_i \in \{0, 1\}$
- Under logit we have

$$p = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \tag{15}$$

- Then the likelihood

$$\begin{aligned} L(\theta; y_1, \ldots, y_n) &= \Pi_{y_i=1} p_i \Pi_{y_i \neq 1} (1 - p_i) && (16) \\ &= \Pi_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} && (17) \\ &= \Pi_{i=1}^{n} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) && (18) \end{aligned}$$

MLE Logit

- The log likelihood is then

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

- FOC

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

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

- 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) $i = 1, \ldots, n$
- Logit

$$p_i = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \tag{22}$$

- Prediction

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

- Classification

$$\hat{Y}_i = 1[\hat{p}_i > 0.5] \tag{24}$$

Example

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