

# Data Mining

COIS 4400H / AMOD 5440H

## Classification: Model Evaluation



## Metrics for Performance Evaluation

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)

## Metrics for Performance Evaluation

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

Most widely-used metric:

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

## Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(ij)	+	-
	+	-1	100
	-	1	0

Model M <sub>1</sub>	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%  
Cost = 3910

Model M <sub>2</sub>	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 90%  
Cost = 4255

## Cost-Sensitive Measures

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

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

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

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

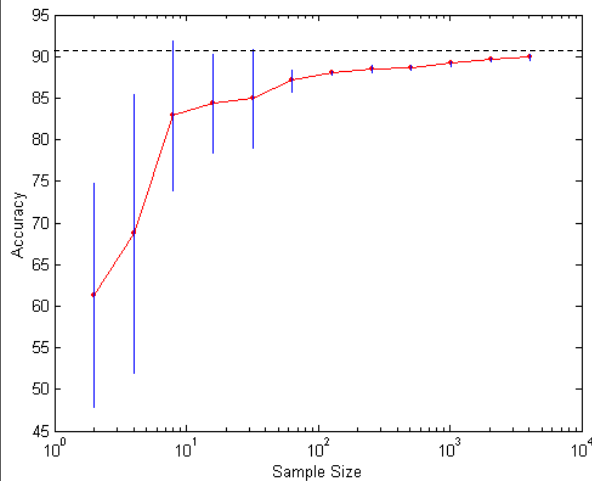
## Methods for Performance Evaluation

How to obtain a reliable estimate of performance?

Performance of a model may depend on other factors besides the learning algorithm:

- Class distribution
- Cost of misclassification
- Size of training and test sets

## Learning Curve



- Learning curve shows how accuracy changes with varying sample size
- Requires a sampling schedule for creating learning curve:

Effect of small sample size:

- Bias in the estimate
- Variance of estimate

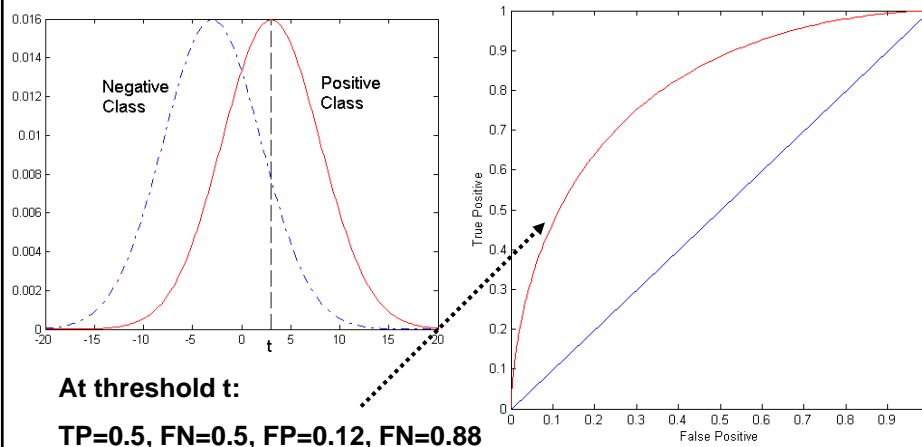
## Methods of Estimation

- Reserve 2/3 for training and 1/3 for testing
- Repeated holdout
- Cross validation: Partition data into  $k$  disjoint subsets, train on  $k-1$  partitions, test on the remaining one
- Leave-one-out:  $k=n$
- Sampling with replacement

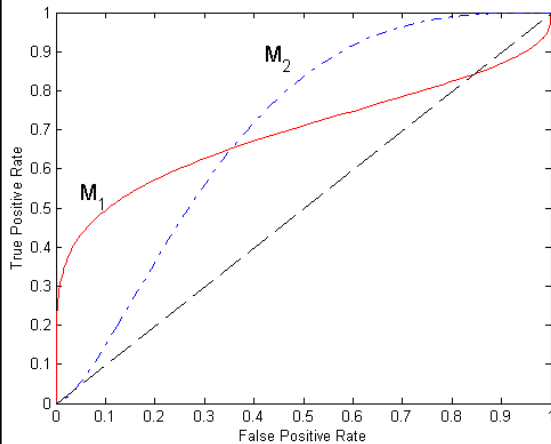
# Receiver Operating Characteristic (ROC)

- Developed in 1950s for signal detection theory to analyze noisy signals
- Characterizes the trade-off between positive hits and false alarms
- Performance of each classifier represented as a point on the ROC curve
- Changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point

## ROC Curve



## Using ROC for Model Comparison



- No model consistently outperforms 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

Class	+	-	+	-	-	-	+	-	+	+	
Threshold $\geq$	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:

