Classification

Machine Learning

Ignacio Sarmiento-Barbieri

Universidad de La Plata

Agenda

Recap

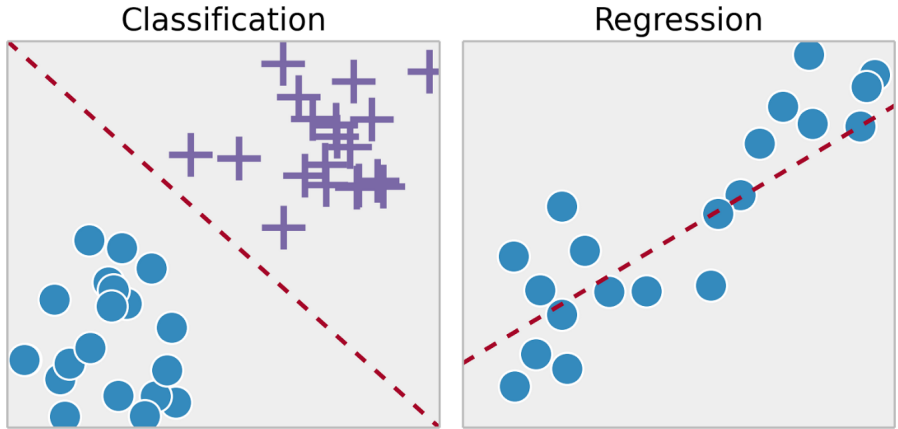
- Queremos predecir y en funcion de observables ($\mathbf{x}$)

$$y = f(\mathbf{x}) + u \tag{1}$$

- donde la estimación de f implica la que minimize el riesgo (prediga mejor fuera de muestra):

$$\hat{f} = \underset{f}{\operatorname{argmin}} \{E\left[L(y, f(\mathbf{x}; \Theta))\right]\} \tag{2}$$

Classification



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

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 $\hat{y}$

Agenda

1. Motivation
2. Risk, Probability, and Classification
3. Missclassification
   - Accuracy
   - TNR
   - TNR
   - ROC curve
   - Imbalanced Classification

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

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 $L(i,j)$

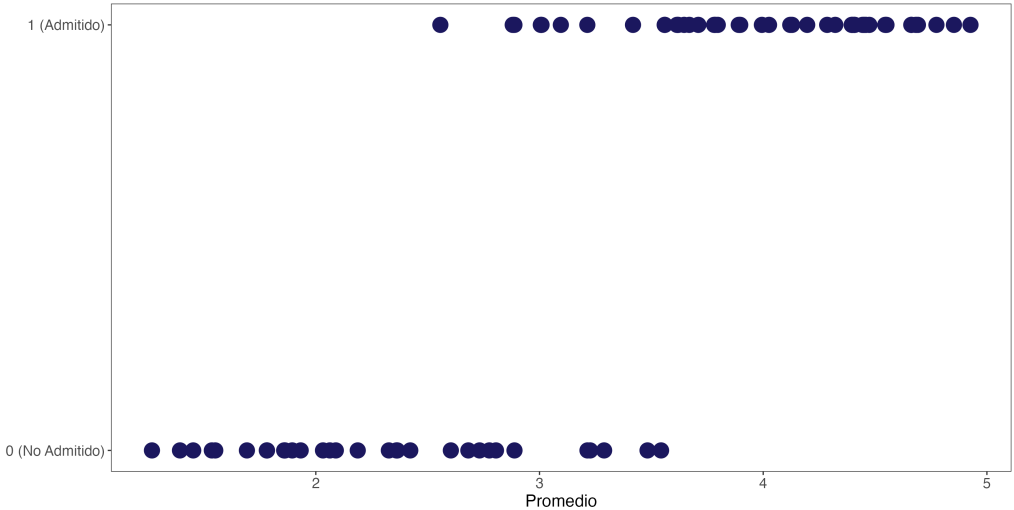
$$L(i,j) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases} \tag{3}$$

- Risk: expected loss of taking action j

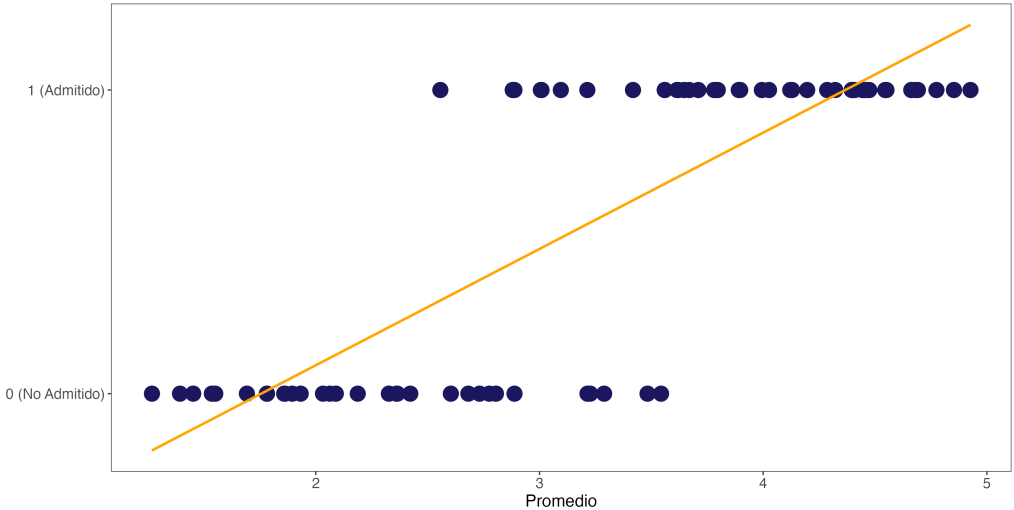
Bayes classifier

$$R(1) \quad < R(0) \tag{4}$$

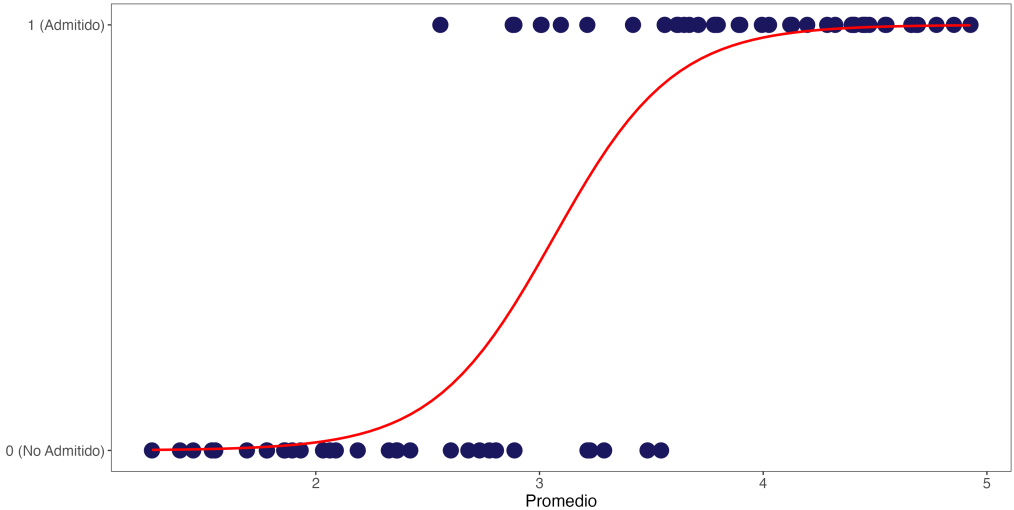
Estimating probabilities



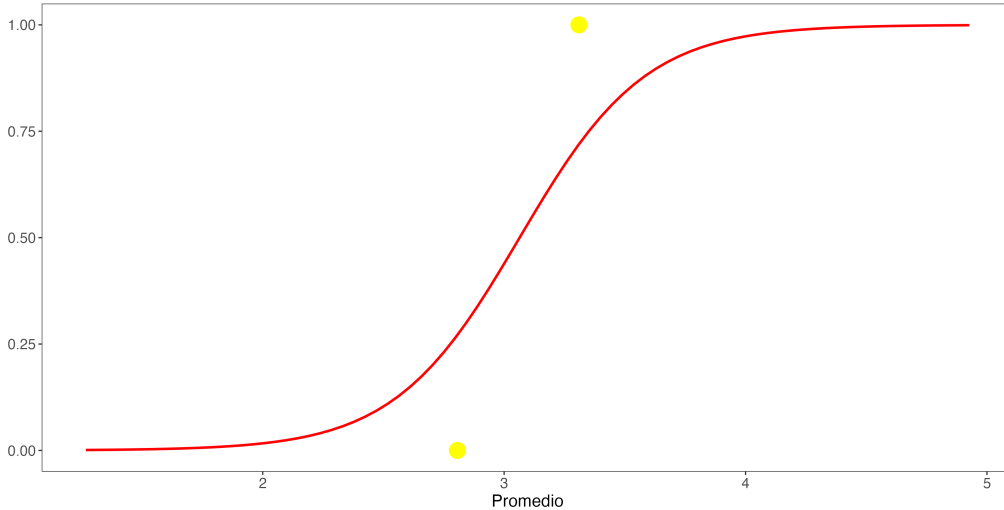
Estimating probabilities



Estimating probabilities



Estimating probabilities



Estimating probabilities

Summary

- We observe (y_i, X_i) $i = 1, \ldots, n$
- Estimate probabilities, e.g. LPM, Logit, etc

$$p_i = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \tag{5}$$

- Predictict probabilities

$$\hat{p}_i = \frac{e^{X_i\hat{\beta}}}{1 + e^{X_i\hat{\beta}}} \tag{6}$$

- Classify

$$\hat{Y}_i = 1[\hat{p}_i > 0.5] \tag{7}$$

Agenda

1. Motivation
2. Risk, Probability, and Classification
3. Missclassification
   - Accuracy
   - TNR
   - TNR
   - ROC curve
   - Imbalanced Classification

Confusion Matrix: Metrics

		y_i = 1	y_i = 0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

Accuracy

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

- Accuracy: the fraction of predictions our model got right.

$$\frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

TNR

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$P[\hat{y} = 0|y = 0] = \frac{TN}{TN + FP} \tag{9}$$

TPR

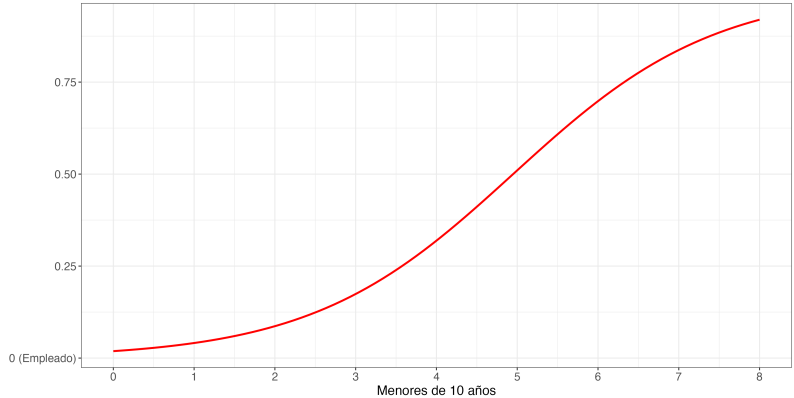
		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$P[\hat{y} = 1|y = 1] = \frac{TP}{TP + FN} \tag{10}$$

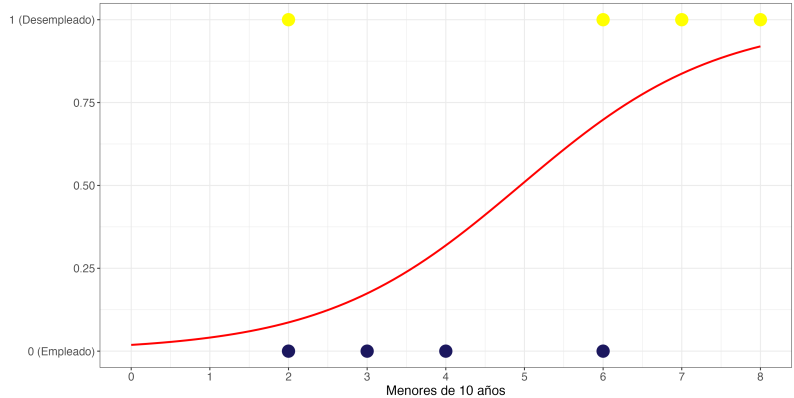
Example: Unemployment

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

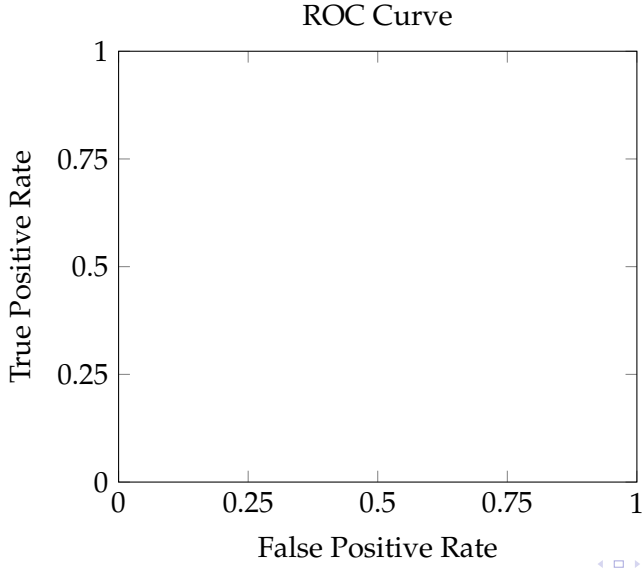
Trade-Off between Different Classification Thresholds



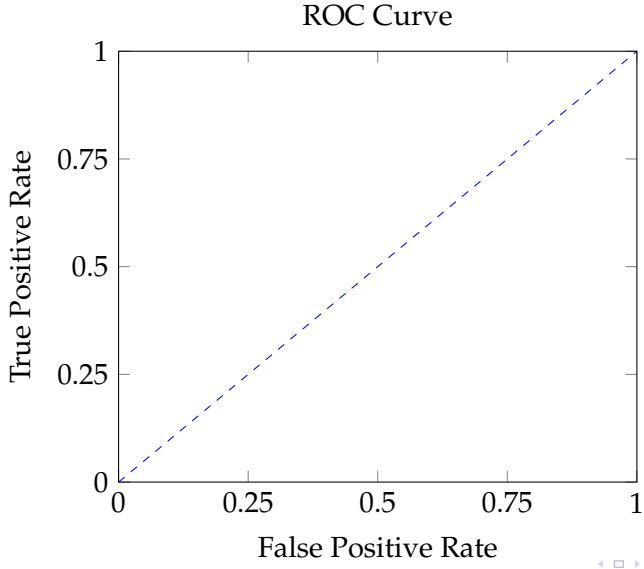
Trade-Off between Different Classification Thresholds



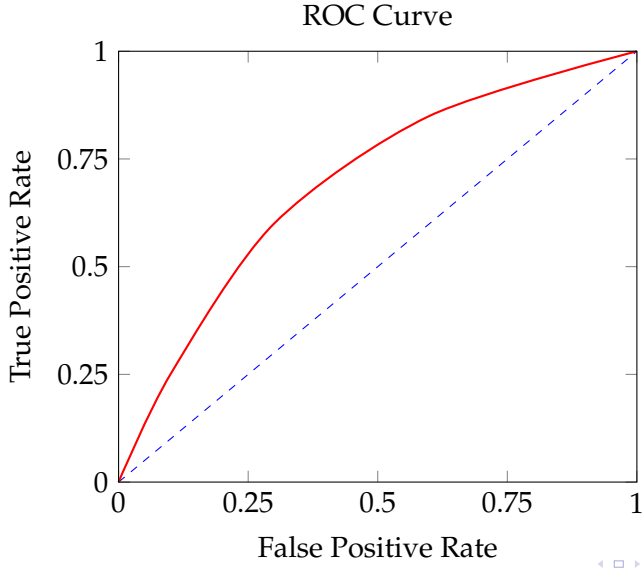
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1

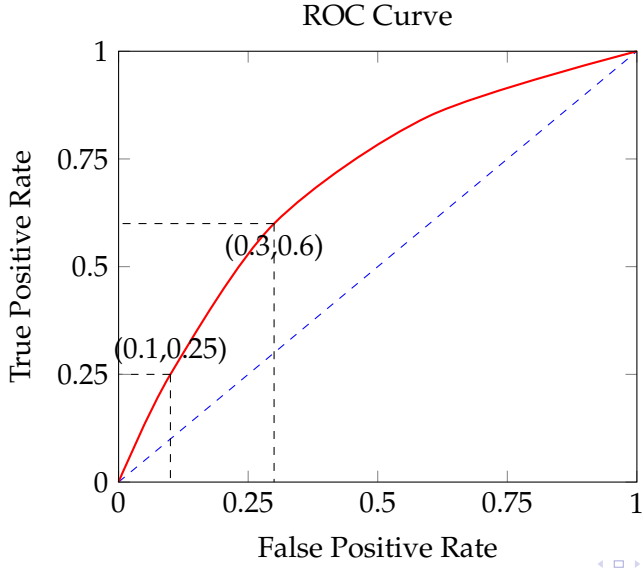
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1

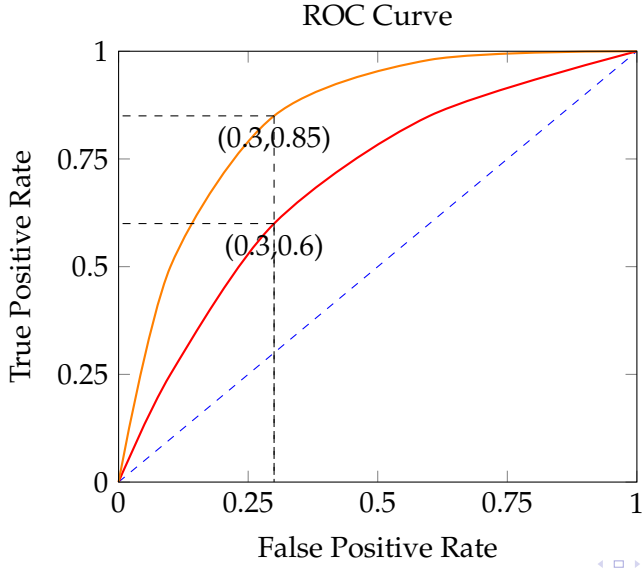
ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1
(0.3,0.6)
(0.1,0.25)

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
(0.3,0.85)
(0.3,0.6)
0
0.25
0.5
0.75
1

Example: Unemployment

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

Imbalanced Classification: Motivation

- Interest in one of the classes: Poor, Default, Unemployed, Fraud
- Imbalanced classes pose a challenge

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	$<$1% of the data set

TPR & PPV

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$P[\hat{y} = 1|y = 1] = \frac{TP}{TP + FN} \tag{11}$$

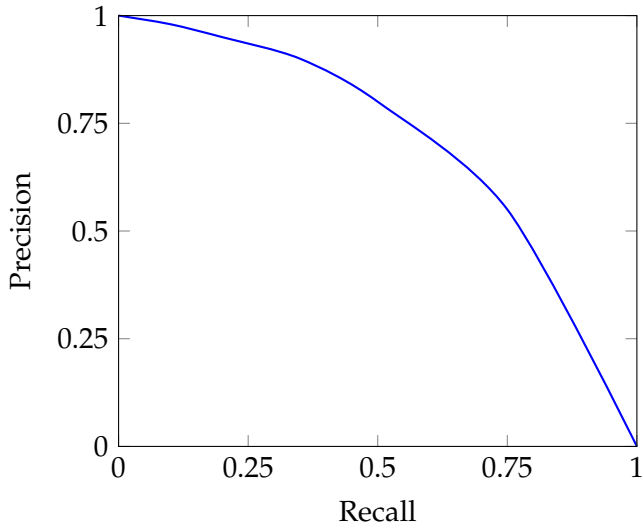
TPR & PPV

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$P[\hat{y} = 1|y = 1] = \frac{TP}{TP + FN} \tag{11}$$

$$P[y = 1|\hat{y} = 1] = \frac{TP}{TP + FP} \tag{12}$$

PR-Curve



F-Scores

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

F-Scores

		y_i	
		1	0
$\hat{y}_i$	1	TP	FP
	0	FN	TN

$$F_\beta = (1 + \beta^2) \frac{Precision \times Recall}{(\beta^2 \times Precision + Recall)} \tag{14}$$

Example: Unemployment

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