

# **Data Mining**

## **Classification: Alternative Techniques**

---

Imbalanced Class Problem

Introduction to Data Mining, 2<sup>nd</sup> Edition

by

Tan, Steinbach, Karpatne, Kumar

# Class Imbalance Problem

- Lots of classification problems where the classes are skewed (more records from one class than another)
  - Credit card fraud
  - Intrusion detection
  - Defective products in manufacturing assembly line

# Challenges

- Evaluation measures such as accuracy is not well-suited for imbalanced class
- Detecting the rare class is like finding needle in a haystack

# Confusion Matrix

- Confusion Matrix:

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a	b
	c	d

**a: TP (true positive)**

**b: FN (false negative)**

**c: FP (false positive)**

**d: TN (true negative)**

# Accuracy

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	a (TP)	b (FN)
	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$



# Problem with Accuracy

- Consider a 2-class problem  
(total number of test samples 10.000)
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

# Problem with Accuracy

- Consider a 2-class problem
  - Number of Class NO examples = 990
  - Number of Class YES examples = 10
- If a model predicts everything to be class NO, accuracy is  $990/1000 = 99\%$ 
  - This is misleading because the model does not detect any class YES example
  - Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS	
		Class=Yes      Class=No
	Class=Yes	a      b
	Class=No	c      d

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

Harmonic Mean of P and R



# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	1	9
	Class=No	0	990

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F - measure (F)} = \frac{2*0.1*1}{1+0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	40	10
	10	40

Precision ( $p$ ) = 0.8

Recall ( $r$ ) = 0.8

F - measure ( $F$ ) = 0.8

Accuracy = 0.8

# Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	10	40

Precision (p) = 0.8

Recall (r) = 0.8

F - measure (F) = 0.8

Accuracy = 0.8

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	1000	4000

Precision (p) = ~ 0.04

Recall (r) = 0.8

F - measure (F) = ~ 0.08

Accuracy = ~ 0.8

# Measures of Classification Performance

ACTUAL CLASS	PREDICTED CLASS		
		Yes	No
	Yes	TP	FN
	No	FP	TN

$\alpha$  is the probability that we reject the null hypothesis when it is true. This is a Type I error or a false positive (FP).

$\beta$  is the probability that we accept the null hypothesis when it is false. This is a Type II error or a false negative (FN).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$ErrorRate = 1 - accuracy$$

$$Precision = Positive Predictive Value = \frac{TP}{TP + FP}$$

$$Recall = Sensitivity = TP Rate = \frac{TP}{TP + FN}$$

$$Specificity = TN Rate = \frac{TN}{TN + FP}$$

$$FP Rate = \alpha = \frac{FP}{TN + FP} = 1 - specificity$$

$$FN Rate = \beta = \frac{FN}{FN + TP} = 1 - sensitivity$$

$$Power = sensitivity = 1 - \beta$$



# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	40	10
	10	40

Precision (p) = 0.8

TPR = Recall (r) = 0.8

FPR = 0.2

F - measure (F) = 0.8

Accuracy = 0.8

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	40	10
	1000	4000

Precision (p) = ~ 0.04

TPR = Recall (r) = 0.8

FPR = 0.2

F - measure (F) = ~ 0.08

Accuracy = ~ 0.8

# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	10	40
	Class=No	10	40

Precision (p) = 0.5

TPR = Recall (r) = 0.2

FPR = 0.2

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	25	25
	Class=No	25	25

Precision (p) = 0.5

TPR = Recall (r) = 0.5

FPR = 0.5

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	40	10

Precision (p) = 0.5

TPR = Recall (r) = 0.8

FPR = 0.8

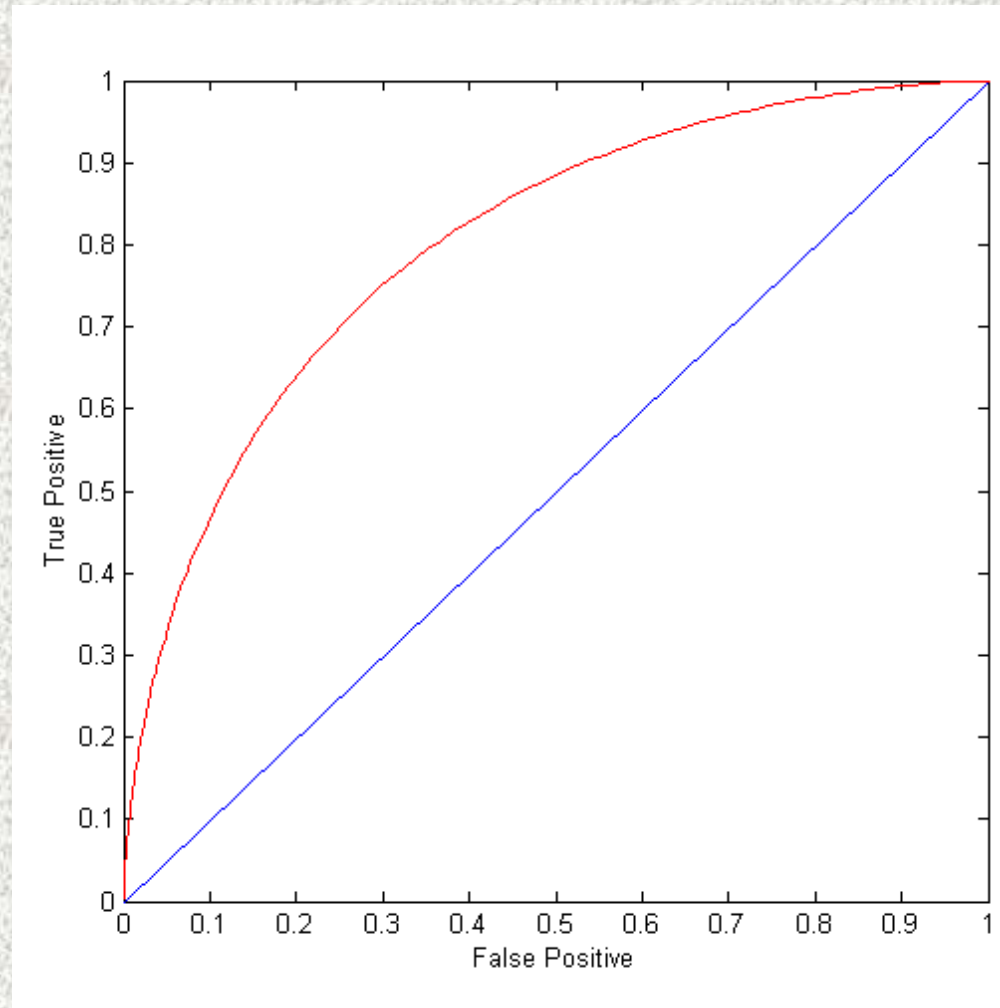
# ROC (Receiver Operating Characteristic)

- A graphical approach for displaying trade-off between detection rate and false alarm rate
- Developed in 1950s for signal detection theory to analyze noisy signals
- ROC curve plots TPR against FPR
  - Performance of a model represented as a point in an ROC curve
  - Changing the threshold parameter of classifier changes the location of the point

# ROC Curve

(TPR,FPR):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal
- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - ◆ prediction is opposite of the true class





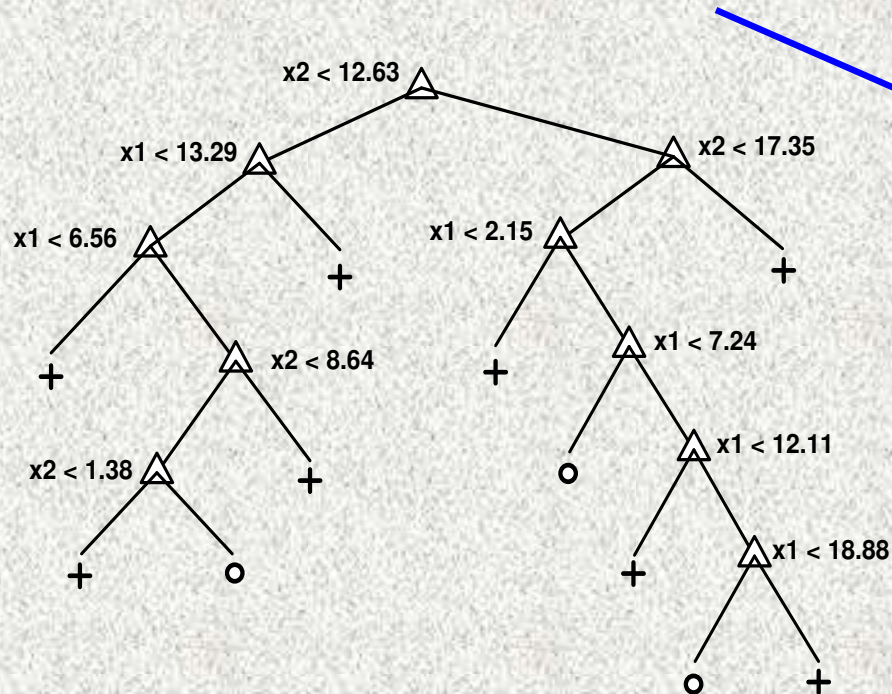
# ROC (Receiver Operating Characteristic)

- To draw ROC curve, classifier must produce continuous-valued output
  - Outputs are used to rank test records, from the most likely positive class record to the least likely positive class record
- Many classifiers produce only discrete outputs (i.e., predicted class)
  - How to get continuous-valued outputs?
    - ◆ Decision trees, rule-based classifiers, neural networks, Bayesian classifiers, k-nearest neighbors, SVM

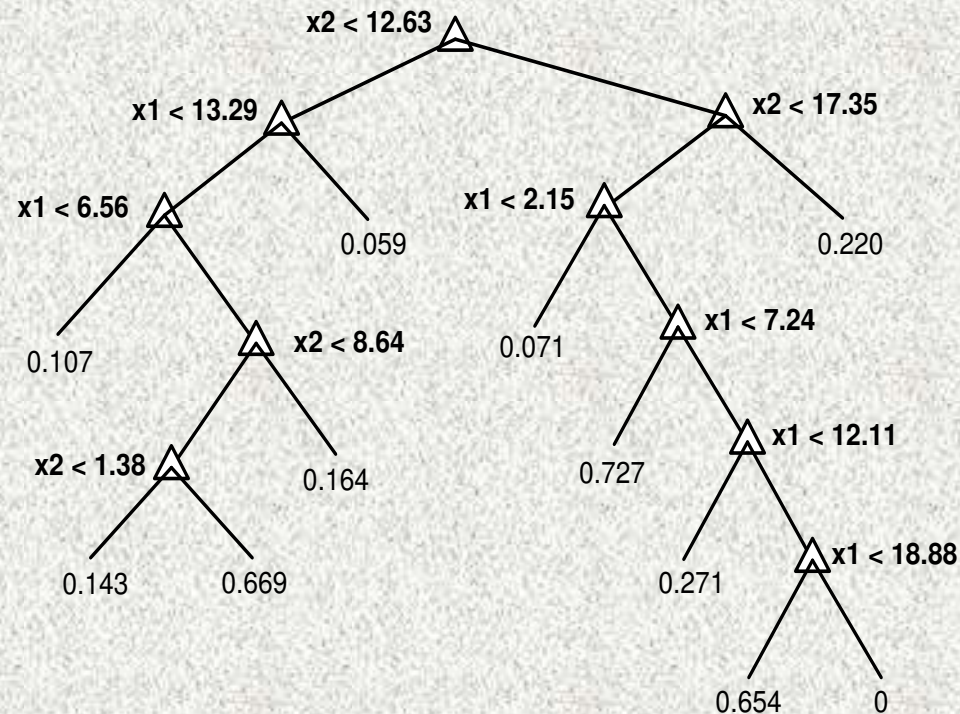


# Example: Decision Trees

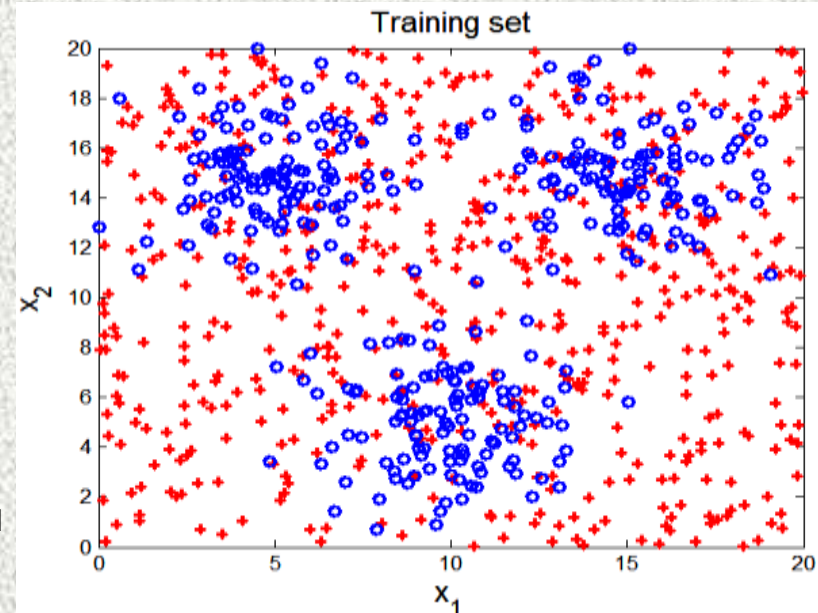
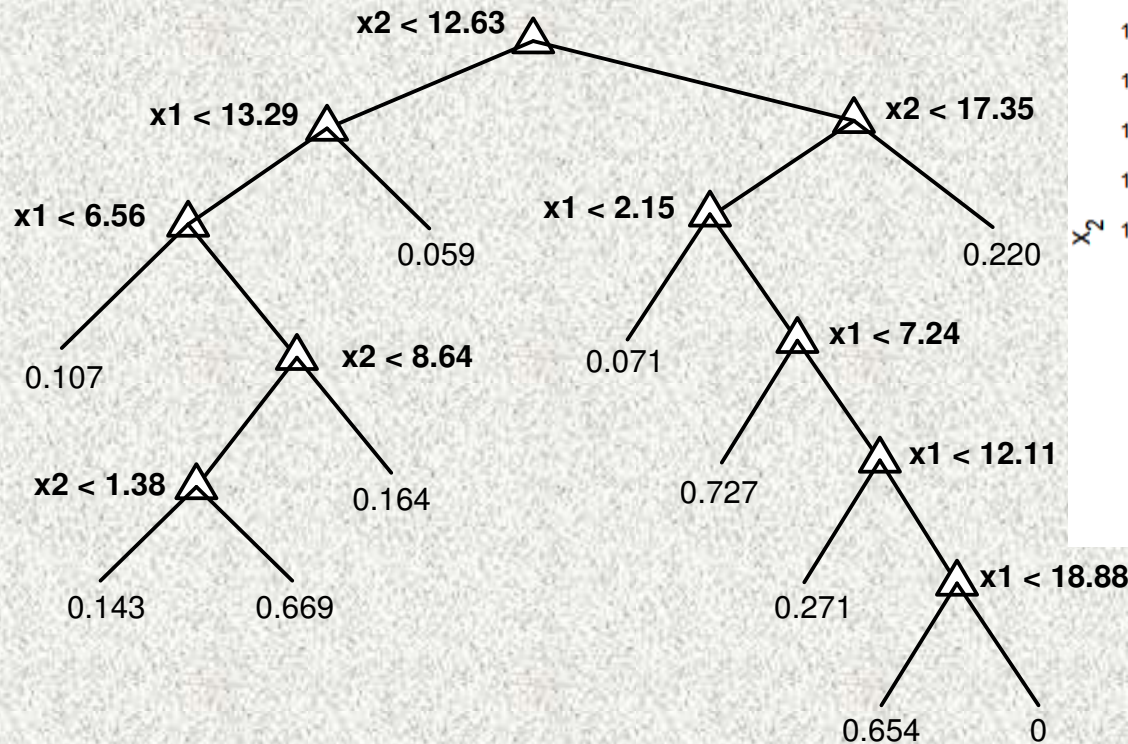
## Decision Tree



## Continuous-valued outputs



# ROC Curve Example

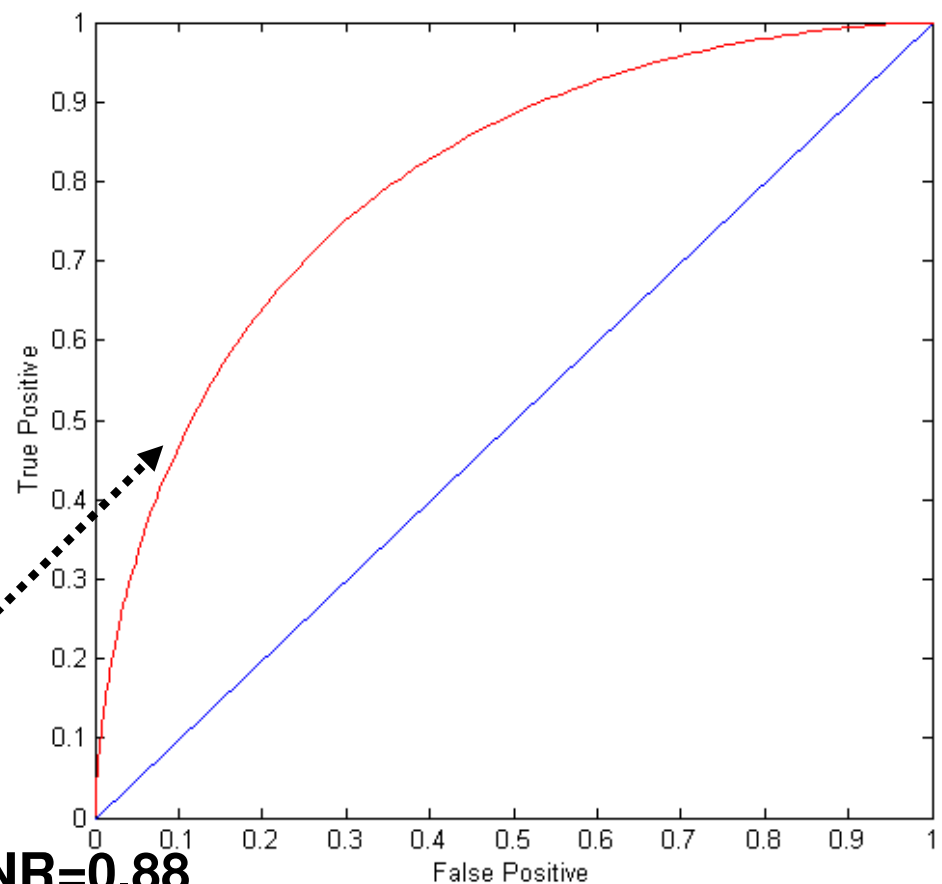
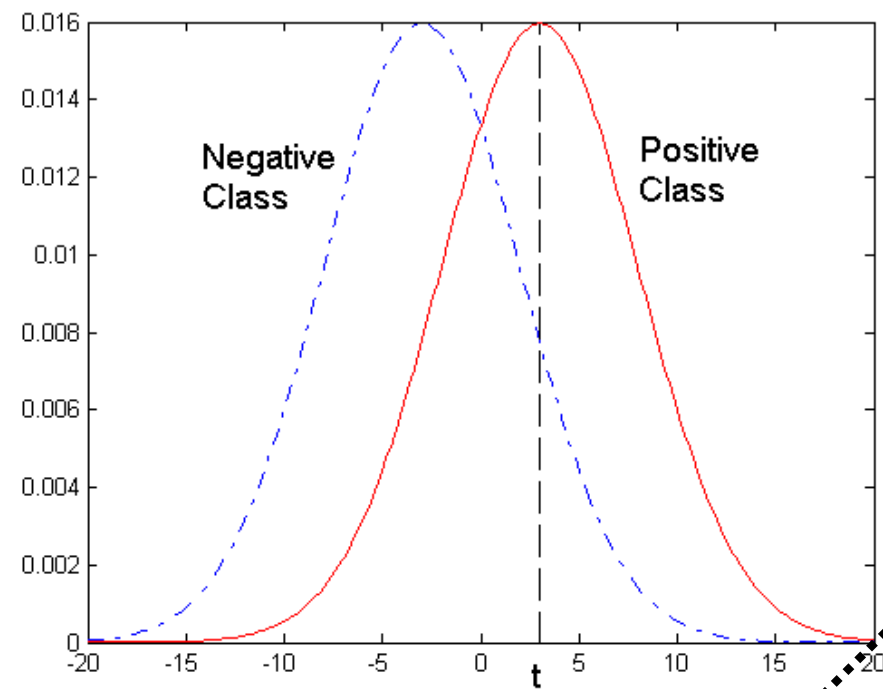


$\alpha = 0.3$		Predicted Class	
		Class 0	Class +
Actual Class	Class 0	645	209
	Class +	298	948

$\alpha = 0.7$		Predicted Class	
		Class 0	Class +
Actual Class	Class 0	181	673
	Class +	78	1168

# ROC Curve Example

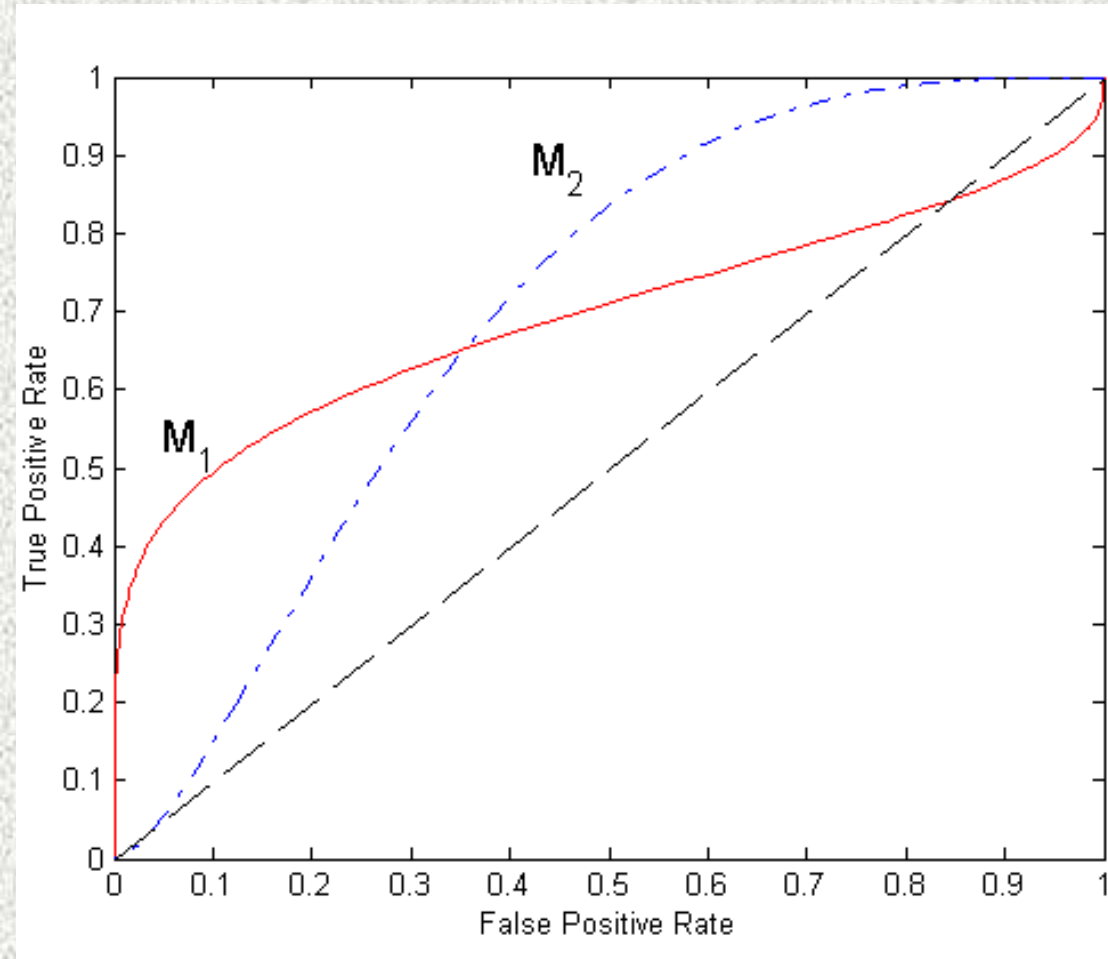
- 1-dimensional data set containing 2 classes (positive and negative)
- Any points located at  $x > t$  is classified as positive



At threshold  $t$ :

**TPR=0.5, FNR=0.5, FPR=0.12, TNR=0.88**

# Using ROC for Model Comparison



- No model consistently outperform the other
  - $M_1$  is better for small FPR
  - $M_2$  is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5



# How to Construct an ROC curve

Instance	Score	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use a classifier that produces a continuous-valued score for each instance
  - The more likely it is for the instance to be in the + class, the higher the score
- Sort the instances in decreasing order according to the score
- Apply a threshold at each unique value of the score
- Count the number of TP, FP, TN, FN at each threshold
  - $TPR = TP / (TP + FN)$
  - $FPR = FP / (FP + TN)$



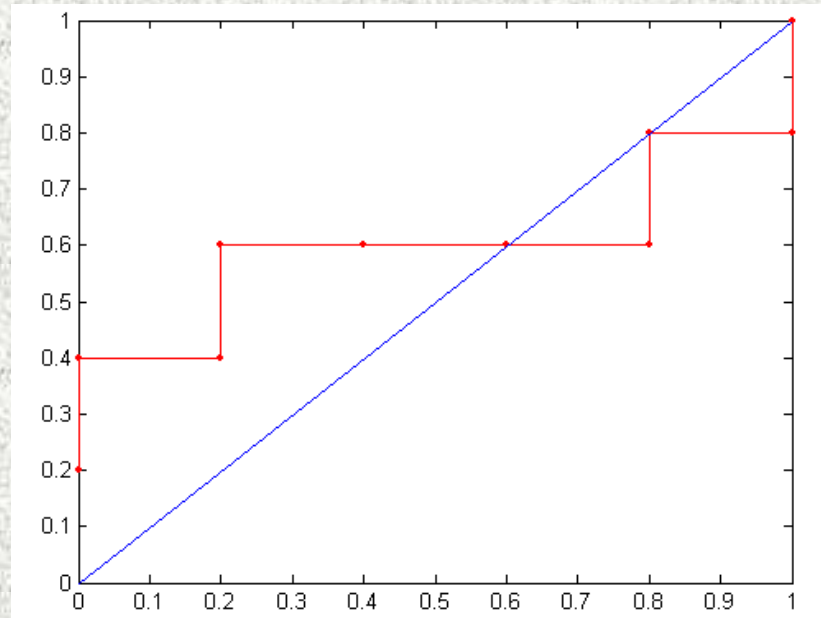
# How to construct an ROC curve

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

→

→

ROC Curve:



# Handling Class Imbalanced Problem

- Class-based ordering (e.g. RIPPER)
  - Rules for rare class have higher priority
- Cost-sensitive classification
  - Misclassifying rare class as majority class is more expensive than misclassifying majority as rare class
- Sampling-based approaches

# Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	f(Yes, Yes)	f(Yes, No)
	Class=No	f(No, Yes)	f(No, No)

$C(i,j)$ : Cost of misclassifying class  $i$  example as class  $j$

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	$C(i, j)$	Class=Yes	Class=No
	Class=Yes	$C(\text{Yes}, \text{Yes})$	$C(\text{Yes}, \text{No})$
	Class=No	$C(\text{No}, \text{Yes})$	$C(\text{No}, \text{No})$

$$\text{Cost} = \sum C(i, j) \times f(i, j)$$

# Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i,j)	+	-
	+	-1	100
	-	1	0

Model $M_1$	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model $M_2$	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

# Cost Sensitive Classification

- Example: Bayesian classifier
  - Given a test record  $x$ :
    - ◆ Compute  $p(i|x)$  for each class  $i$
    - ◆ Decision rule: classify node as class  $k$  if

$$k = \arg \max_i p(i | x)$$

- For 2-class, classify  $x$  as  $+$  if  $p(+|x) > p(-|x)$ 
  - ◆ This decision rule implicitly assumes that  $C(+|+) = C(-|-) = 0$  and  $C(+|-) = C(-|+)$



# Cost Sensitive Classification

- General decision rule:
  - Classify test record  $x$  as class  $k$  if

$$k = \arg \min_j \sum_i p(i | x) \times C(i, j)$$

- 2-class:
  - $\text{Cost}(+) = p(+|x) C(+,+) + p(-|x) C(-,+)$
  - $\text{Cost}(-) = p(+|x) C(+,-) + p(-|x) C(-,-)$
  - Decision rule: classify  $x$  as  $+$  if  $\text{Cost}(+) < \text{Cost}(-)$

- ◆ if  $C(+,+) = C(-,-) = 0$ :

$$p(+ | x) > \frac{C(-,+)}{C(-,+) + C(+,-)}$$

# Sampling-based Approaches

- Modify the distribution of training data so that rare class is well-represented in training set
  - Undersample the majority class
  - Oversample the rare class
- Advantages and disadvantages