

Classification: Performance Metrics & Class Imbalance

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1 Recap

- Computational algorithms
- Probit

2 Confusion Matrix

- Accuracy
- TNR
- TNR

3 ROC curve

4 Imbalanced Classification

- Metrics

Agenda

1 Recap

- Computational algorithms
- Probit

2 Confusion Matrix

- Accuracy
- TNR
- TNR

3 ROC curve

4 Imbalanced Classification

- Metrics

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

$$1[p_i \geq c]$$

Logit

- The log likelihood is

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

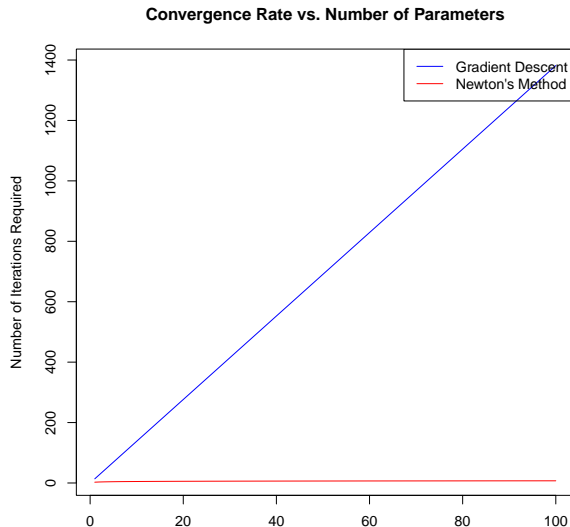
where $p_i = P(y_i = 1 | X_i) = f(X_i)$

- 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

Computational algorithms

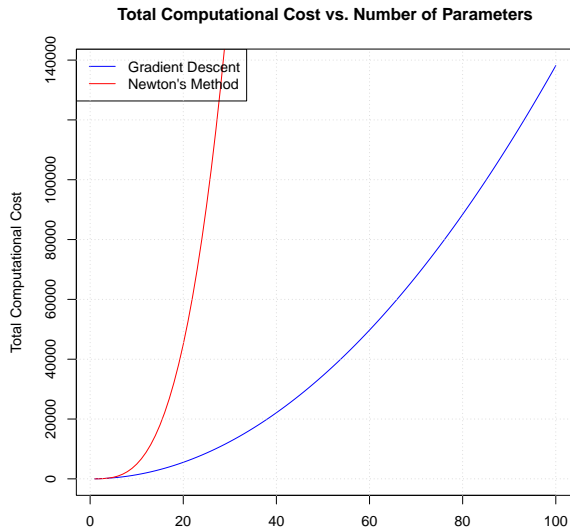
Computational algorithms

Gradient Descent vs Newton's Method



Computational algorithms

Gradient Descent vs Newton's Method



Probit

- ▶ $Pr(y = 1|X) = \Phi(X'\beta)$ 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

1 Recap

- Computational algorithms
- Probit

2 Confusion Matrix

- Accuracy
- TNR
- TNR

3 ROC curve

4 Imbalanced Classification

- Metrics

Confusion Matrix

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

Accuracy

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

$$\frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

TNR

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

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

TPR

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

$$P[\hat{y} = 1 | y = 1] = \frac{TP}{TP + FN} \quad (3)$$

Example: Unemployment

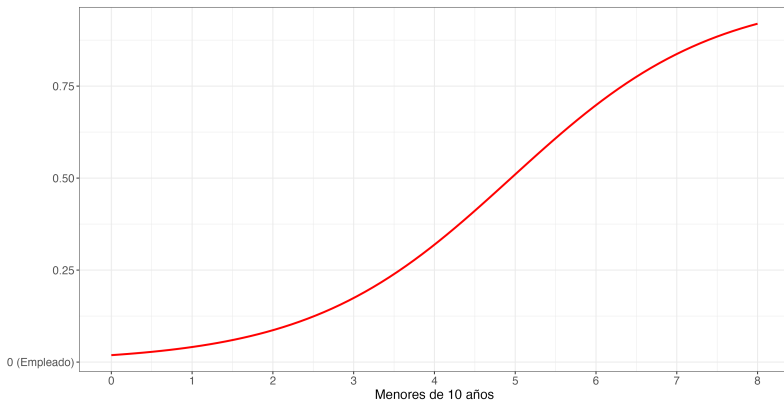


photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>

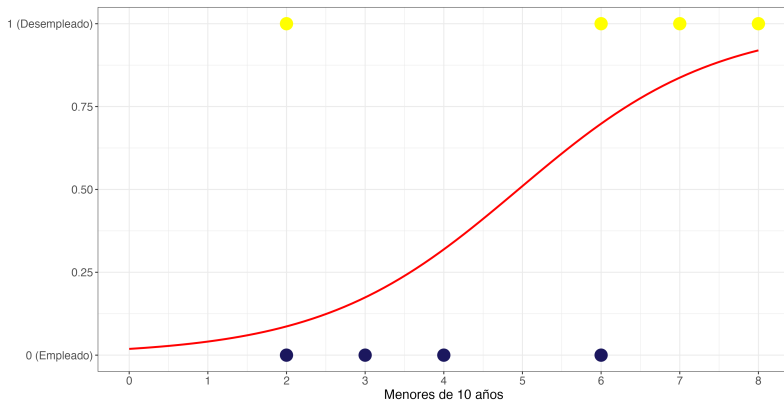
Agenda

- ① Recap
 - Computational algorithms
 - Probit
- ② Confusion Matrix
 - Accuracy
 - TNR
 - TNR
- ③ ROC curve
- ④ Imbalanced Classification
 - Metrics

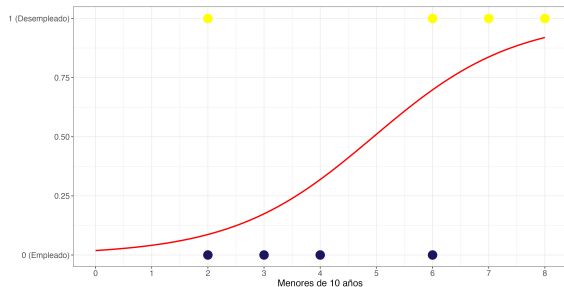
Trade-Off between Different Classification Thresholds



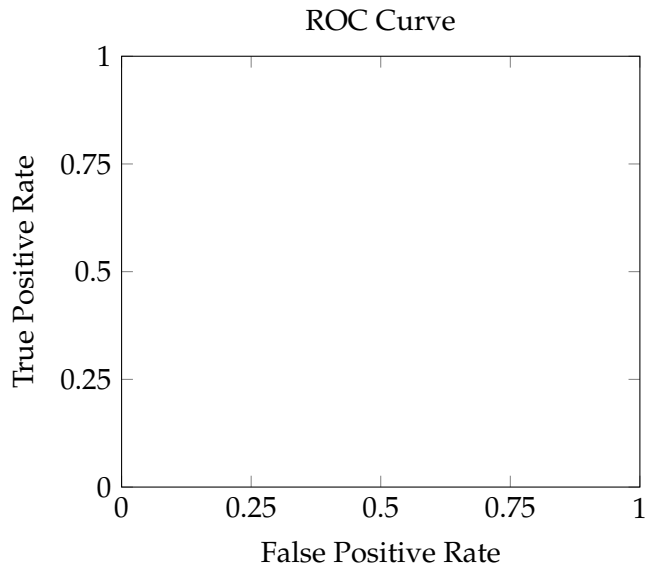
Trade-Off between Different Classification Thresholds



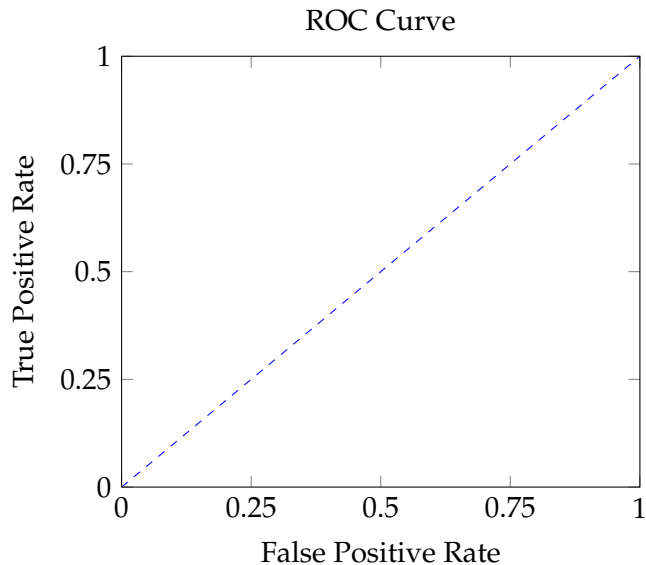
Trade-Off between Different Classification Thresholds



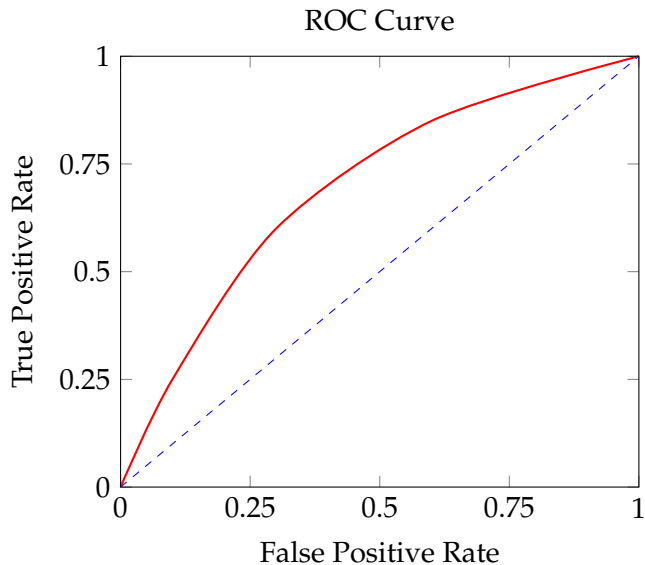
ROC Plot



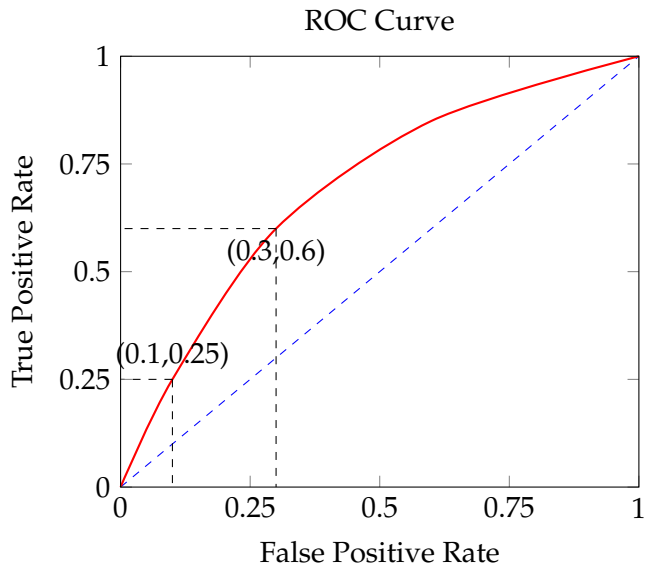
ROC Plot



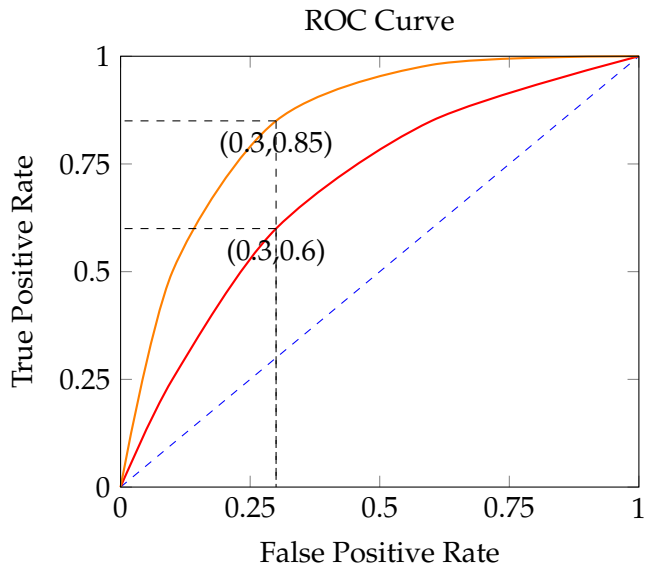
ROC Plot



ROC Plot



ROC Plot



Example: Unemployment



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

Agenda

- ① Recap
 - Computational algorithms
 - Probit
- ② Confusion Matrix
 - Accuracy
 - TNR
 - TNR
- ③ ROC curve
- ④ Imbalanced Classification
 - Metrics

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

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

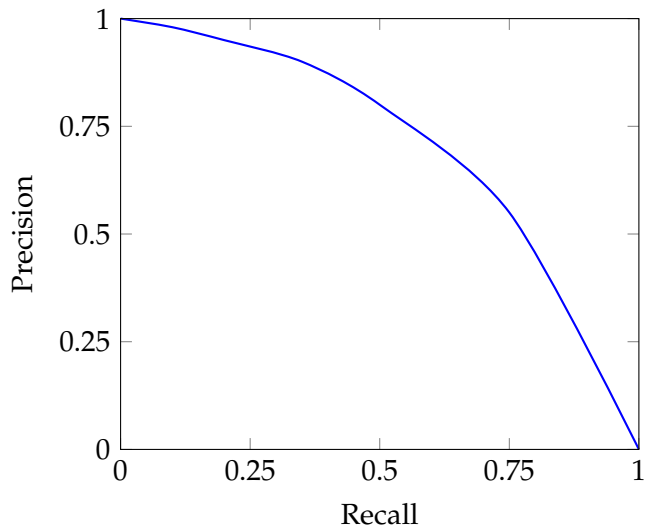
TPR & PPV

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

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

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

PR-Curve



F-Scores

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

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

F-Scores

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

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

Example: Unemployment



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