بسمه تعالی

Data Mining: Model Evaluation

Dr O. Pournik ***MD, MPH, MSc, PhD*** **pournik@gmail.com**

Model Evaluation

- ***Metrics for Performance Evaluation***
  - ***How to evaluate the performance of a model?***
- ***Methods for Performance Evaluation***
  - ***How to obtain reliable estimates?***
- ***Methods for Model Comparison***
  - ***How to compare the relative performance among competing models?***

Model Evaluation

- ***Metrics for Performance Evaluation***
  - ***How to evaluate the performance of a model?***
- ***Methods for Performance Evaluation***
  - ***How to obtain reliable estimates?***
- ***Methods for Model Comparison***
  - ***How to compare the relative performance among competing models?***

Metrics for Performance Evaluation

- ***Focus on the predictive capability of a model***
  - ***Rather than how fast it takes to classify or build models, scalability, etc.***
- ***Confusion Matrix***

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Most widely-used metric:

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

Limitation of Accuracy

- *Consider a 2-class problem*
  - *Number of Class 0 examples = 9990*
  - *Number of Class 1 examples = 10*

- *If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %*
  - *Accuracy is misleading because model does not detect any class 1 example*

Cost Matrix

- ***C(i|j): Cost of misclassifying class j example as class i***

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

Cost vs Accuracy

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

Cost	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	p	q
	Class=No	q	p

Accuracy is proportional to cost if

1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

$N = a + b + c + d$

$\text{Accuracy} = (a + d)/N$

$\text{Cost} = p\,(a + d) + q\,(b + c)$

$= p\,(a + d) + q\,(N - a - d)$

$= q\,N - (q - p)(a + d)$

$= N\,[q - (q-p) \times \text{Accuracy}]$

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)

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

Model Evaluation

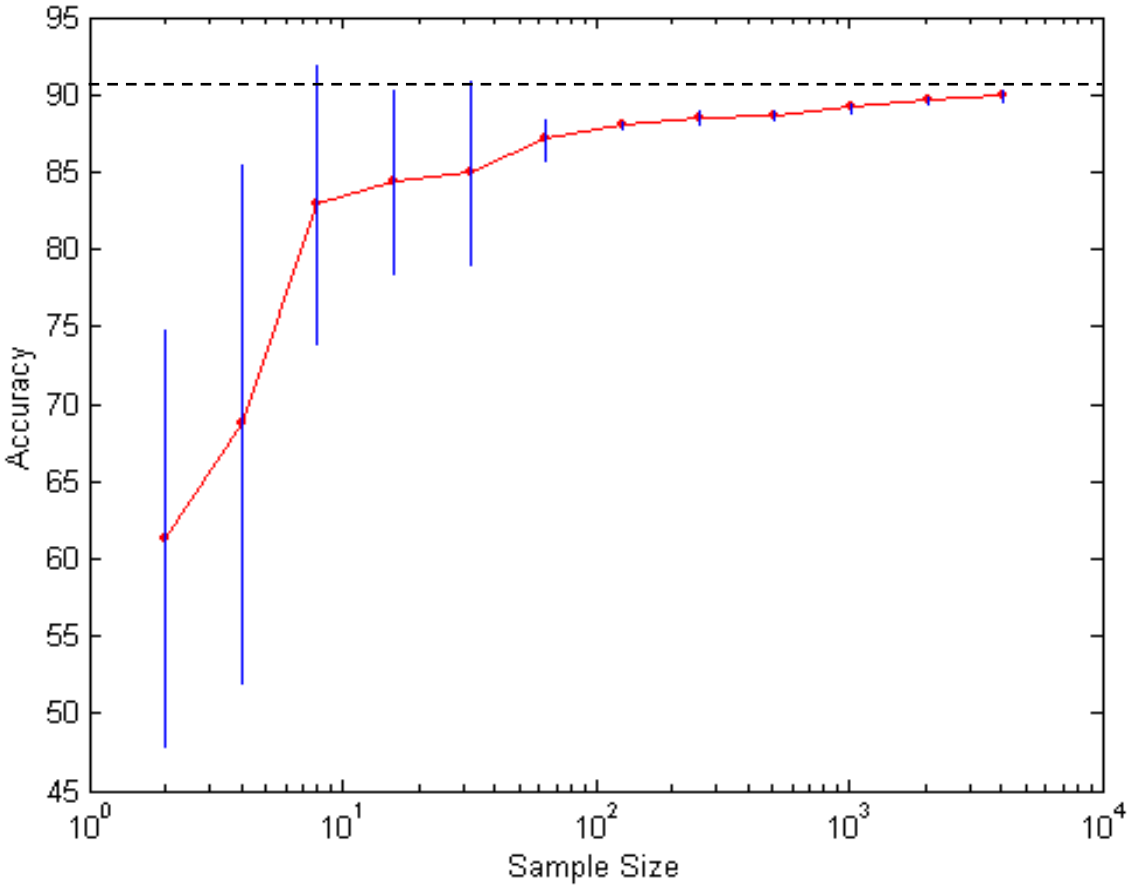
- ***Metrics for Performance Evaluation***
  - ***How to evaluate the performance of a model?***
- ***Methods for Performance Evaluation***
  - ***How to obtain reliable estimates?***
- ***Methods for Model Comparison***
  - ***How to compare the relative performance among competing models?***

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:***
  - ***Arithmetic sampling (Langley, et al)***
  - ***Geometric sampling (Provost et al)***

Effect of small sample size:

- ***Bias in the estimate***
- ***Variance of estimate***

Methods of Estimation

Sampling strategies

- *Holdout*
  - *Reserve 2/3 for training and 1/3 for testing*
- *Random subsampling*
  - *Repeated holdout*
- *Cross validation*
  - *Partition data into k disjoint subsets*
  - *k-fold: train on k-1 partitions, test on the remaining one*
  - *Leave-one-out: k=n*
- *Stratified sampling*
  - *oversampling vs. under sampling*
- *Bootstrap*
  - *Sampling with replacement*

Model Evaluation

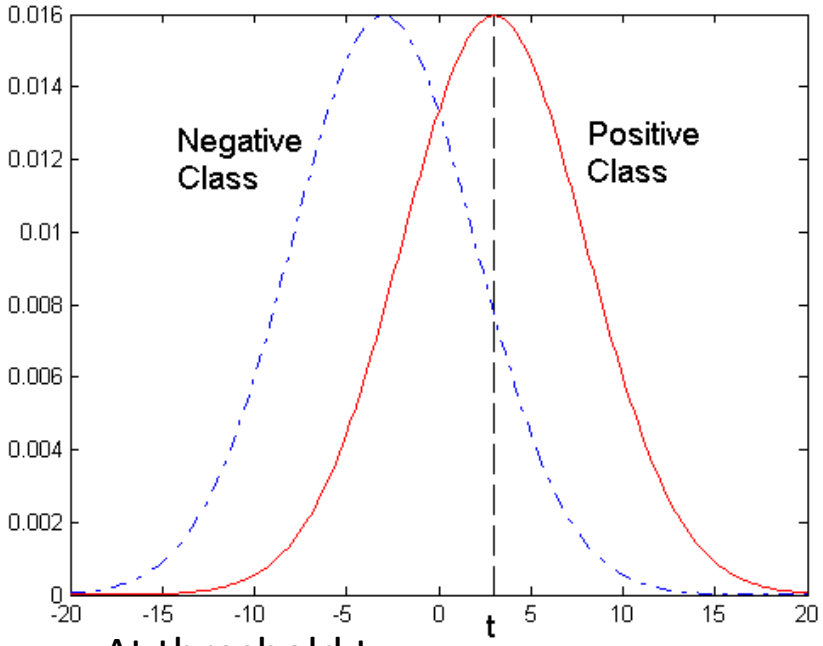
- ***Metrics for Performance Evaluation***
  - ***How to evaluate the performance of a model?***
- ***Methods for Performance Evaluation***
  - ***How to obtain reliable estimates?***
- ***Methods for Model Comparison***
  - ***How to compare the relative performance among competing models?***

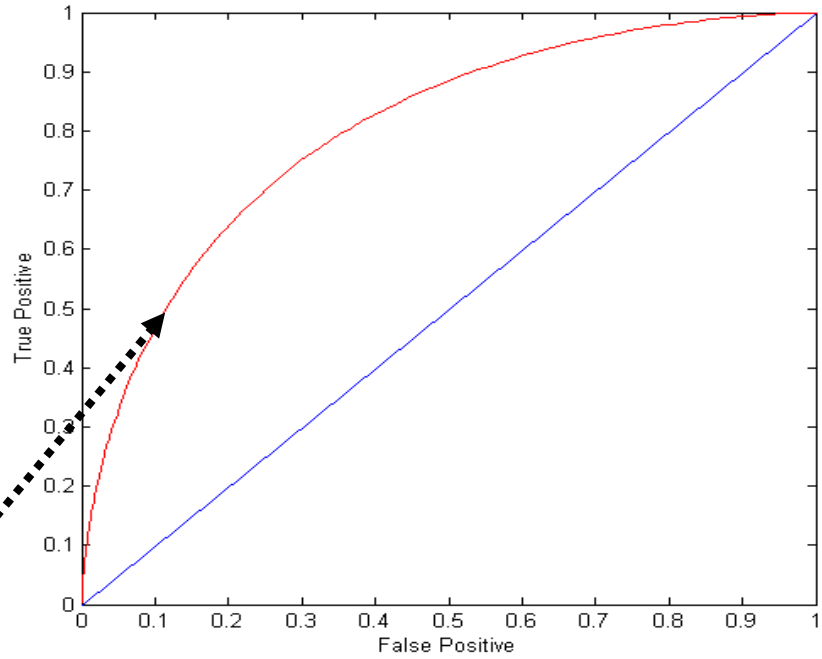
ROC (Receiver Operating Characteristic)

- *Developed in 1950s for signal detection theory to analyze noisy signals*
  - *Characterize the trade-off between positive hits and false alarms*
- *ROC curve plots TP (on the y-axis) against FP (on the x-axis)*
- *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

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





At threshold t:

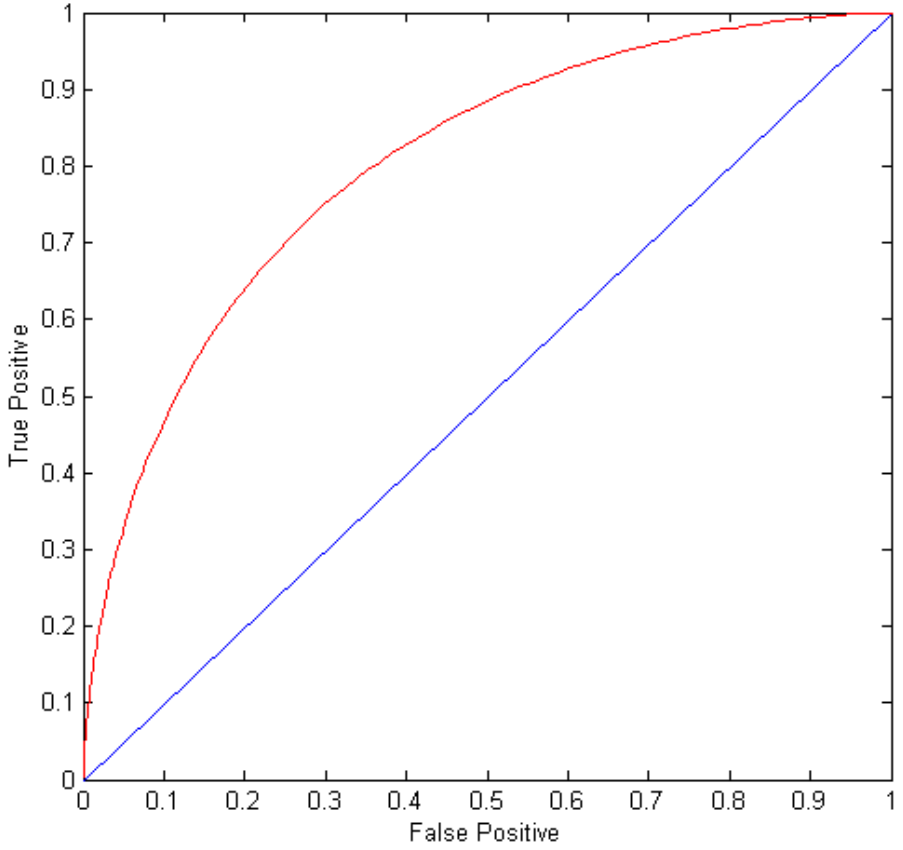
TP=0.5, FN=0.5, FP=0.12, FN=0.88

ROC Curve

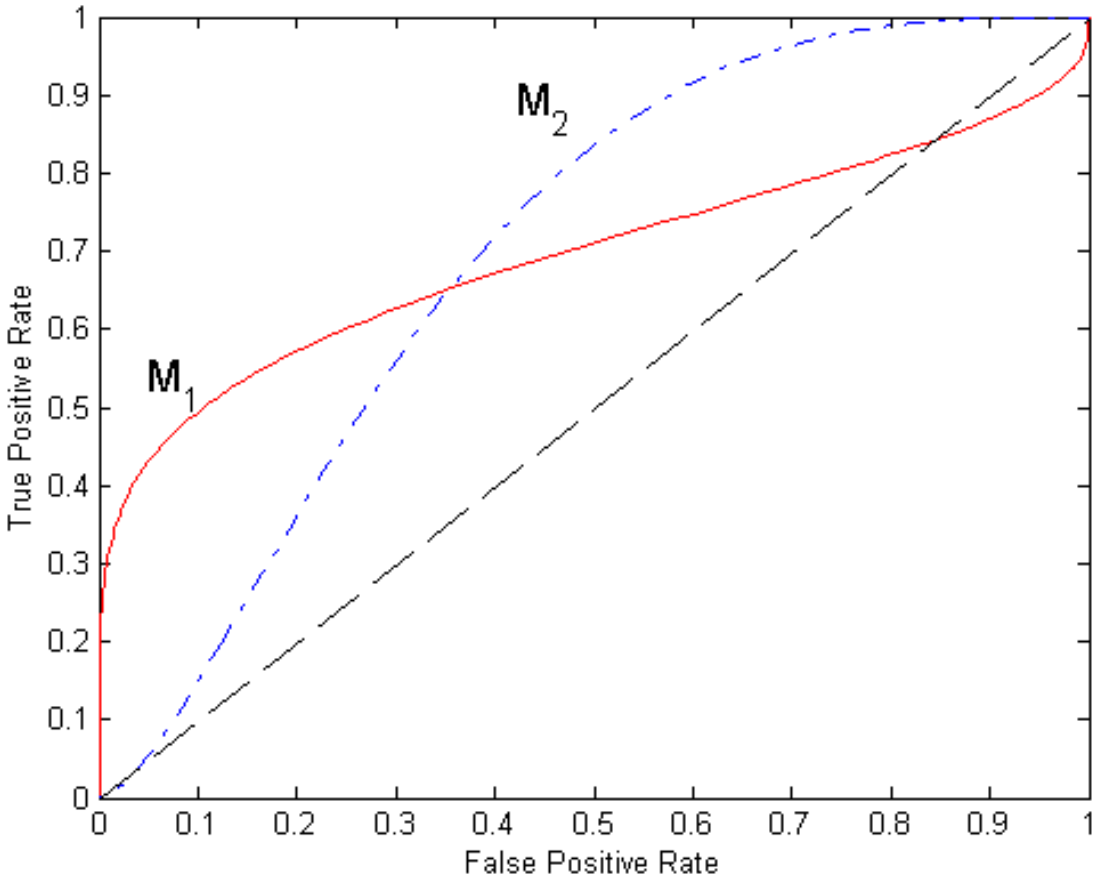
(TP,FP):

- (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



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

Any Questions?