Classification: Performance Metrics & Class Imbalance

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Classification: Motivation

- Many predictive questions are about classification
  - Credit, Poverty, Firm default, Fraud, Unemployment, etc.
- Aim is to classify y, where y represents membership in a category
  - Qualitative, not necessarily ordered
  - We will focus for now in the binary case

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

Classification: Recap

2 Estados de la Naturaleza $\{0, 1\}$ → {Empleado, Desempleado}

→ {No Pobre, Pobre}

$P_i = Pr(y = 1 | x) = f(x) + u$

- LPM
- Logit, Probit
- No param: KNN
- LDA
- QDA

vínculo con Logit.

- Naive Bayes

Función indicadora. Se cumple [.] → 1

No → 0

$$\overbrace{1[p_i \geq c]}$$

$c = 0.5$ → Sale de una pérdida simétrica.

$\hat{P}_i \geq 0.5 \rightarrow 1$

$\hat{P}_i < 0.5 \rightarrow 0$

Logit

$\ell \to$ concava

$-\ell \to$ convexa.

- The log likelihood is

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]$$

  where $p_i = P(y_i = 1 | X_i) = \underline{f(X_i)}$
- 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 $\to$ Hessiano.

Computational algorithms

Gradiente Descendente

hiperparametro del algoritmo

$$\beta^{j+1} = \beta^{j} - \boxed{\varepsilon} \nabla L(\beta^{j})$$

→ Aproximar lineal

Newton Method.

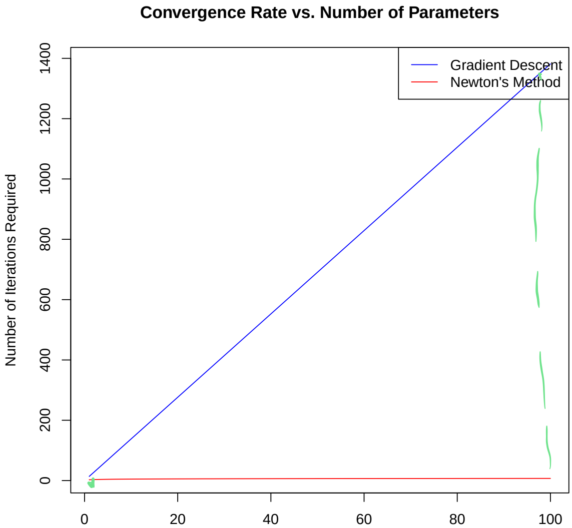
$$L(\overbrace{\beta^{j}+d}^{\beta^{j+1}}) \approx L(\beta^{j}) + d \nabla L(\beta^{j}) + \frac{1}{2} d' \nabla^{2} L(\beta) d$$

$$d = \underline{H}^{-1} \nabla L(\beta^{j})$$

$$\beta^{j+1} = \beta^{j} - \boxed{H^{-1}} \nabla L(\beta^{j})$$

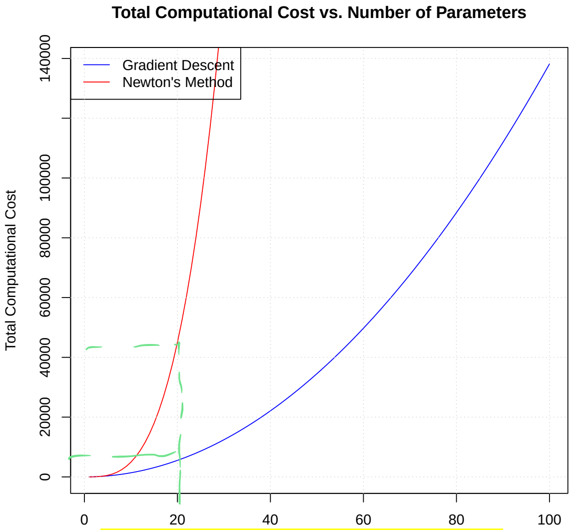
Computational algorithms

Gradient Descent vs Newton's Method



Computational algorithms

Gradient Descent vs Newton's Method



Probit

d

logistica

- $Pr(y = 1|X) = \Phi(X'\beta)$ where Φ is the standard normal cdf.
- In practice, the probit and logit models generally yield very similar predicted probabilities,
- There are practical reasons for favoring one or the other in some cases for mathematical convenience, in other computational convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds.

Example: Unemployment

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

Agenda

Confusion Matrix

y_i : Estado Nat (Referencia)

		1 Desemplado	0 Empleado	
$\hat{y}_i$ Predicción	1	True Positive TP Verdadero Positivo	FP False Positivo	Predichos Positivos $\hat{P}$
	0	FN False Negative	TN True Negative.	Predichos Negativos $\hat{N}$
		Actual Positivos P	Actual Negatives N	

Accuracy

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

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} = \frac{\text{Aciertos}}{\text{Total de Obs.}} \tag{1}$$

TNR

: - TRUE Negative Rate

- Specificity.

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

4
2
4

$$P[\hat{y}=0|y=0] = \frac{TN}{TN+FP} = \frac{\text{TRUE Neg}}{\text{Ac Negativos}} \quad (2)$$

False Positive Rate $= 1 - TNR = \frac{FP}{N}$

Error tipo 1 $= P(\hat{y}=1 \mid y=0)$

TPR : - True positive Rate
- Sensitivity
- Recall

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

$$P[\hat{y} = 1 | y = 1] = \frac{TP}{TP + FN} = \frac{TP}{\text{Positives}} \tag{3}$$

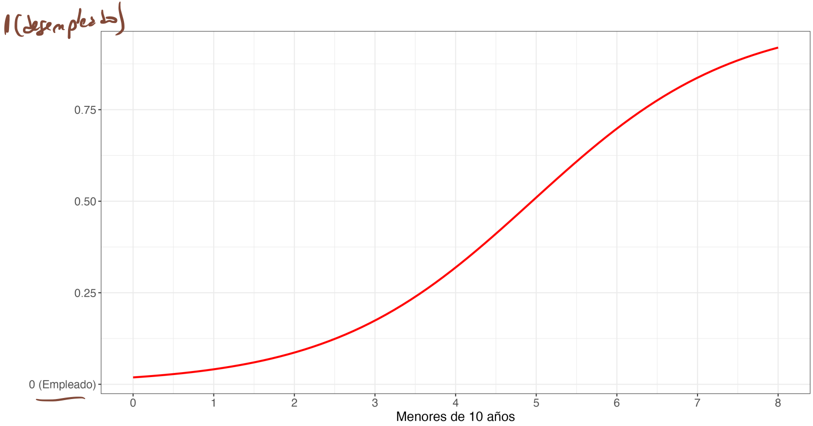
Error tipo II $\Rightarrow$ $P(\hat{y}=0 | y=1) = 1 - TPR = \frac{FN}{\text{Positives}}$

Example: Unemployment

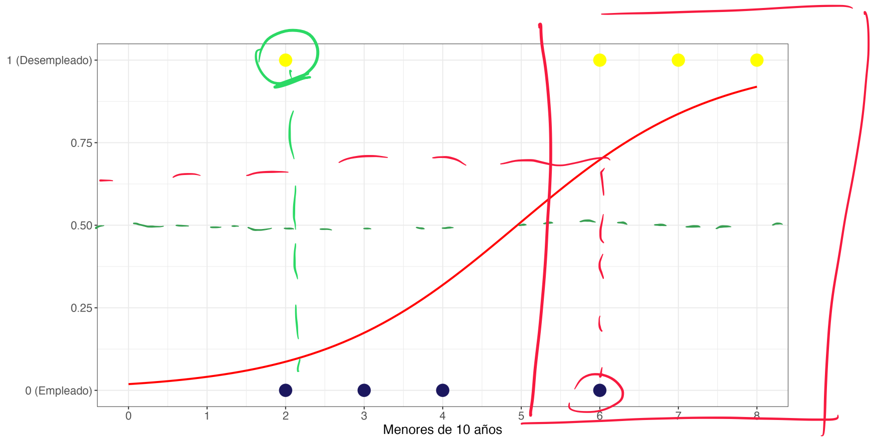
photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

Trade-Off between Different Classification Thresholds

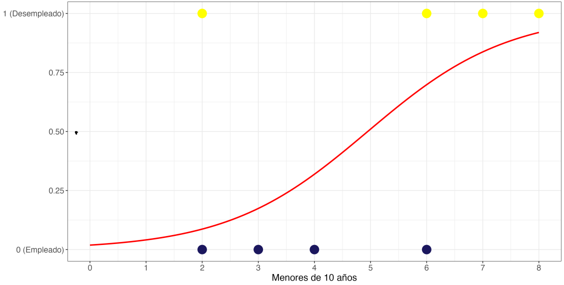


Trade-Off between Different Classification Thresholds



Trade-Off between Different Classification Thresholds

$$\mathbb{1}[\hat{p}_i > c]$$



$c = 1$

$$\begin{array}{c|cc} & y=1 & y=0 \\ \hline \hat{y}=1 & 4 & 4 \\ \hat{y}=0 & 0 & 0 \end{array}$$

$TPR = 1$

$FPR = 1 \leftarrow TNR = 0$

$c = 0$

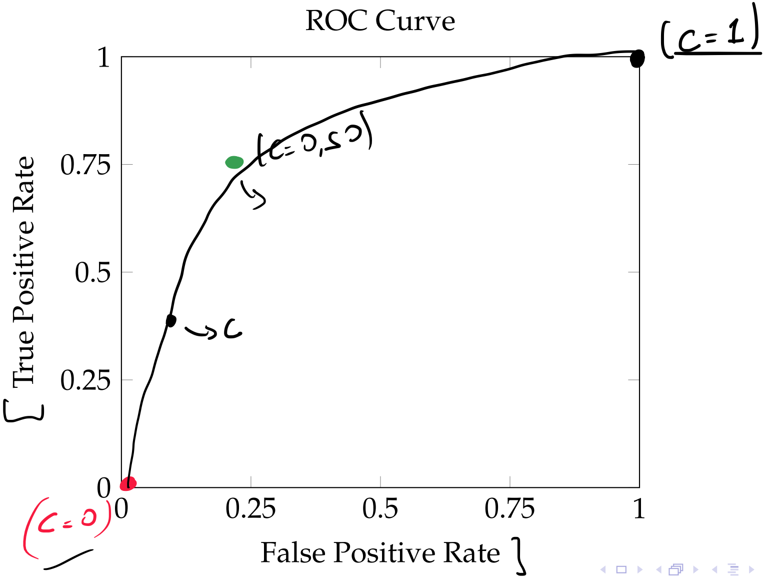
$$\begin{array}{c|cc} & y=1 & y=0 \\ \hline \hat{y}=1 & 0 & 0 \\ \hat{y}=0 & 4 & 4 \end{array}$$

$TPR = 0$

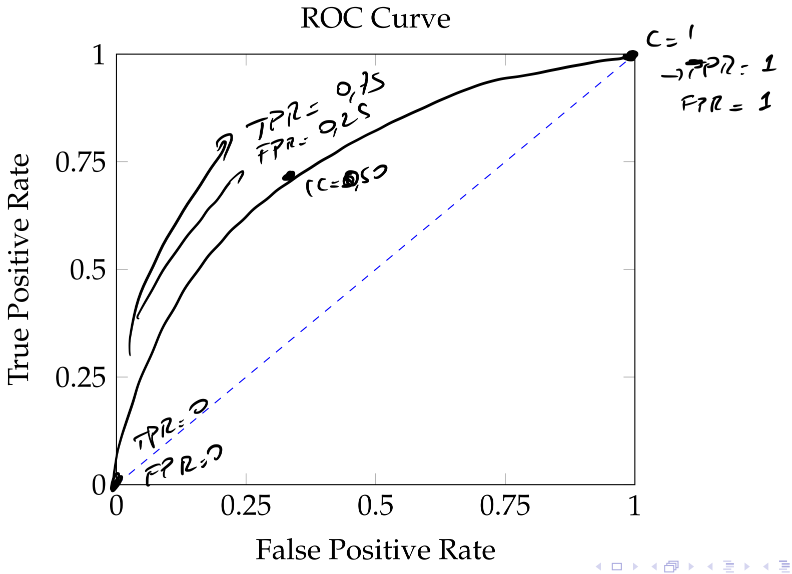
$FPR = 1 - TNR = 0$

$TNR = 1$

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1
(c=1)
(c=0,50)
c
(c=0)

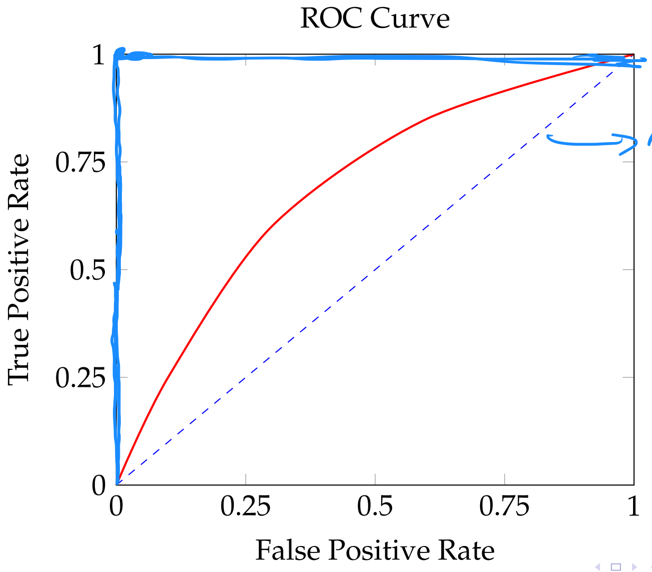
ROC Plot



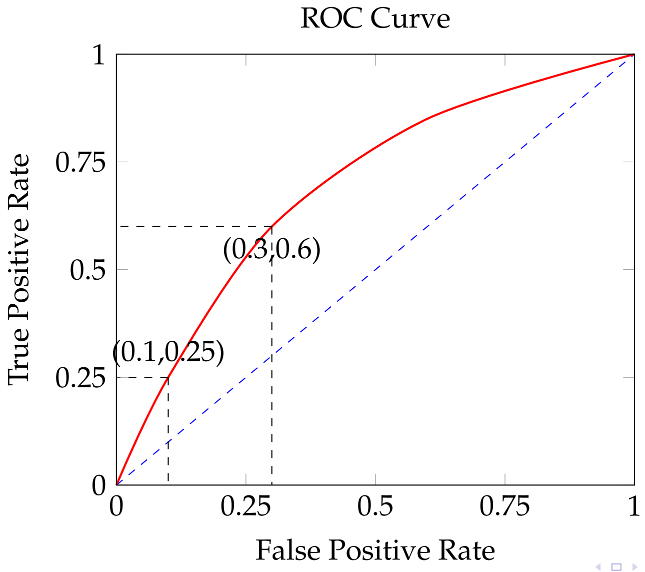
ROC Curve
True Positive Rate
False Positive Rate
0
0.25
0.5
0.75
1
c = 1
TPR = 1
FPR = 1
TPR = 0,75
FPR = 0,25
c = 0,50
TPR = 0
FPR = 0

ROC Plot

AUC = 1

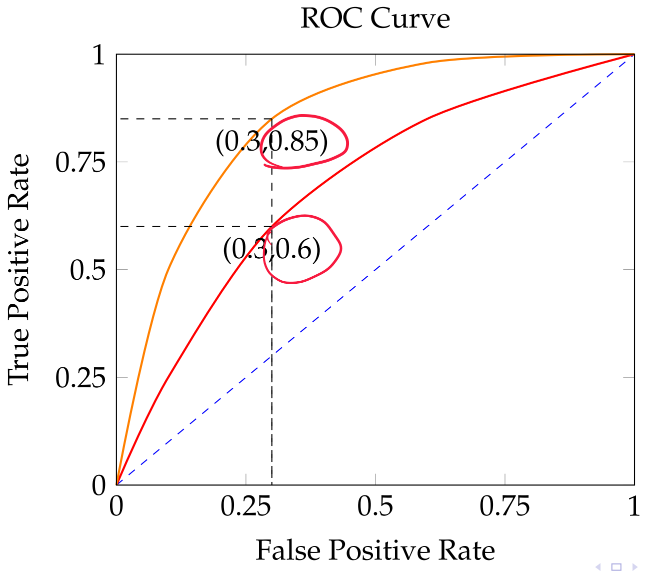


ROC Plot



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

ROC Plot



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

Example: Unemployment

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

Agenda

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

$$
\begin{array}{ccccc}
 & & & y_i & \\
 & & 1 & & 0 \\
\hat{y}_i & 1 & \boxed{TP} & & FP \\
 & 0 & FN & & \boxed{TN}
\end{array}
$$

$$\boxed{P[\hat{y} = 1 | y = 1] = \frac{TP}{TP + FN} = \begin{matrix} TPR \\ Recall \end{matrix}} \tag{4}$$

$$TNR = \frac{TN}{FP + TN}$$

$$\boxed{FPR = \frac{FP}{FP + \boxed{TN}}}$$

TPR & PPV

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

$$P[\hat{y} = 1 | y = 1] = \frac{TP}{TP + FN} \tag{4}$$

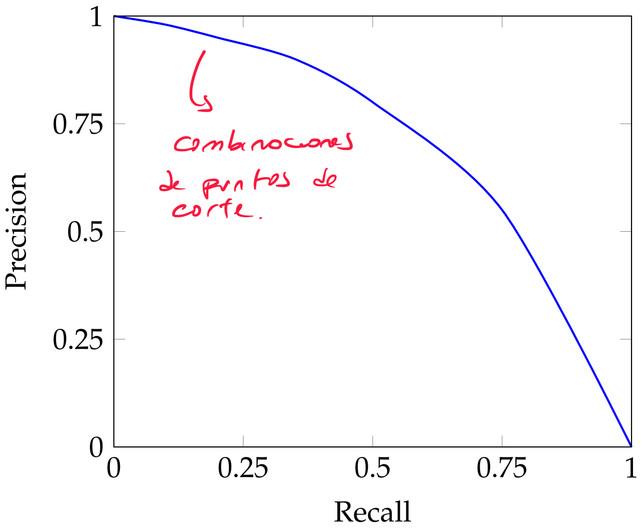
= TPR

Recall

$$P[y = 1 | \hat{y} = 1] = \frac{TP}{TP + FP} \tag{5}$$

= Precision

PR-Curve



combinaciones de puntos de corte.

Desbalance de clase

Vamos a pensar como remediarlo. para predecir Bien los desempleados.

F-Scores

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

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Precision = 0,90

Recall = 0,10

$F_1 = 0{,}18$

F-Scores

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

$$F_\beta = (1+\underbrace{\beta^2}_{1})\frac{Precision \times Recall}{(\beta^2 \times Precision + Recall)} \tag{7}$$

$\beta = 2$

$\beta = 0,5$

$\Big\}$ cual medida es enfatizada.

H.W.

Example: Unemployment

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