Classification: Performance Metrics & Class Imbalance

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Recap
2. Confusion Matrix
3. ROC curve
4. Imbalanced Classification
   - Metrics

Agenda

1. Recap
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 - Unemployment

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

Agenda

1 Recap

2 Confusion Matrix

3 ROC curve

4 Imbalanced Classification
- Metrics

Confusion Matrix

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

Agenda

1 Recap

2 Confusion Matrix

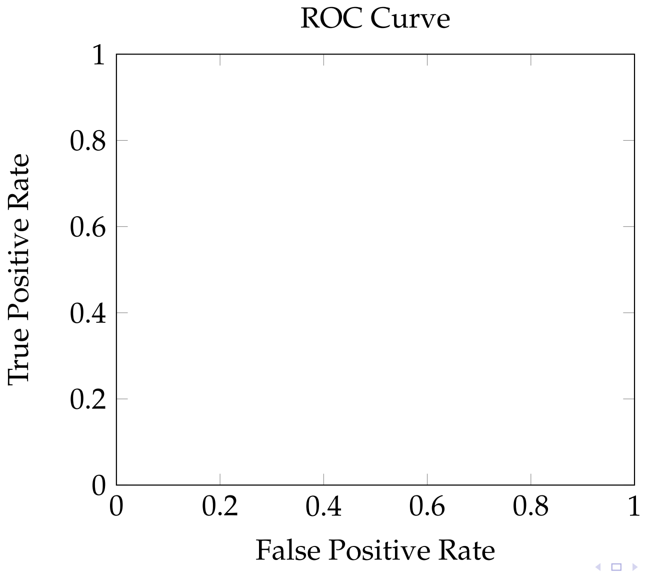
3 ROC curve

4 Imbalanced Classification
   - Metrics

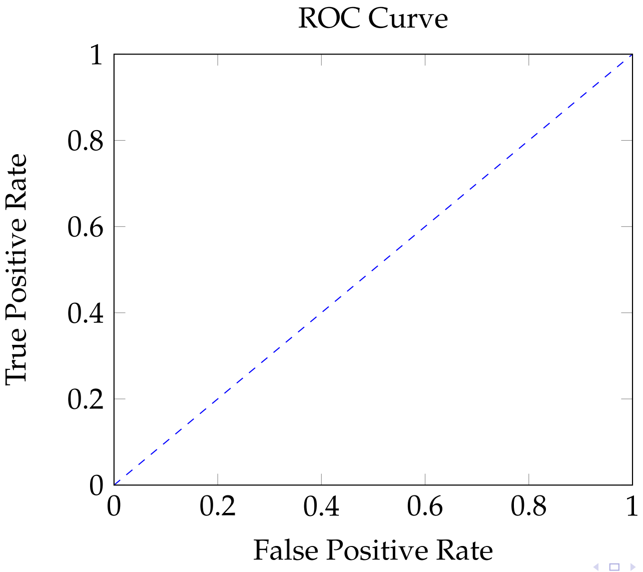
Trade-Off between Different Classification Thresholds

$$\hat{y}_i = 1[p_i \geq c]$$

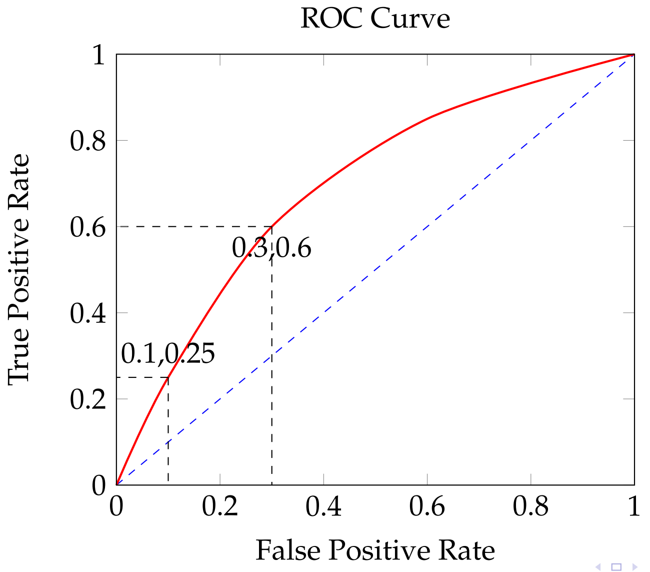
ROC Plot


ROC Curve
True Positive Rate
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
False Positive Rate

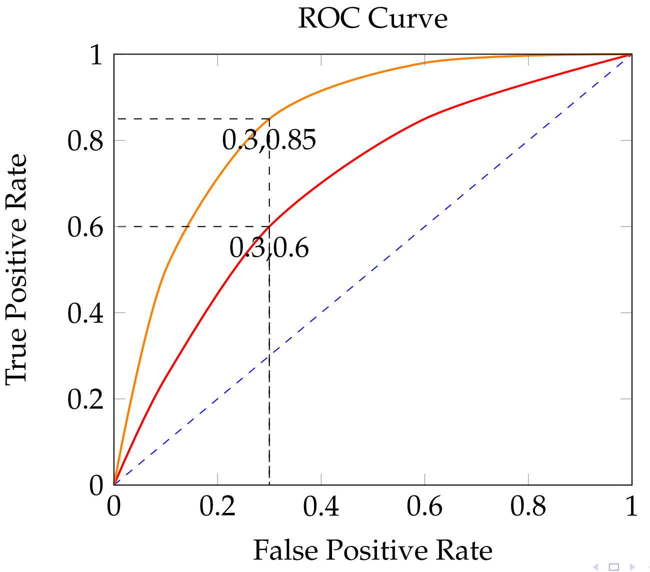
ROC Plot


ROC Curve
1
0.8
0.6
0.4
0.2
0
True Positive Rate
0
0.2
0.4
0.6
0.8
1
False Positive Rate

ROC Plot


ROC Curve
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1
0.3,0.6
0.1,0.25

ROC Plot


ROC Curve
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
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

1 Recap

2 Confusion Matrix

3 ROC curve

4 Imbalanced Classification
- Metrics

Imbalanced Classification: Motivation

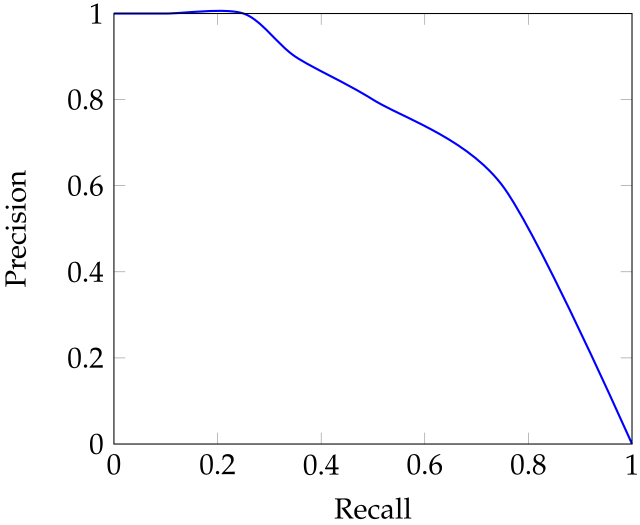
- Interest in one of the classes: Poor, Default, Unemployed, Fraud
- Imbalanced classes pose a challenge

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

PR-Curve



Imbalanced Classification: Solutions

- Model Tuning
- Alternative Cutoffs
- Weights
- Adjust Prior Probabilities
- Class rebalancing

Example: Unemployment

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