Valutazione di modelli

venerdì, 03 Novembre 2006

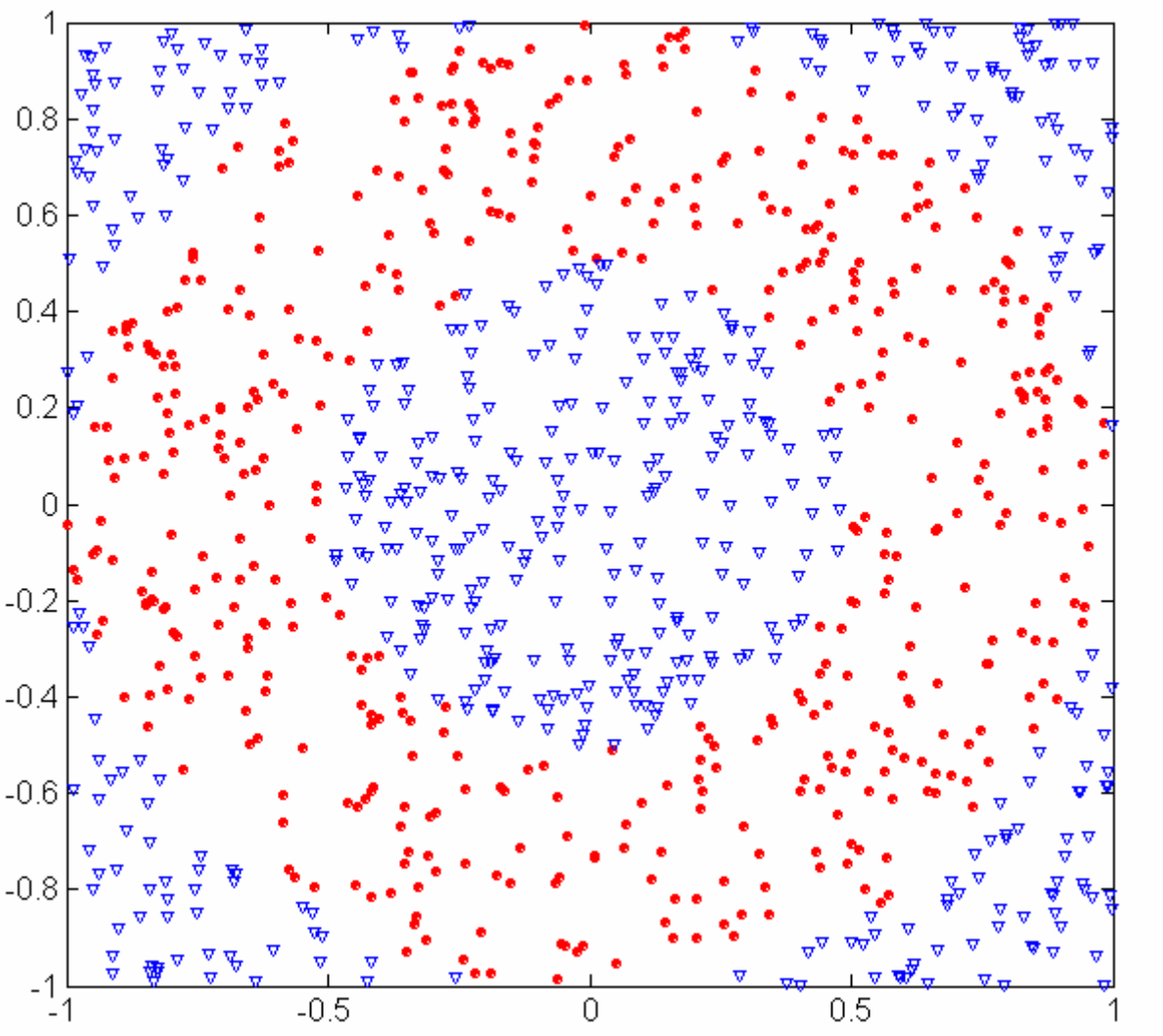
Giuseppe Manco

References:

Chapter 3, Mitchell

Chapters 4.5, 5.7, Tan, Steinbach, Kumar

Underfitting, Overfitting



500 cerchi, 500 triangoli.

Cerchi:

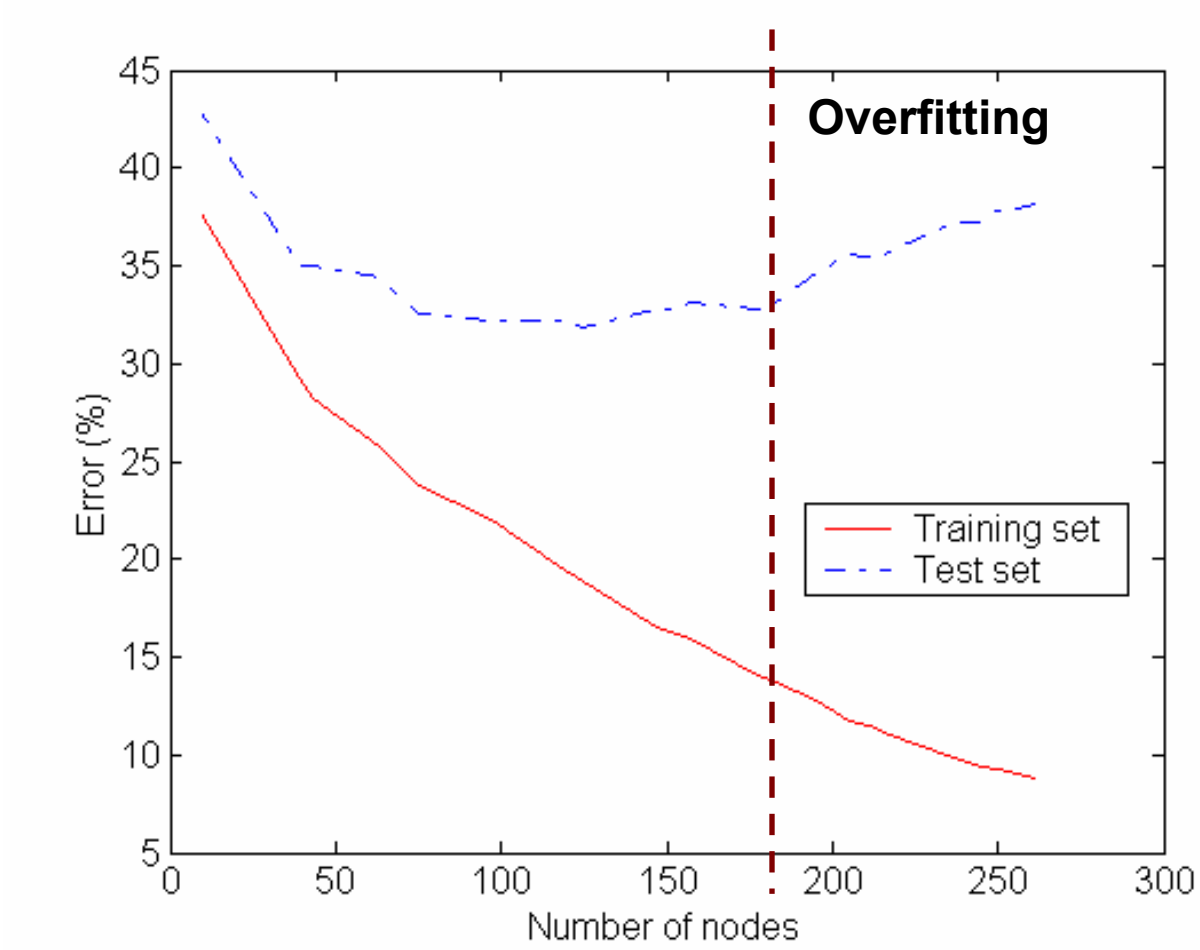
$0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$

Triangoli:

$\text{sqrt}(x_1^2+x_2^2) > 0.5$ o

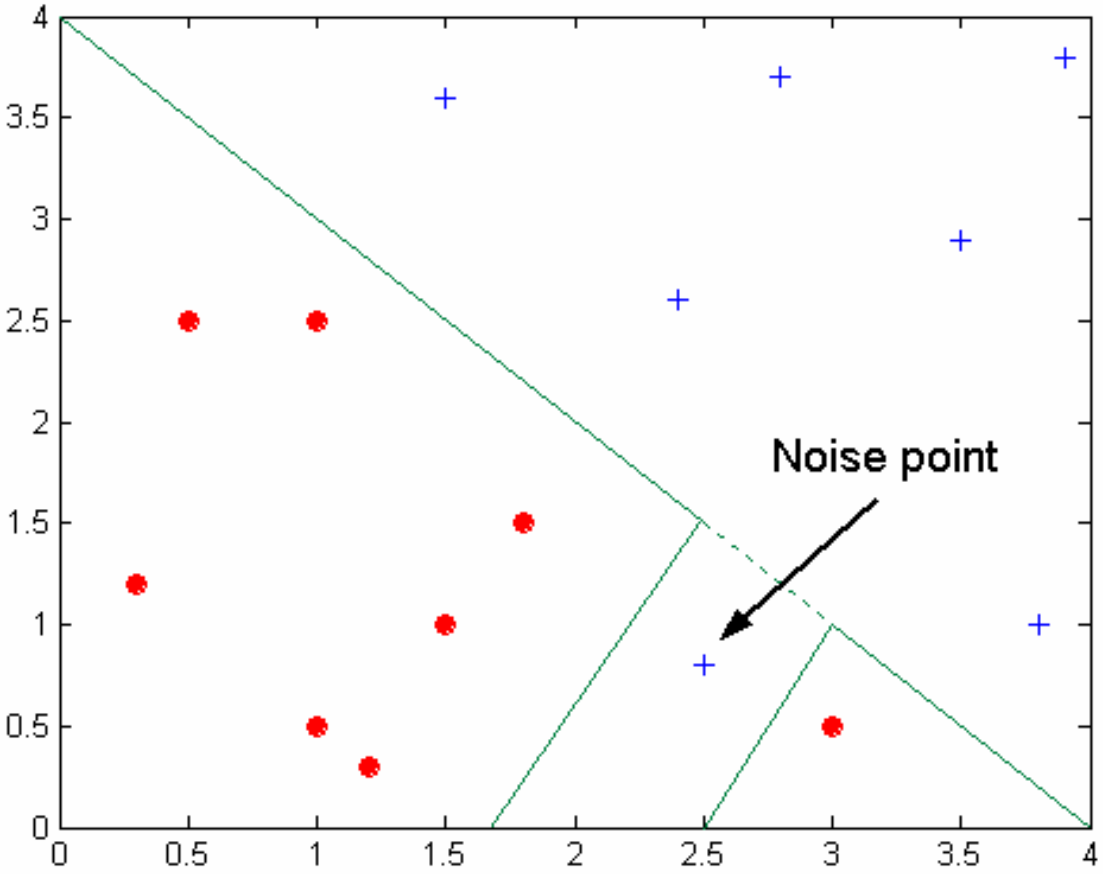
$\text{sqrt}(x_1^2+x_2^2) < 1$

Underfitting, Overfitting



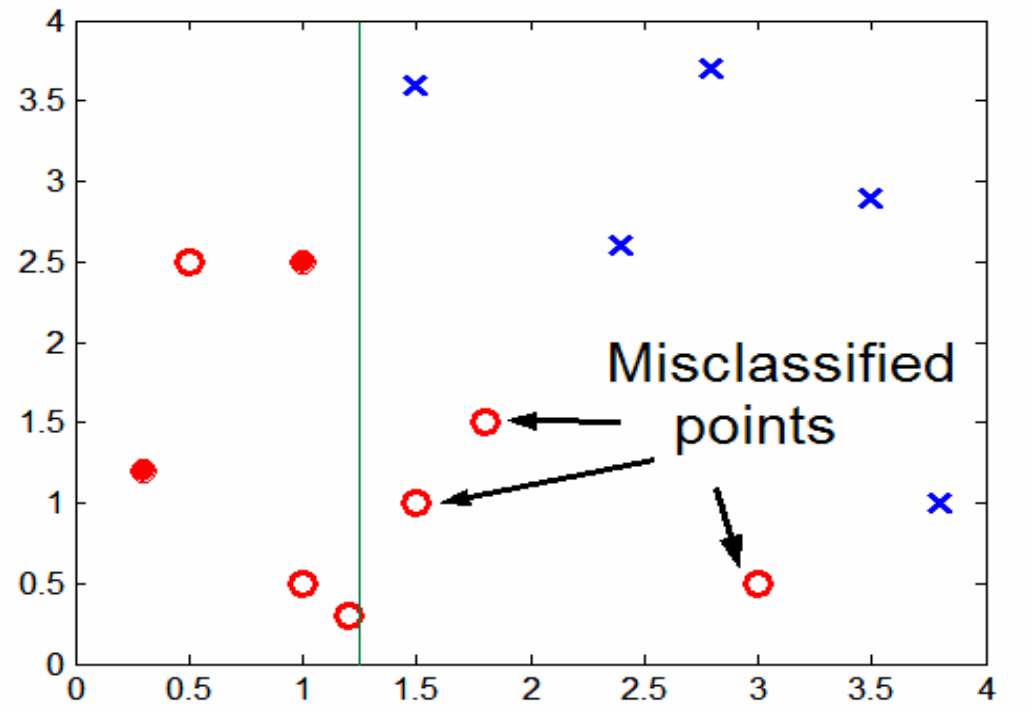
Underfitting: modello stroppo semplice -> errori su training e test grandi

Overfitting dovuto al rumore



Il decision boundary è distorto a causa del rumore

Overfitting dovuto a un training set ridotto



La mancanza di dati nella metà inferiore del grafico rende difficoltosa la predizione dell'etichetta di classe in quella regione

La mancanza di dati sul training set causa la predizione del test set sulla base di dati irrilevanti

Commenti sull'Overfitting

- Overfitting rende i modelli più complessi del necessario
- L'errore sul training set non fornisce una stima appropriata di come il modello si comporterà sui dati non visti
- Need new ways for estimating errors

Errore

- Si possono ottenere ipotesi consistenti?
  - È auspicabile?
  - $error_D(h)=|\{x|h(x) \neq c(x), <x,c(x)> \in D\}|$
- È sempre possibile ottenere un modello con l'errore minimale
  - Perché?
  - È auspicabile?

Overfitting

- **Definizione**
  - *h* presenta overfitting su *D* se ∃ un'ipotesi alternativa *h'* per la quale
    - $error_D(h) < error_D(h')$ but $error_{test}(h) > error_{test}(h')$
  - Cause tipiche: traiing set troppo piccolo (le decisioni sono basate su pochi dati); rumore
- **Come si allevia l'overfitting?**
  - Prevenzione
    - Selezionare solo gli attributi *rilevanti* (utili nel modello)
    - Richiede una misura della rilevanza
  - aggiramento
    - Schivare il problema quando c'è sentore che sta per avvenire
    - Valutare h su un insieme di test e fermare la costruzione del modello quando le performances scadono
  - Riabilitazione
    - "terapia di recupero"
    - Costruzione del modello, eliminazione degli elementi che contribuiscono all'overfitting

Due definizioni di errore

- **Errore "vero"**
  - **Visione probabilistica**

$$error_D(h) = \underset{x \in D}{P}(c(x) \neq h(x))$$

- **Errore sul campione**
  - **Visione frequentistica**

$$error_s(h) = \frac{1}{n} \sum_{x \in S} \delta(c(x) \neq h(x))$$

- **Quanto *error*$_S$*(h)* approssima *error*$_D$*(h)*?**

Esempio

- **h misclassifica 12 esempi su 40 S**

$$error_S(h) = \frac{12}{40} = .30$$

- **Qual'è $error_D(h)$?**

Stime, Previsioni

- Dato *S* di dimensione *n*
- Si valuti $error_S(h)$
  - $error_S(h)$ è una variabile casuale
- Cosa possiamo concludere?

Intervalli di confidenza [1]

- **Se**
  - **S contiene n istanza**
  - **n>30**
- **allora**
  - **Con probabilità 95%, $error_D(h)$ si trova nell'intervallo**

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Intervalli di confidenza [2]

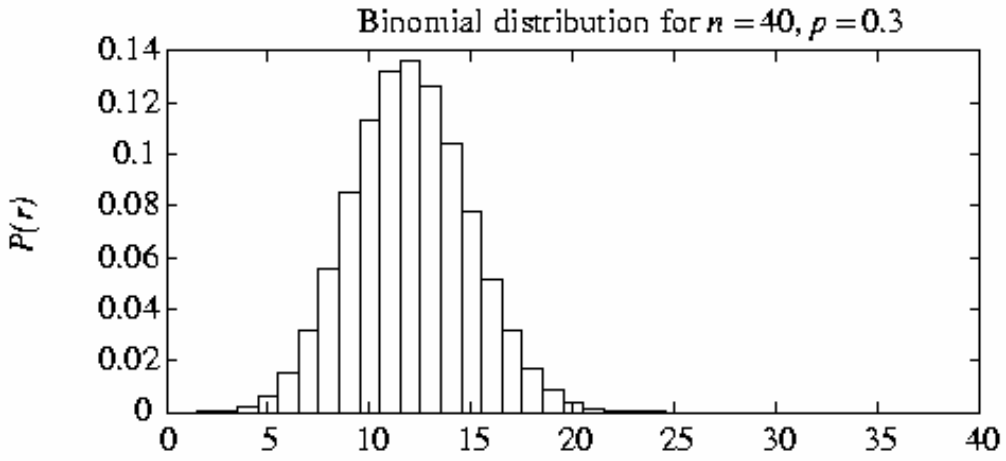
- **Se**
  - **S contiene n istanza**
  - **n>30**
- **allora**
  - **Con probabilità N%, $error_D(h)$ si trova nell'intervallo**

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

N%	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

$error_S(h)$ è una variabile casuale

- **La probabilità di osservare r misclassificazioni:**

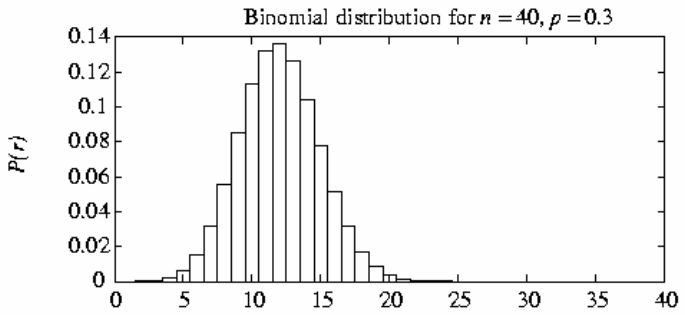


$$P(r) = \frac{n!}{r!(n-r)!} error_D(h)^r (1 - error_D(h))^{n-r}$$

Probabilità Binomiale

- P(r) = probabilità di avere r teste nel lancio della monetina
  - P(head) = p
- Media

$$E[X] = \sum_i P(i) = np$$



- Varianza

$$Var[X] = E\left[\left(X - E[X]\right)^2\right] = np(1-p)$$

- Devianza

$$\sigma_X = \sqrt{Var[X]} = \sqrt{np(1-p)}$$

$error_S(h)$

- ***$error_S(h)$ segue una distribuzione binomiale***
  - ***Per definizione,***

$$error_S(h) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X_i = \begin{cases} 0 & \text{se } c(x_i) = h(x_i) \\ 1 & \text{altrimenti} \end{cases}$$

  - ***Assumendo***

$$E[X_i] = \mu$$

$$Var[X_i] = \sigma^2$$

  - ***Otteniamo***

$$E[\overline{X}] = \mu$$

$$Var[\overline{X}] = \frac{\sigma^2}{n}$$

Approssimiamo $error_S(h)$

- **Media**

$$\mu_{error_S(h)} = error_D(h)$$

- **devianza**

$$\sigma_{error_S(h)} = \sqrt{\frac{error_D(h)(1-error_D(h))}{n}}$$

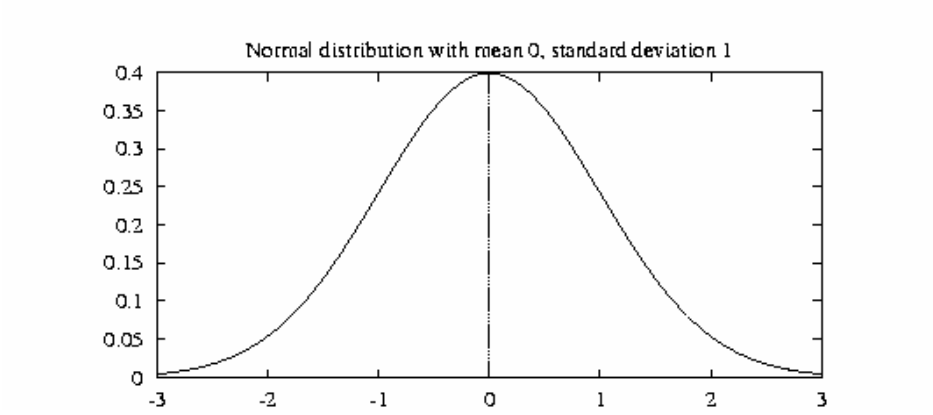
- **Utilizzando la distribuzione normale**
  - **media**

$$\mu_{error_S(h)} = error_D(h)$$

  - **varianza**

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Distribuzione Normale



- **densità**

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **distribuzione**

$$P(a \leq X < b) = \int_a^b p(x)\mathrm{d}x$$

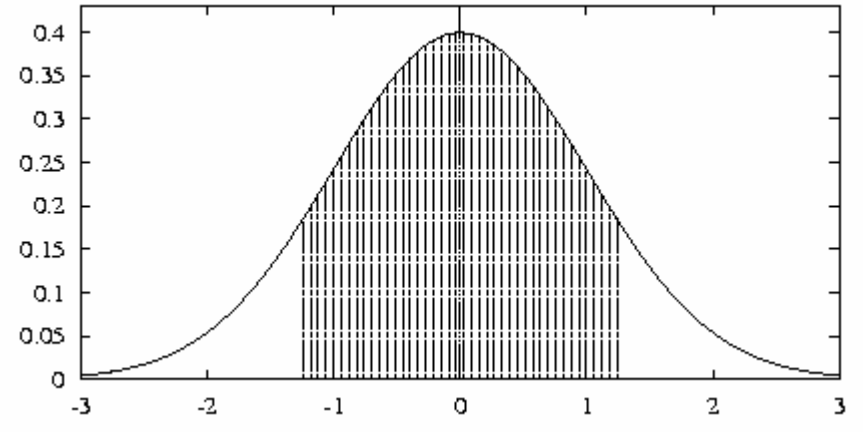
- **media**

$$E[X] = \mu$$

- **varianza**

$$Var[X] = \sigma^2$$

Distribuzione Normale



- 80% dell'area (probabilità) si trova in $\mu+1.28\sigma$
- N% dell'area (probabilità) si trova in $\mu+z_N\sigma$

N%	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Intervalli di confidenza

- Se S contiene n istanze, n>30
- allora
  - Con probabilità N%, $error_S(h)$ si trova nell'intervallo

$$error_D(h) \pm z_N \sqrt{\frac{error_D(h)(1-error_D(h))}{n}}$$

  - equivalentemente, $error_D(h)$ si trova nell'intervallo

$$error_S(h) \pm z_N \sqrt{\frac{error_D(h)(1-error_D(h))}{n}}$$

  - In base al teorema del Limite Centrale,

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Calcolo degli intervalli di confidenza

- Si sceglie il parametro da stimare
  - $error_D(h)$
- Si sceglie un'approssimazione
  - $error_S(h)$
- Si determina la probabilità che governa l'approssimazione
  - $error_S(h)$ è binomiale, approssimata dalla distribuzione normale per n>30
- Si trovano gli intervalli (L,U) per cui N% della probabilità ricade in [L,U]
  - Si usa la tabella dei valori z_N

Test di significatività

- **Dati due modelli:**
  - M1: accuratezza = 85%, test su 30 instanze
  - M2: accuratezza = 75%, test 5000 instanze

- **Possiamo dire che M1 è migliore di M2?**
  - Con che confidenza possiamo affidarci alle relative accuratezze?
  - Potrebbe la differenza nelle performance essere spiegata come un risultato delle fluttuazioni random nel test set?

Confrontare due modelli

- **Qual'è il migliore tra M1 e M2?**
  - **M1 testato su D1 (size=n1), error = e_1**
  - **M2 testato su D2 (size=n2), error = e_2**
  - **Assumendo D1 e D2 indipendenti,**
  - **If n1 and n2 are sufficiently large, then**

$$e_1 \sim N(\mu_1, \sigma_1)$$
$$e_2 \sim N(\mu_2, \sigma_2)$$

  - **Con**

$$\hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

Confrontare due modelli

- **Misura: d = e1 – e2**
  - d ~ *N*(d_t,σ_t), dove d_t è la differenza effettiva
  - Poiché D1 e D2 sono indipendenti, la loro varianza si somma:

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$

$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

  - Al livello di confidenza (1-α),

$$d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$$

Un esempio

- **Dati: M1: n1 = 30, e1 = 0.15**
  M2: n2 = 5000, e2 = 0.25
- **d = |e2 – e1| = 0.1**

$$\hat{\sigma}_d = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- **Al 95% confidence level, $Z_{\alpha/2}$=1.96**

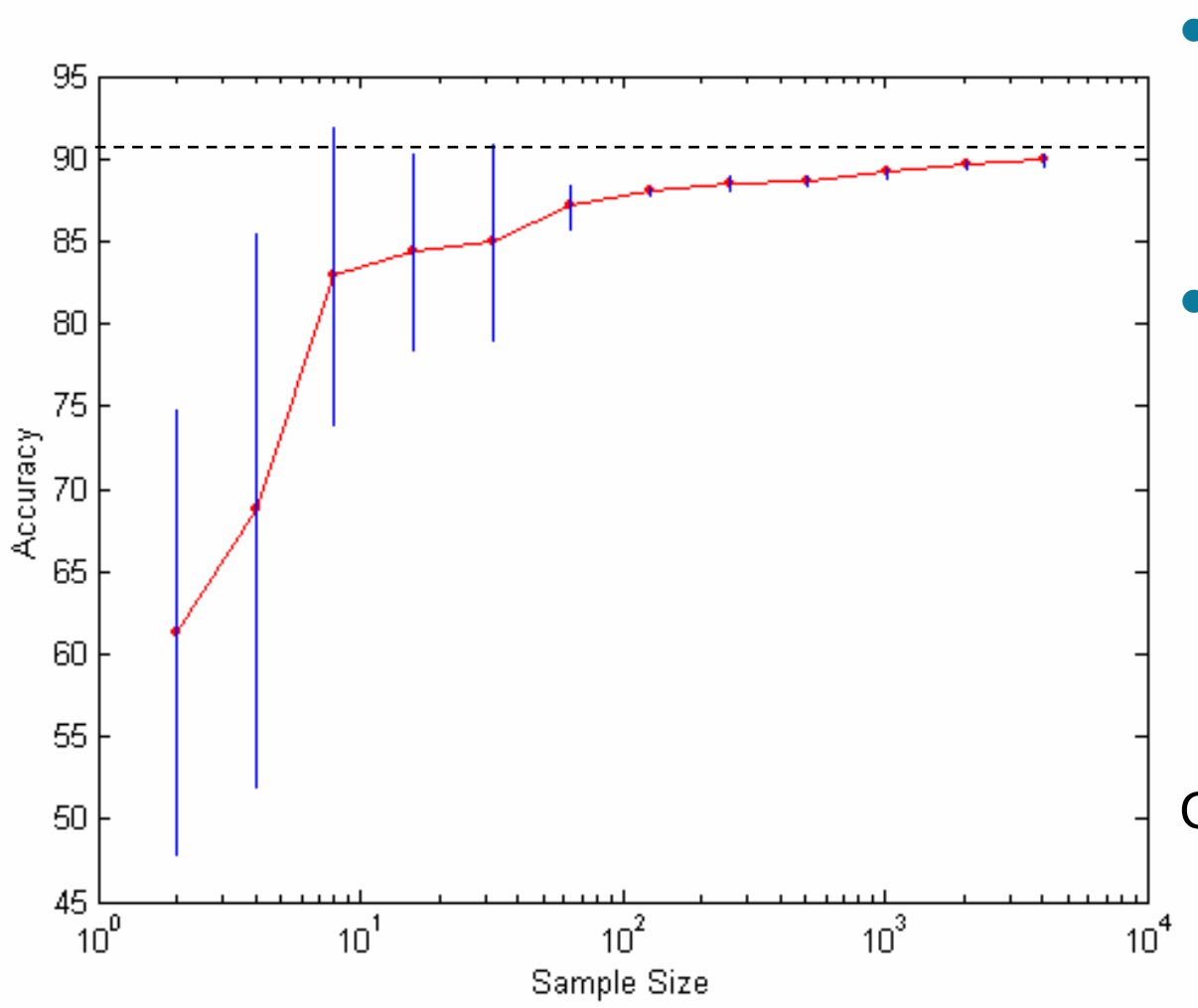
$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> L'intervallo contiene 0 => la differenza non è significativa dal punto di vista statistico

Metodi per valutare le performance

- L'efficacia di un modello può dipendere da più fattori indipendenti dall'algoritmo di learning
  - Distribuzione di Classe
  - Costo della misclassificazione
  - Dimensione di training e test set

La curva di learning



- Mostra come l'accuratezza cambia con campioni di dimensioni differenti
- Generabile tramite meccanismi di sampling:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)

Gli effetti di campioni piccoli:

- Bias nella stima
- Varianza alta

Stima del modello

- Holdout
  - 2/3 per il training, 1/3 per il test
- Random subsampling
  - Holdout ripetuto
- Cross validation
  - Partizioniamo i dati in k sottoinsiemi disgiunti
  - k-fold: apprendi su k-1 partizioni, valuta sulla partizione rimanente
  - Leave-one-out: k=n
- Stratified sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling con rimpiazzamento

Metriche per la valutazione di modelli

- **Il focus è sulla capacità predittiva di un modello**
  - **Piuttosto che su velocità, scalabilità, ecc.**

- **Matrice di confusione:**

		CLASSE PREDETTA	
		Class=Yes	Class=No
CLASSE ATTUALE	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metriche per la valutazione di modelli …

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

- **La metrica d'uso comune:**

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

Limiti dell'accuratezza

- **Classificazione binaria**
  - **Numero di esempi in classe 0 = 9990**
  - **Numero di esempi in classe 1 = 10**

- **Se il modello predice ogni cosa in classe 0, l'accuratezza è 9990/10000 = 99.9 %**
  - **L'accuratezza è fuorviante perché il modello non predice nessun esempio di clase 1**

La matrice di costo

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

C(i|j): Quanto mi costa misclassificare un esempio di classe j in classe i

I costi della classificazione

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

Accuratezza = 80%

Costo = 3910

Accuratezza = 90%

Costo = 4255

Costo vs accuratezza

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

L'accuratezza è proporzionale al costo se
1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

N = a + b + c + d

Accuratezza = (a + d)/N

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

Costo = p (a + d) + q (b + c)

= p (a + d) + q (N – a – d)

= q N – (q – p)(a + d)

= N [q – (q-p) × Accuratezza]

Misure Cost-Sensitive

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

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

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

- Precision sbilanciata verso C(Yes|Yes) & C(Yes|No)
- Recall sbilanciata verso C(Yes|Yes) & C(No|Yes)
- F-measure sbilanciata verso tutto tranne che C(No|No)

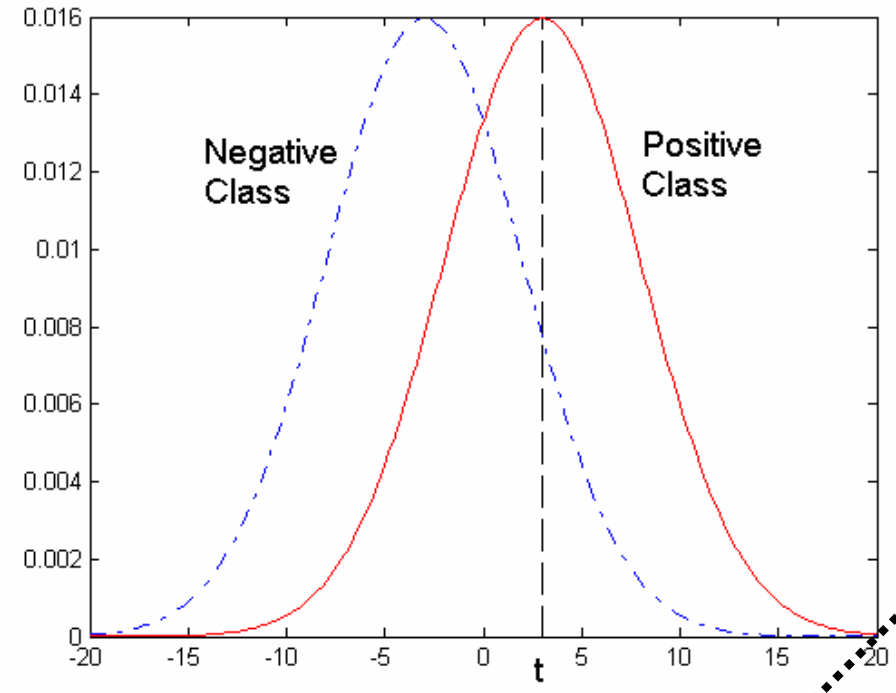
$$\text{Accuratezza pesata} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

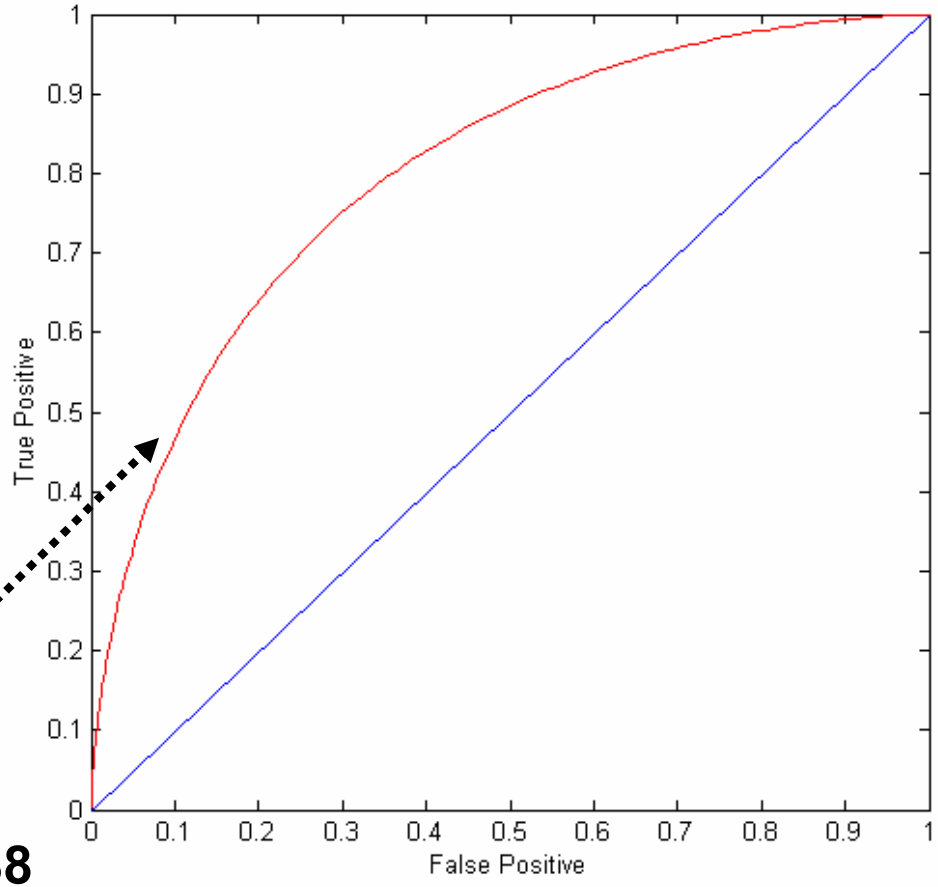
ROC (Receiver Operating Characteristic)

- Sviluppata negli anni ’50 per l’analisi dei segnali
  - Caratterizza il trade-off tra classificazioni positive e falsi allarmi
- La curva traccia TP (sull’asse y) e FP (sull’asse x)
- La performance di un classificatore è rappresentata come un punto sulla curva di ROC
  - Le variazioni del threshold (in classificazioni con soglia), la distribuzione del campione o la matrice di costo cambia la collocazione del punto
    - Traccia la curva

ROC Curve

- 1-dimensional data set contenente 2 classi (positiva and negativa)
- Ogni punto x > t è classificato come positivo





Al threshold t:

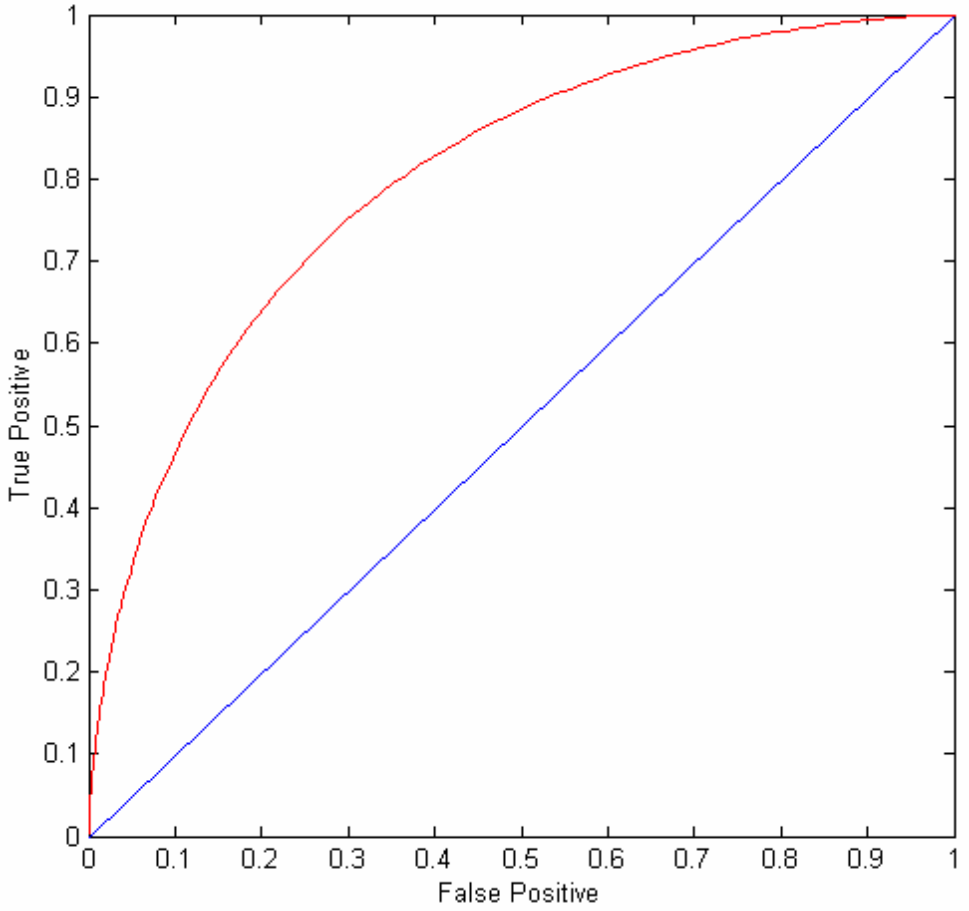
TP=0.5, FN=0.5, FP=0.12, FN=0.88

ROC Curve

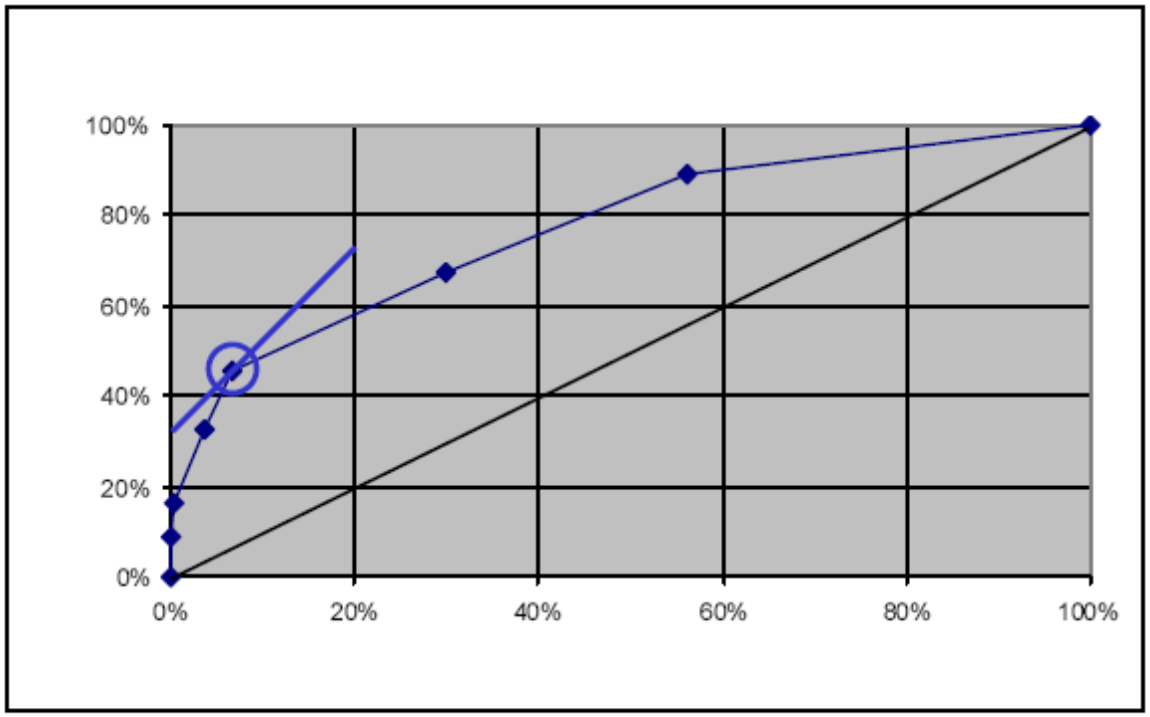
(TP,FP):

- (0,0): ogni elemento di classe negativa
- (1,1): ogni elemento di classe positiva
- (1,0): classificazione ideale

- Linea diagonale:
  - Scelta casuale
  - Sotto la linea diagonale:
    - predizione opposta alla classe effettiva

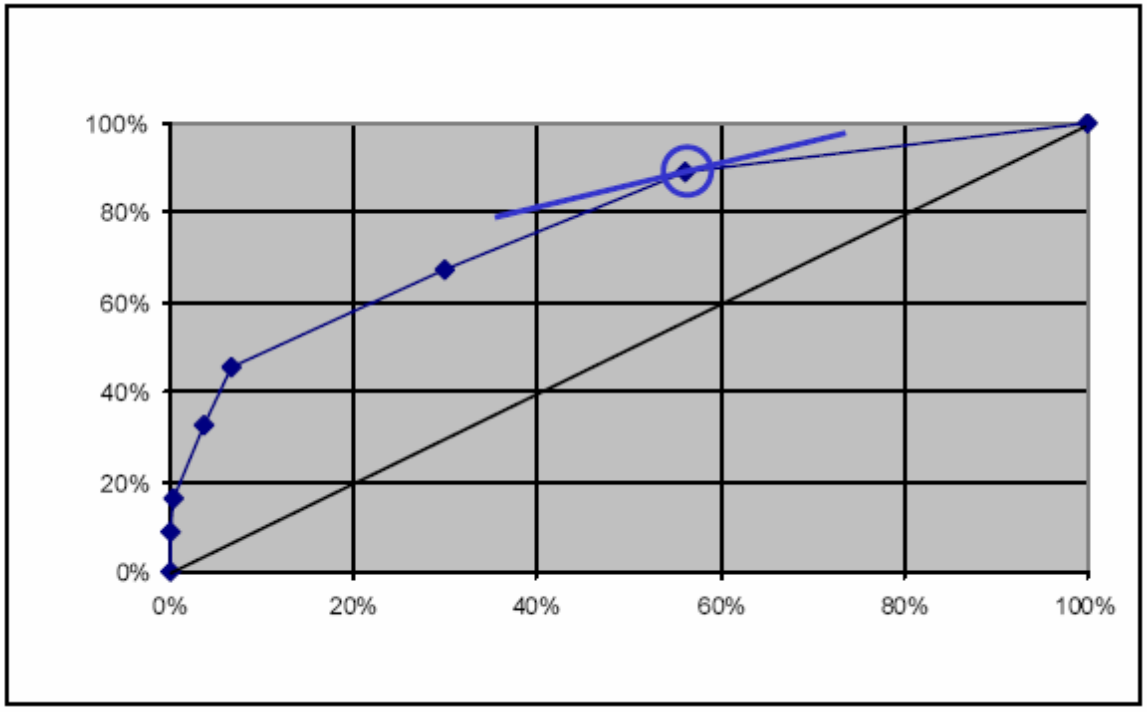


Confronto di Modelli



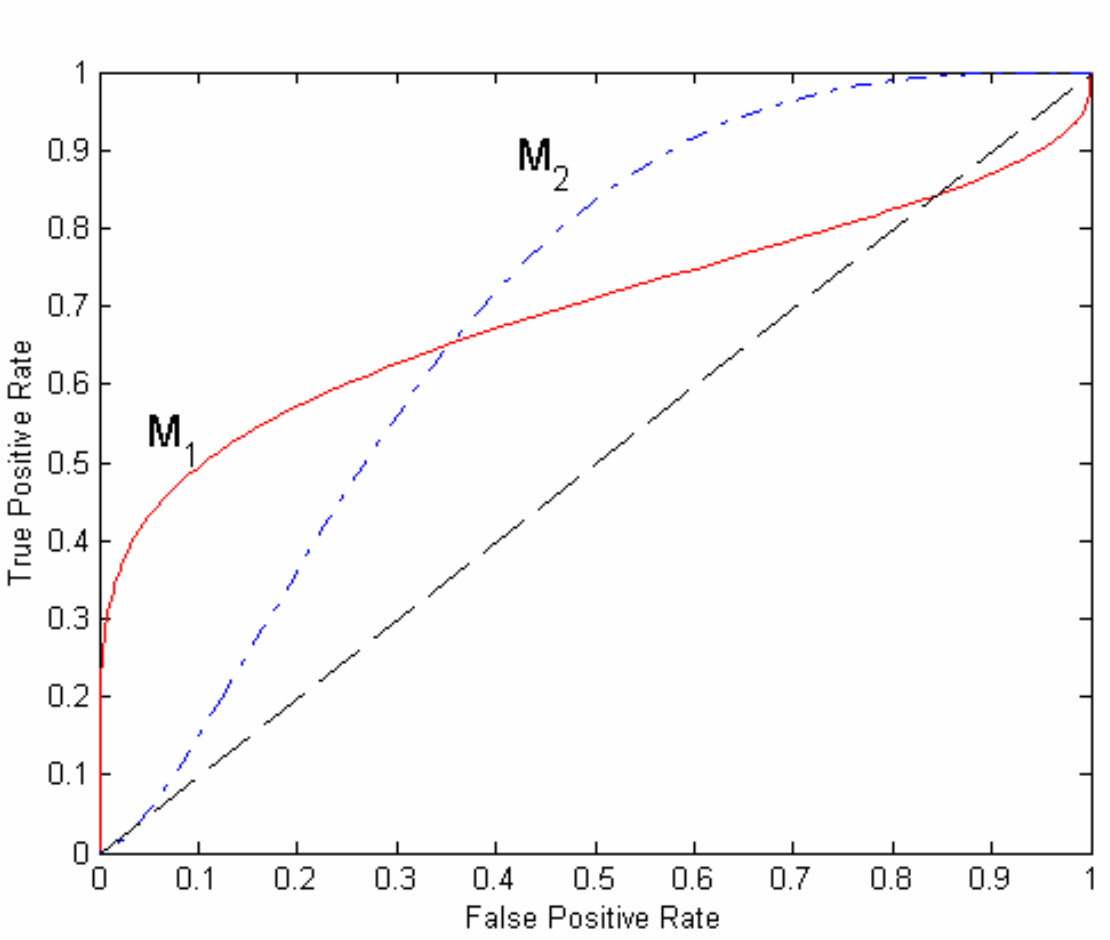
- Contesto applicativo 1
- FPCost/FNCost = ½
- Neg/Pos=2
- Slope = 4/2=2

Confronto di Modelli



- Contesto applicativo 2
- FPCost/FNCost = 1/8
- Neg/Pos=4
- Slope = 4/8=.5

Utilizzo di ROC per il confronto di modelli



- Nessun modello ha il sopravvento
  - M_1 migliore su FPR bassi
  - M_2 migliore su FPR alti
- Area Under the ROC curve
  - Ideale:
    - Area = 1
  - Classificazione a caso:
    - Area = 0.5

Come costruire una ROC curve

Instance	P(+|x)	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	+

- Assunzione: un classificatore produce una posterior probability $P(+|x)$ per ogni istanza di test x
- Ordiniamo le istanze in base per valori decrescenti di $P(+|x)$
- Applichiamo un threshold per ogni valore di $P(+|x)$
- Contiamo il numero di TP, FP, TN, FN per ogni threshold
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

Come costruire una 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: