Classification: Performance Metrics & Class Imbalance

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

1. Recap
   - Probit
   - Generative Models for Classification
2. Confusion Matrix
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

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

Logit

- The log likelihood is

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \Big[y_i \log Pr(y_i = 1|X_i) + (1 - y_i) \log(1 - Pr(y_i = 1|X_i)) \Big]$$

  where $p_i = Pr(y_i = 1|X_i) = \frac{e^{X}\beta}{1+e^{X}\beta}$

- 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

Probit

- $Pr(y_i = 1|X_i) = \Phi(X_i'\beta)$ 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.

Generative Models for Classification

- LDA
- QDA
- Naive Bayes

Example: Unemployment

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

Agenda

Confusion Matrix: Metrics

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

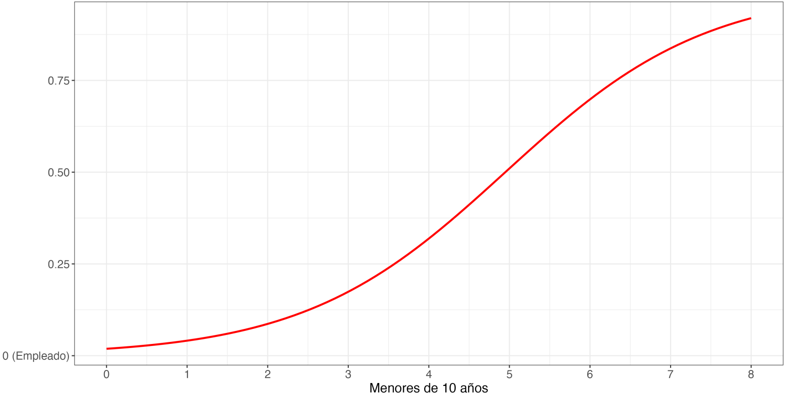
Example: Unemployment

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

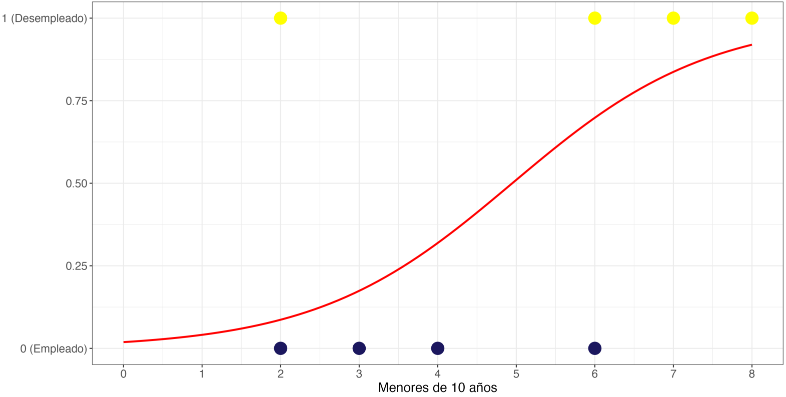
Agenda

1. Recap
   - Probit
   - Generative Models for Classification
2. Confusion Matrix
3. ROC curve
4. Imbalanced Classification
   - Metrics

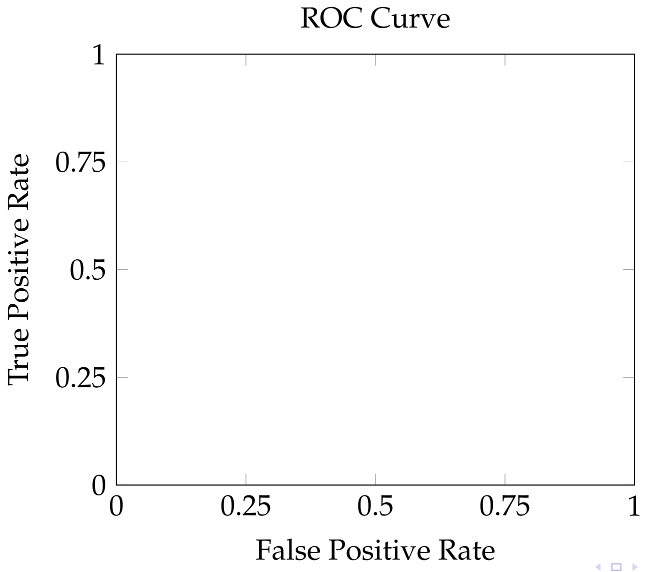
Trade-Off between Different Classification Thresholds



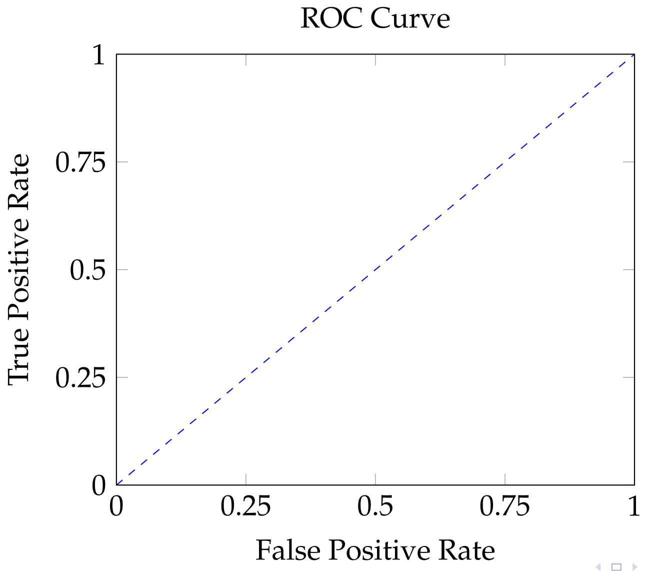
Trade-Off between Different Classification Thresholds



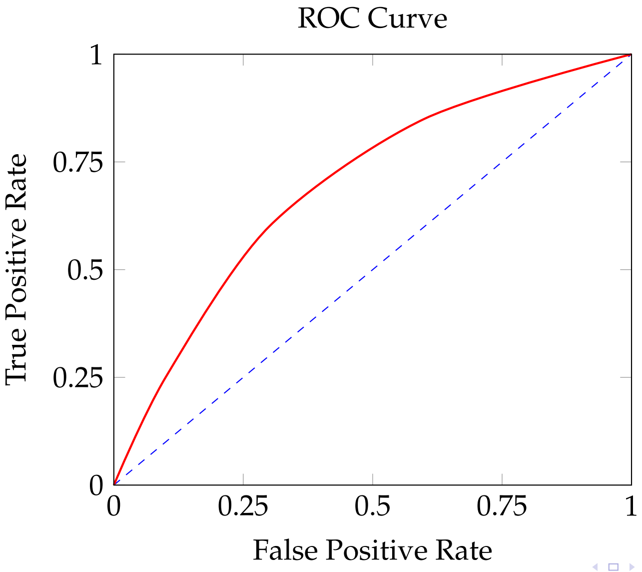
ROC Plot


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

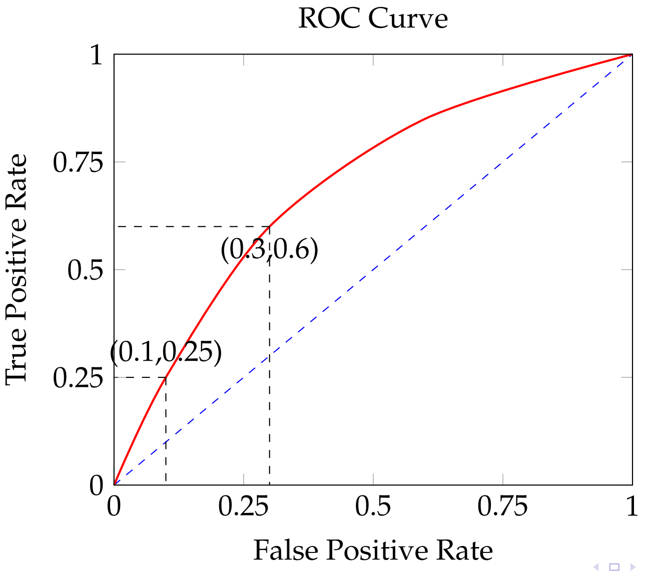
ROC Plot


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

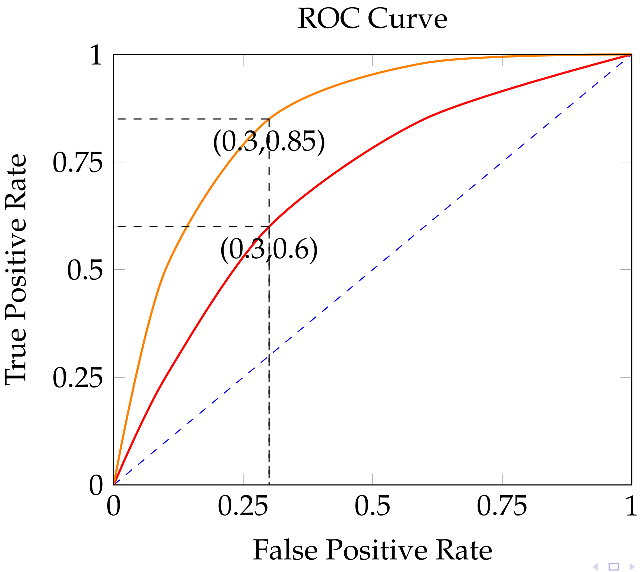
ROC Plot


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

ROC Plot


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

ROC Plot


ROC Curve
1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
(0.3,0.85)
(0.3,0.6)
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

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

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

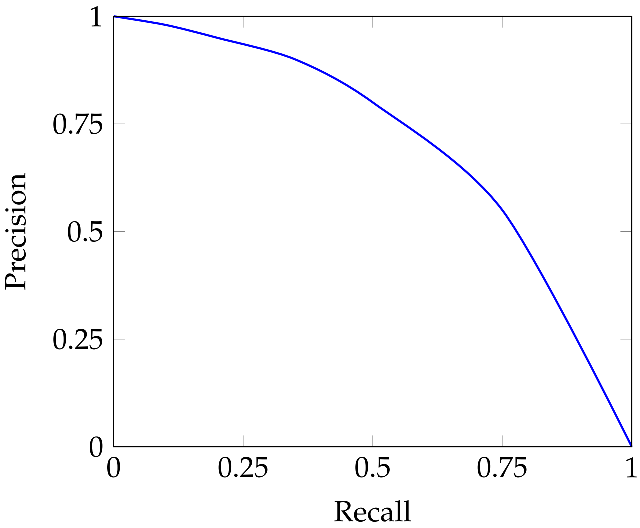
TPR & PPV

$$
\begin{array}{cccc}
 & & \multicolumn{2}{c}{y_i} \\
 & & 1 & 0 \\
\hat{y}_i & 1 & \text{TP} & \text{FP} \\
 & 0 & \text{FN} & \text{TN}
\end{array}
$$

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

$$P[y = 1 | \hat{y} = 1] = \frac{TP}{TP + FP} \tag{2}$$

PR-Curve



F-Scores

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

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

F-Scores

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

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{(\beta^2 \times Precision + Recall)} \tag{4}$$

Example: Unemployment

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