Outline

- Evaluation

Outline

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

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Confusion Matrix

	PREDICTED CLASS		
ACTUAL CLASS		Cat	Dog
	Cat	a	b
	Dog	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Measure

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

All samples of "yes"

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

Measure

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

All samples predicted as yes”

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

Measure

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

All samples of "yes"

All samples predicted as "yes"

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

$$\frac{2a}{2a+b+c}$$

Accuracy

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

- 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

	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)

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

Computing Cost of Classification

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

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

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

Accuracy = 80%

Cost = 3910

Accuracy = 90%

Cost = 4255

Weighted Accuracy

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

Quiz: evaluation

We developed a Random Decision algorithm to classify an email to be spam and non-spam.
Data: 100 spam emails, 200 nonspam emails.
With the following results, please calculate: i) recall; ii) precision; iii) F- measure; and iv) accuracy

	Prediction		
Ground-truth lael		Spam	Non-spam
	spam	70	30
	Non-spam	20	180

Quiz: accuracy

Animal classification problem. Given the following confusion matrix, can you calculate the accuracy?

	Dog	Cat	monkey
Dog	80	10	5
Cat	30	60	10
monkey	10	8	82

Outline

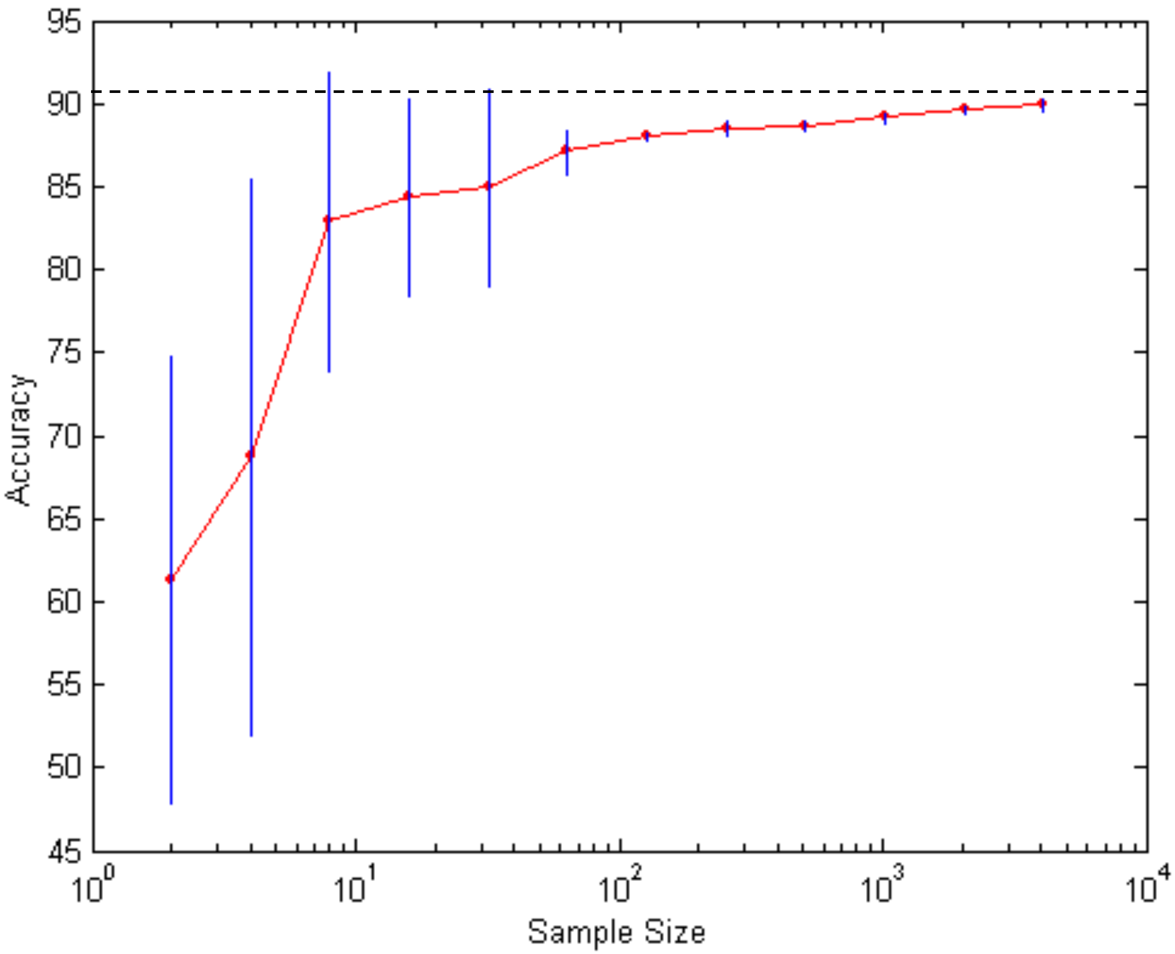
- Metrics for Performance Evaluation
  – How to evaluate the performance of a model?
- Methods for Performance Evaluation
  – How to obtain reliable estimates?
- Methods for Model Comparison
  – How to compare the relative performance among competing models?

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

1. Learning Curve



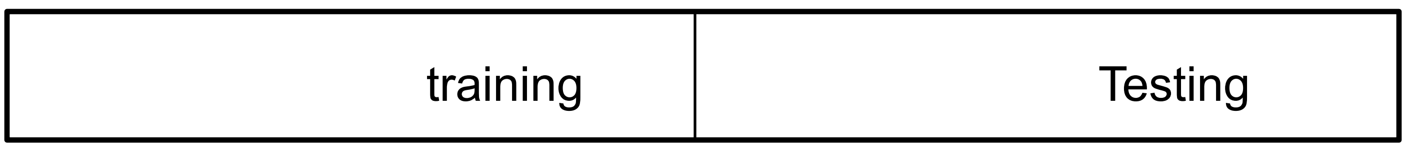
- Learning curve shows how accuracy changes with varying sample size
- Requires a sampling schedule for creating learning curve:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)

Effect of small sample size:

- Bias in the estimate
- Variance of estimate

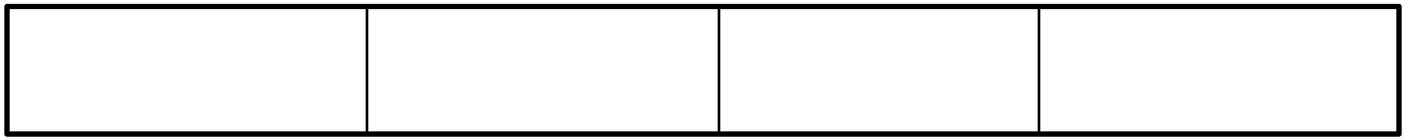
2. Holdout

Reserve 2/3 for training and 1/3 for testing



Reserve ½ for training and ½ for testing

3. Cross validation



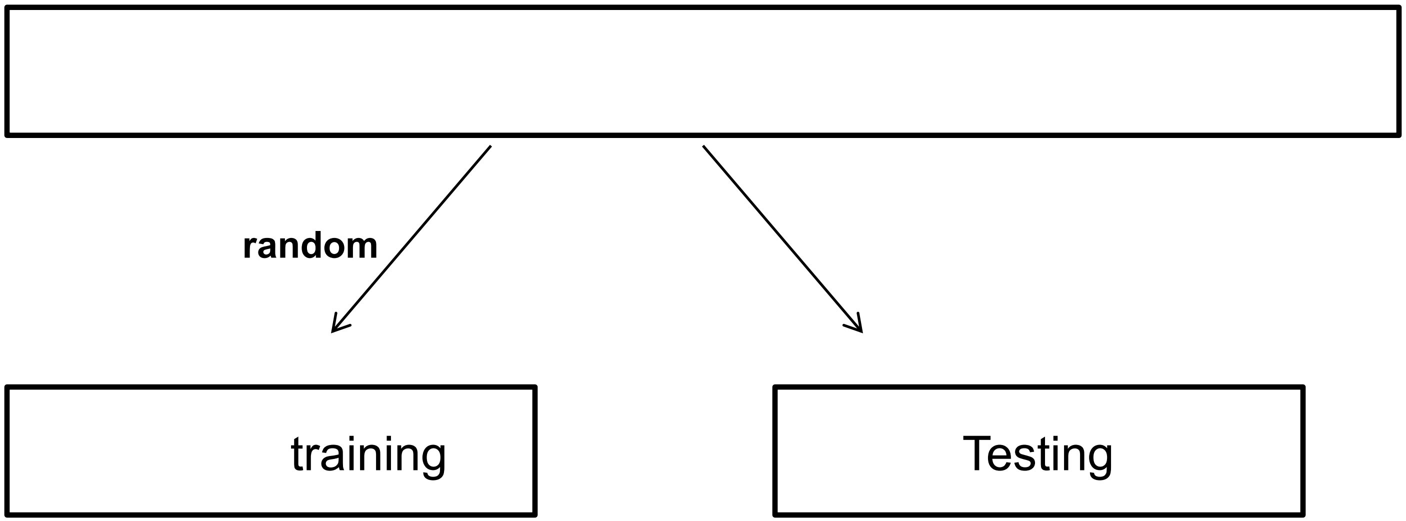
Partition data into k disjoint subsets

k-fold: train on k-1 partitions, test on the remaining one;

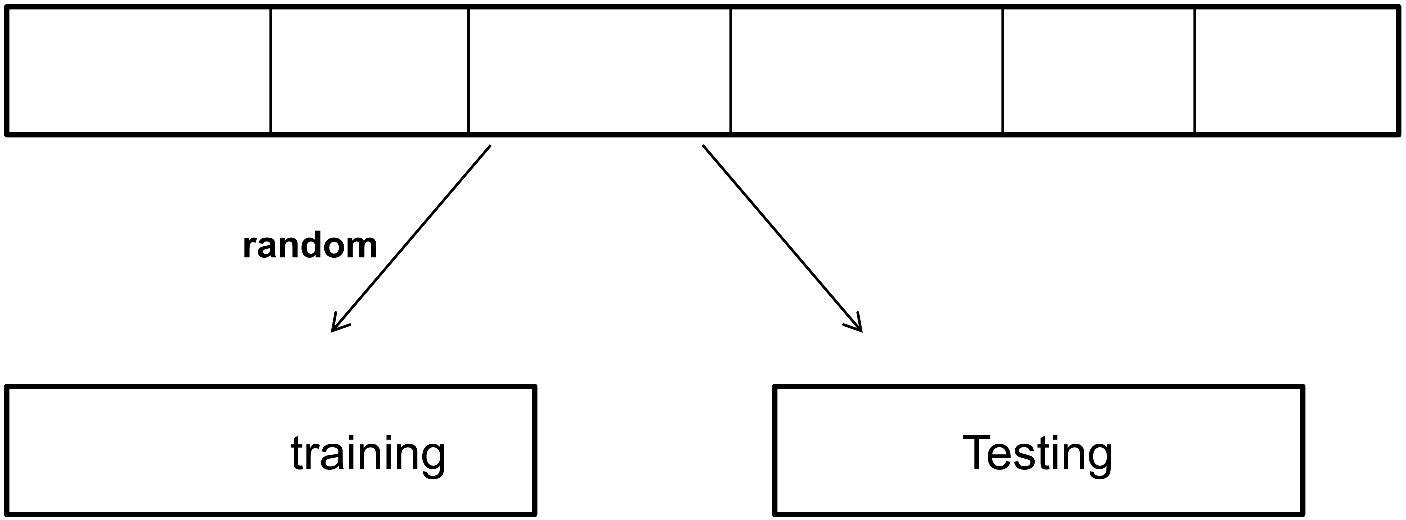
Leave-one-out

4. Sampling: bootstrap

Sampling with replacement



5. Stratified sampling



oversampling vs undersampling

Recap: evaluation methods

1. Learning curve
2. Hold-out
3. Cross-validation
4. Bootstrap
5. Stratified sampling

Outline

- 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)

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

ROC: basic idea

Instance	P(+|A)	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 classifier that produces posterior probability for each test instance P(+|A)
- Sort the instances according to P(+|A) in decreasing order
- What's the appropriate threshold to pick up the positive samples?

Conti.

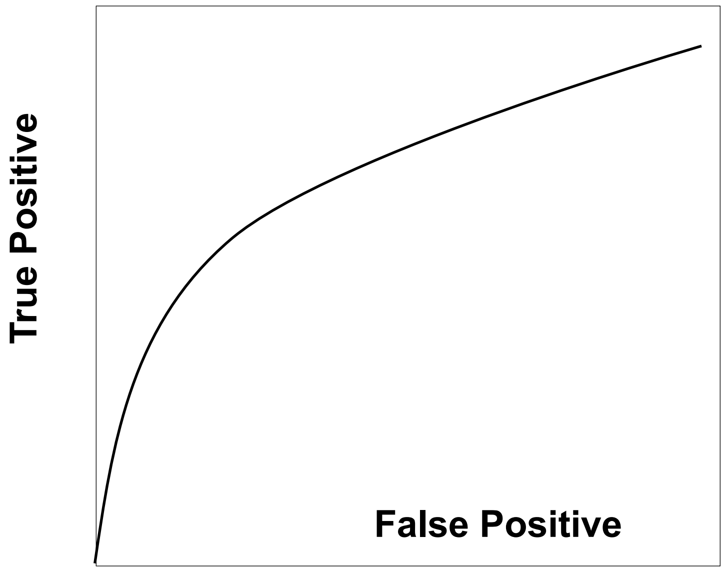
Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

•For each unique value, consider it as a threshold

- Count the number of TP, FP, TN, FN.
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

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

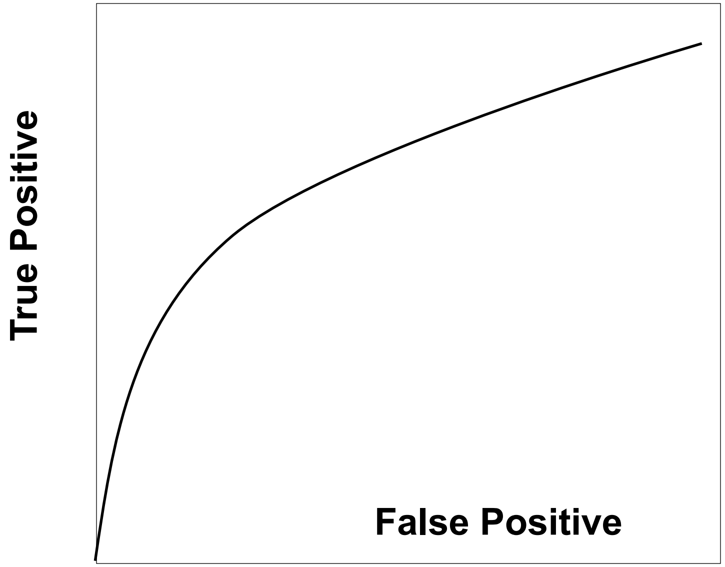
Conti



ROC (Receiver Operating Characteristic)

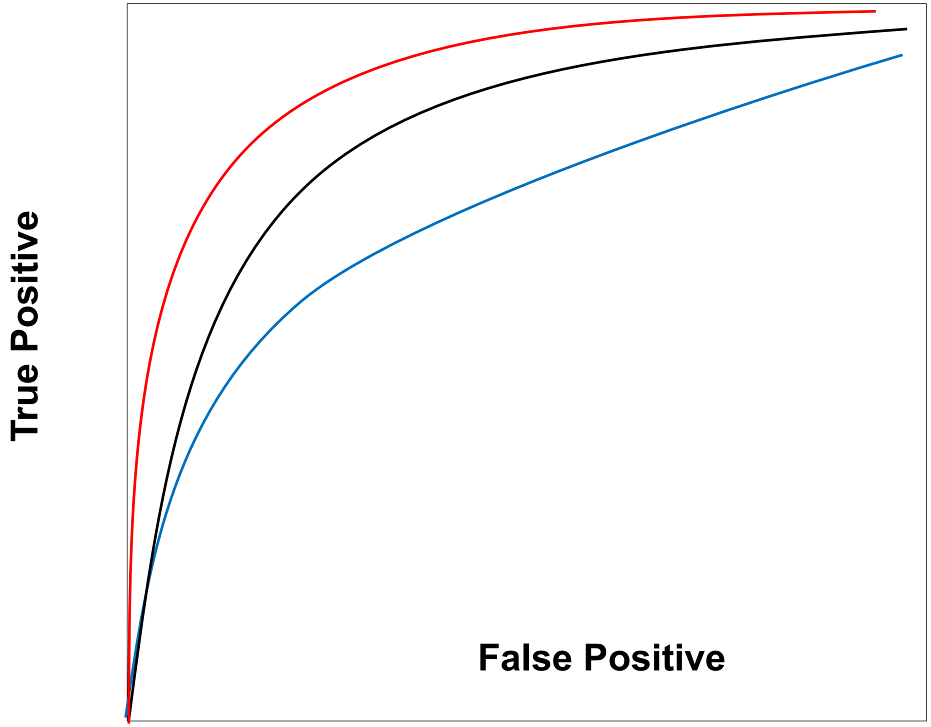
ROC

- 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



Quiz

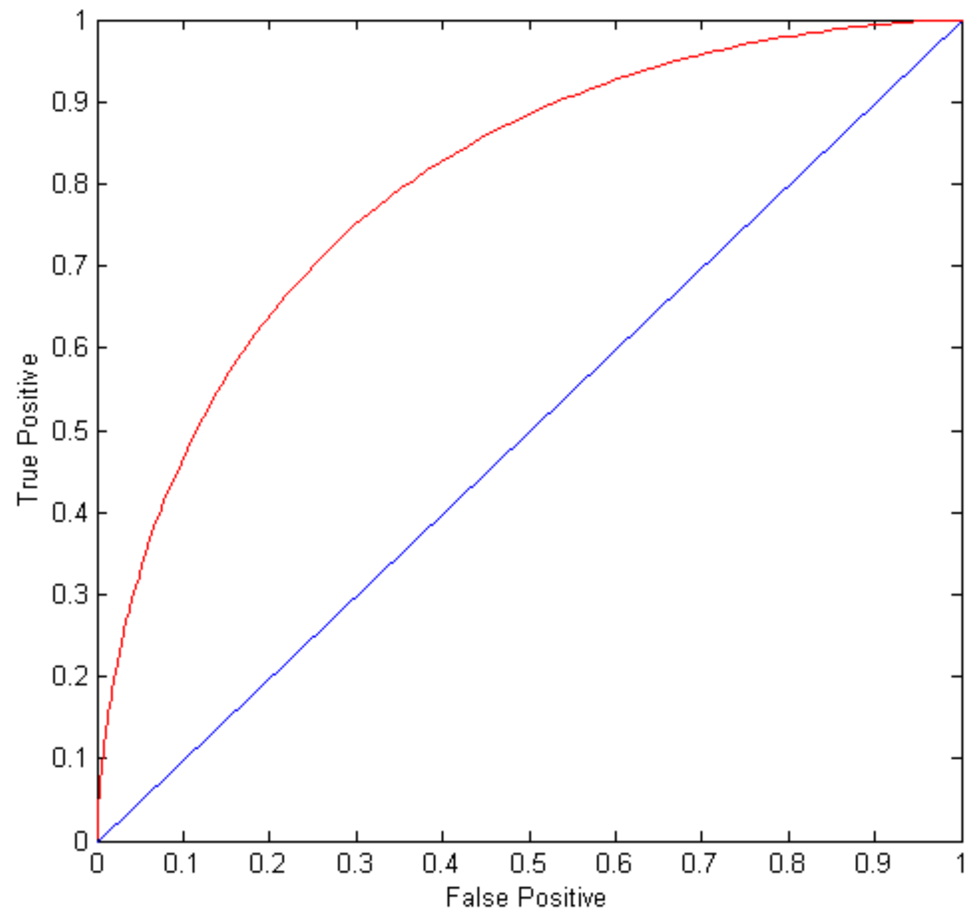
For spam email classification problem, suppose you have three different models with the following ROCs. Which algorithm is the best?



Understanding ROC Curve

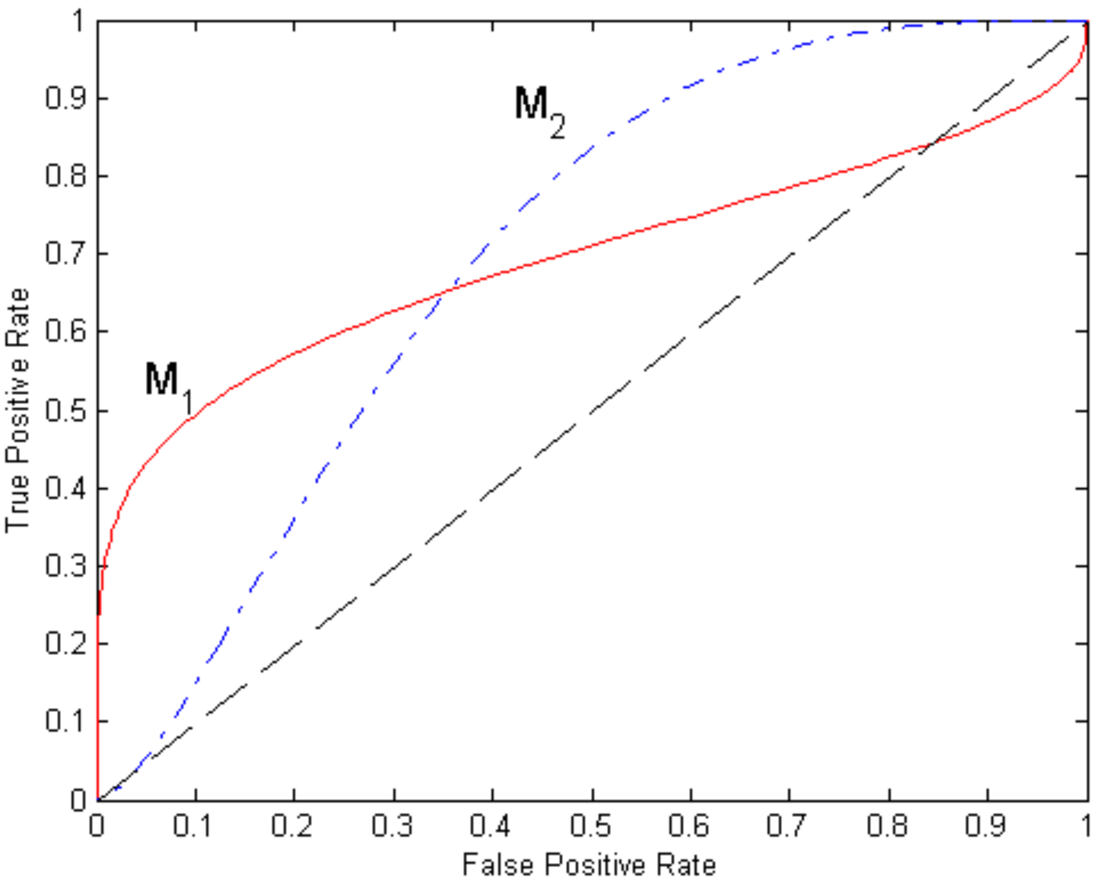
(True Positive,False Positive)

- (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

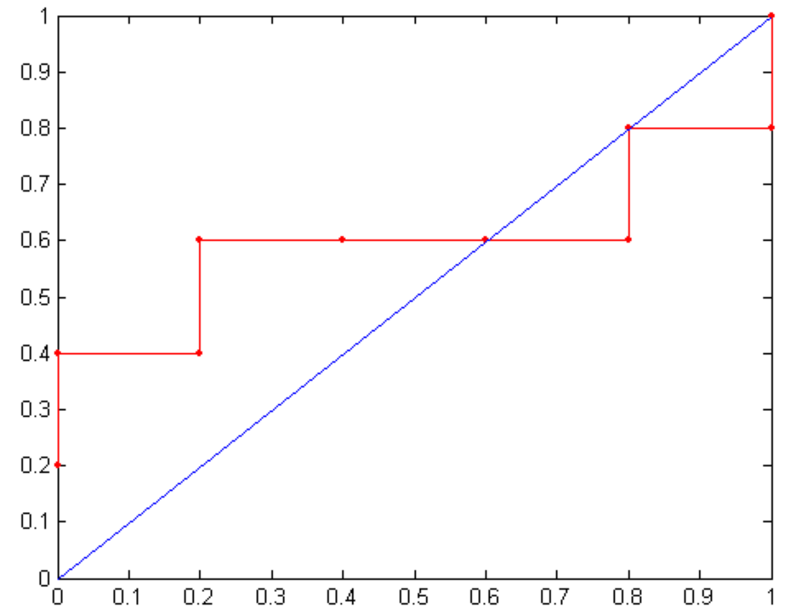


- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

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:



Summery of today

- Metrics for Performance Evaluation
  - Confusion Matrix; recall; precision; F-measure; accuracy
- Methods for Performance Evaluation
  - Learning curve; Hold-out; Cross-validation; Bootstrap; Stratified sampling
- Methods for Model Comparison
  - ROC curves