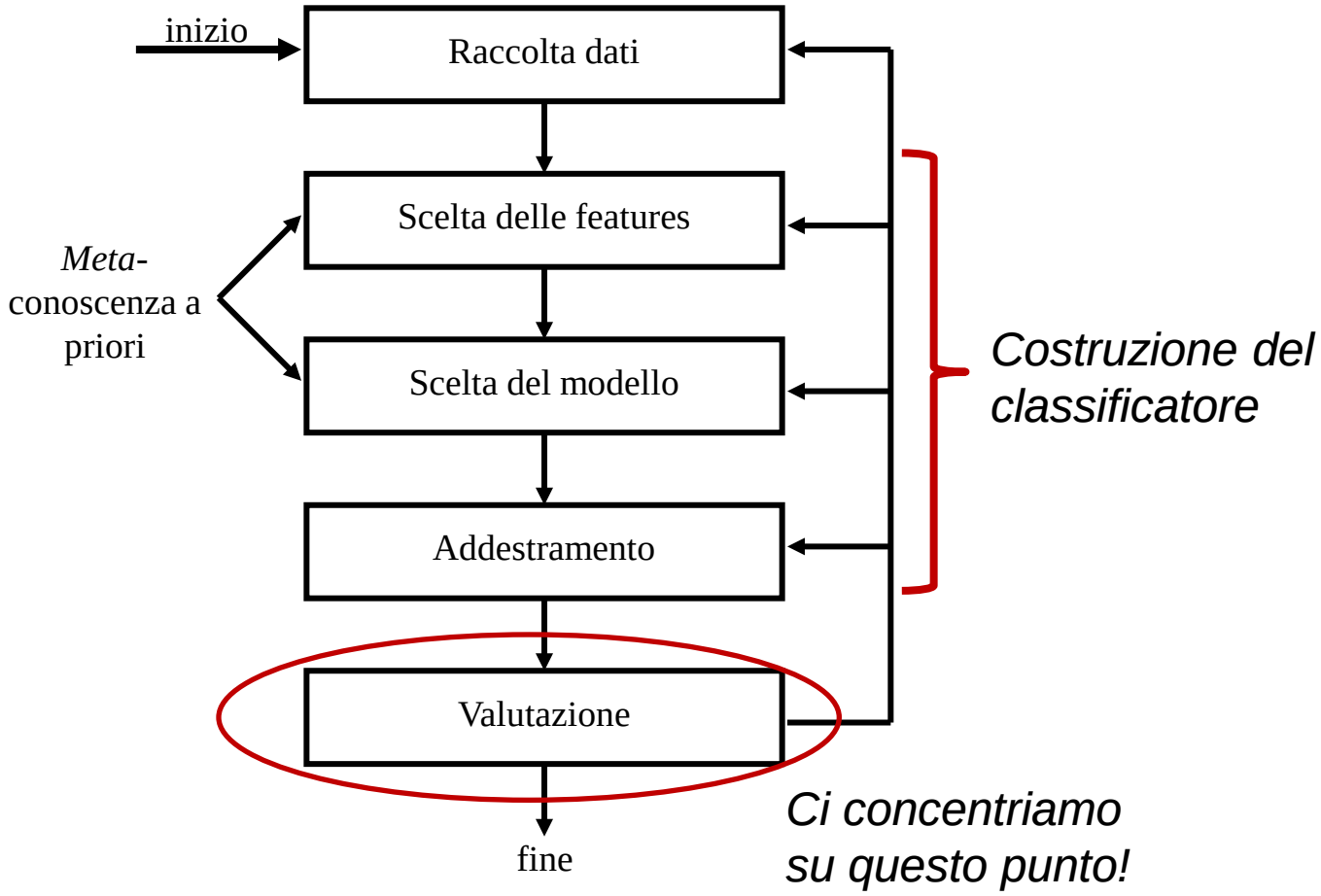
Università di Verona
Dipartimento di Informatica
Corso di Laurea magistrale in Ingegneria e Scienze Informatiche

Sistemi Avanzati per il Riconoscimento

Lez.01 - Valutazione dei Classificatori Supervisionati

Ciclo di vita di un classificatore



Tipologia di un classificatore

$$Y = f(X) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$$

- Y variabile aleatoria (v.a.) da predirre, X la v.a. visibile (che genera le osservazioni **x**), ε **errore irriducibile** (piccolo a piacere)
- **Classificatore** (propriamente detto):

$$Y \in \{\omega_1, ..., \omega_C\}$$

  - lo spazio delle etichette NON è una metrica, ε=0
- **Regressore** $Y \in \mathcal{R}$
  - lo spazio delle misure di output E' una metrica

Valutazione di un classificatore supervisionato

- I classificatori supervisionati vengono addestrati (trainati) su un dataset finito etichettato chiamato **training set**
- Un classificatore addestrato deve essere testato su un dataset differente da quello di training, altrimenti le stime di bontà di classificazione che si ottengono sul training set risultano troppo ottimistiche: esigenza di un dataset di **testing**
- La sperimentazione sul dataset di testing rappresenta un proxy su come si comporterà il classificatore su dati nuovi. In pratica, il test set serve a capire la capacità di generalizzazione del classificatore

Valutazione di un classificatore supervisionato

- Esiste quindi la necessità di capire come un classificatore si comporta in maniera quantitativa
- Sono necessari cioè dei criteri che producano delle misure di valutazione dei classificatori, in modo tale da avere
  - Un riscontro assoluto della bontà di un classificatore
  - Un riscontro relativo della bontà di un classificatore, ossia la capacità di confrontare classificatori diversi operanti sullo stesso dataset di test

Misure di valutazione di un classificatore supervisionato (propriamente detto)

$$Y = f(X) \qquad Y \in \{\omega_1, \ldots, \omega_C\}$$

Valutazione di un classificatore

- Assumiamo di avere un generico problema di classificazione a C classi $\{\omega_1,...,\omega_C\}$ ed un pattern da classificare $\boldsymbol{x}$ con
  - $\mathbf{x} = \{x_1, x_2,..., x_d\}, \mathbf{x} \in \mathrm{R}^d$
  - R^d spazio delle feature
  - $\boldsymbol{x}$ prodotto da una particolare variabile aleatoria X

Misure di Valutazione

- Accuratezza (*accuracy*) $= \frac{\#\,\text{correct classifications}}{\#\,\text{classifications}}$
- Errore (*error rate*) $= \frac{\#\,\text{incorrect classifications}}{\#\,\text{classifications}}$

  $= 1 - \text{error rate}$

- Problemi con l'accuratezza
  - Assume *costi uguali* per gli errori
  - Assume una *distribuzione uniforme dei campioni nelle C classi*

Costi diseguali di errore

- Esempi (C=2 classi)
  - Il costo di classificare erroneamente dei particolari sintomi come tumore al pancreas durante uno screening è *minore* di mancare la segnalazione di un caso di malattia effettivamente presente
  - Il costo per una banca di classificare erroneamente un cliente come affidabile è *maggiore* di mancare la segnalazione di un cliente effettivamente affidabile

Costi diseguali di errore

- Considero quindi la teoria di classificazione di Bayes

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})}$$

con $\{\omega_1, \omega_2, ..., \omega_C\}$ classi

- Suppongo di avere delle azioni

$$\{\alpha_1, \alpha_2, ..., \alpha_C\}$$

associate alla scelta di una classe

Costi diseguali di errore

- Suppongo di avere delle funzioni di costo

$$\lambda(\alpha_i \mid \omega_j)$$

che descrivano il costo (o la perdita) dell'azione α_i quando lo stato è ω_j ;

- Supponiamo di osservare un particolare **x**, e decidiamo di effettuare l'azione α_i : per definizione, saremo soggetti al costo $\lambda(\alpha_i \mid \omega_j)$

Costi diseguali di errore

- RICHIAMO TTR: In fase di classificazione (e quindi non quando devo valutare un errore, ma quando devo scegliere effettivamente una classe) la teoria di Bayes mi indica come usare le funzioni di costo, introducendo la **perdita attesa** (*expected loss*), o **rischio condizionale** (*conditional risk*)

Costi diseguali di errore

- RICHIAMO TTR: Data l'indeterminazione di ω_j (ossia non ho la certezza che ω_j sia la classe corretta) la perdita attesa associata alla scelta di una particolare classe i-esima sarà

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

- la teoria di decisione di Bayes indica di effettuare l'azione che minimizza il rischio condizionale

Costi diseguali di errore

- In fase di valutazione dell'errore, so però quale sia la classe $\omega_{\tilde{j}}$ che ho associato erroneamente ad un particolare **x**, pertanto l'errore di scelta di una particolare classe (appunto, scorretta) sarà

$$\lambda(\alpha_i \mid \omega_{\tilde{j}})P(\omega_{\tilde{j}} \mid \mathbf{x})$$

- Quindi una misura alternativa di errore è quella che somma i costi di tutti gli errori fatti

Costi diseguali di errore (1)

- In forma contratta posso usare

$$\lambda(\alpha_i \mid \omega_j) = c_{ij}$$

dove di solito

$$c_{ij} = 0 \text{ se } i = j,\ c_{ij} \neq 0 \text{ altrimenti}$$

- A partire da questo coefficienti mi posso costruire una

matrice di costo $\underset{C \times C}{Co}$

Distribuzione uniforme dei campioni

- In molte circostanze, fino al momento della classificazione, non si riesce a conoscere la distribuzione di osservazioni nelle varie classi
- Vi sono problemi reali in cui di solito si hanno classi naturalmente sbilanciate
  - Diagnosi medica: 95% sani, 5% malati
  - e-Commerce: 99% non compra, 1% compra
  - Sicurezza: 99.999 % dei cittadini non sono terroristi

Distribuzione uniforme dei campioni - bilanciamento

- Situazioni simili si possono trovare nei problemi di classificazione multiclasse
- Soluzione
  - Costruisci un *training set bilanciato*
    - Campiona senza rimpiazzo dalla classe *meno numerosa (*di cardinalità N*)*, fino ad *M<N* campioni
    - Fai lo stesso con le altre classi, *M* campioni per classe
  - Costruisci un *testing set bilanciato*
    - Prendi i rimanenti *N-M* campioni dalla classe meno numerosa
    - Campiona dalle rimanenti classi *N-M* esemplari

Matrice di confusione

- Permette di catturare la capacità di un classificatore di classificare meglio alcune classi, peggio altre
- Storicamente, introdotta per modellare casi di classificazione binaria (C=2)
- Di solito, classe 1=positiva, classe 2=negativa
- Esempi
  - 1=pedoni, 2=non pedoni; 1=presenza di condizione; 2=assenza di condizione

Matrice di confusione - schema

		Indici predetti	
Indici reali		Classificazione positiva	Classificazione negativa
	Presenza di condizione	**Vero positivo** *tp*	**Falso negativo** *fn* *(type II error)*
	Assenza di condizione	**Falso positivo** *fp* *(type I error)*	**Vero negativo** *tn*

Matrice di confusione - costruzione

- Ad ogni classificazione sul testing set, aggiungo +1 nell'apposita casella (ciò è possibile poiché conosco per ogni campione di test la sua classe predetta e quella reale)
- In caso di classi bilanciate (perfettamente) avro' che la somma su ogni riga mi darà lo stesso numero, ossia la cardinalità della classe
- In caso di classi bilanciate, risulta conveniente normalizzare per righe e ottenere percentuali

Indici predetti

Indici reali	P	N
P	20 (0.66)	10 (0.33)
N	5 (0.17)	25 (0.83)

Matrice di confusione - misure

- Varie misure (**tutte** in [0,1])
  - **Accuratezza** (*accuracy*) $\frac{tp + tn}{tp + tn + fp + fn}$
  - **Precisione** (*precision*, positive predicted value PPV) $\frac{tp}{tp + fp}$
    - Proporzione dei casi predetti come positivi (*tp + fp*) che effettivamente lo sono
    - (SE ALTA) prendo come positivi solo elementi che lo sono
    - (SE BASSA) dico che tutto è positivo, non filtro

Indici reali \ Indici predetti		
	tp	fn
	fp	tn

Matrice di confusione - misure

- **Sensitività**, recupero, richiamo (*recall*, *true positive rate TPR*, sensitivity, hit rate)

$$\frac{tp}{tp + fn}$$

Indici predetti

Indici reali		
	tp	fn
	fp	tn

  - Proporzione di tutti i casi effettivamente positivi in gioco (*tp+fn*) che sono stati classificati effettivamente come tali
  - (SE ALTA) non perdo elementi positivi
  - (SE BASSA) perdo elementi positivi
- **F-measure**
  - Combina precisione e sensitività in un'unica misura
  - Media armonica tra precisione e sensitività

$$2 \cdot \frac{\text{precisione} \cdot \text{sensitività}}{\text{precisione} + \text{sensitività}}$$

Matrice di confusione - misure

- **Specificità** (specificity SPC, true negative rate)

$$\frac{tn}{fp + tn}$$

  - Porzione di tutti i casi negativi (*fp + tn*) che sono stati classificati correttamente
  - Vale il ragionamento fatto per la sensitività, ma fatto per i negativi

- **Fall-out** (*false positive rate FPR*)

$$\frac{fp}{fp + tn}$$

  - Proporzione dei negativi che sono stati scorrettamente classificati come positivi
  - In pratica, quanti dei negativi che ho in gioco (*fp + tn*) sono stati accettati come positivi

Indici reali \ Indici predetti		
	tp	fn
	fp	tn

Matrice di confusione - altre misure

Indici predetti

Indici reali

tp	fn
fp	tn

- **False negative rate FNR**
  - Percentuale di positivi che scorrettamente viene persa

$$\frac{fn}{fn+tp}$$

- **Negative predictive value NPV**
  - Vale il ragionamento fatto per la precisione, ma fatto per i negativi

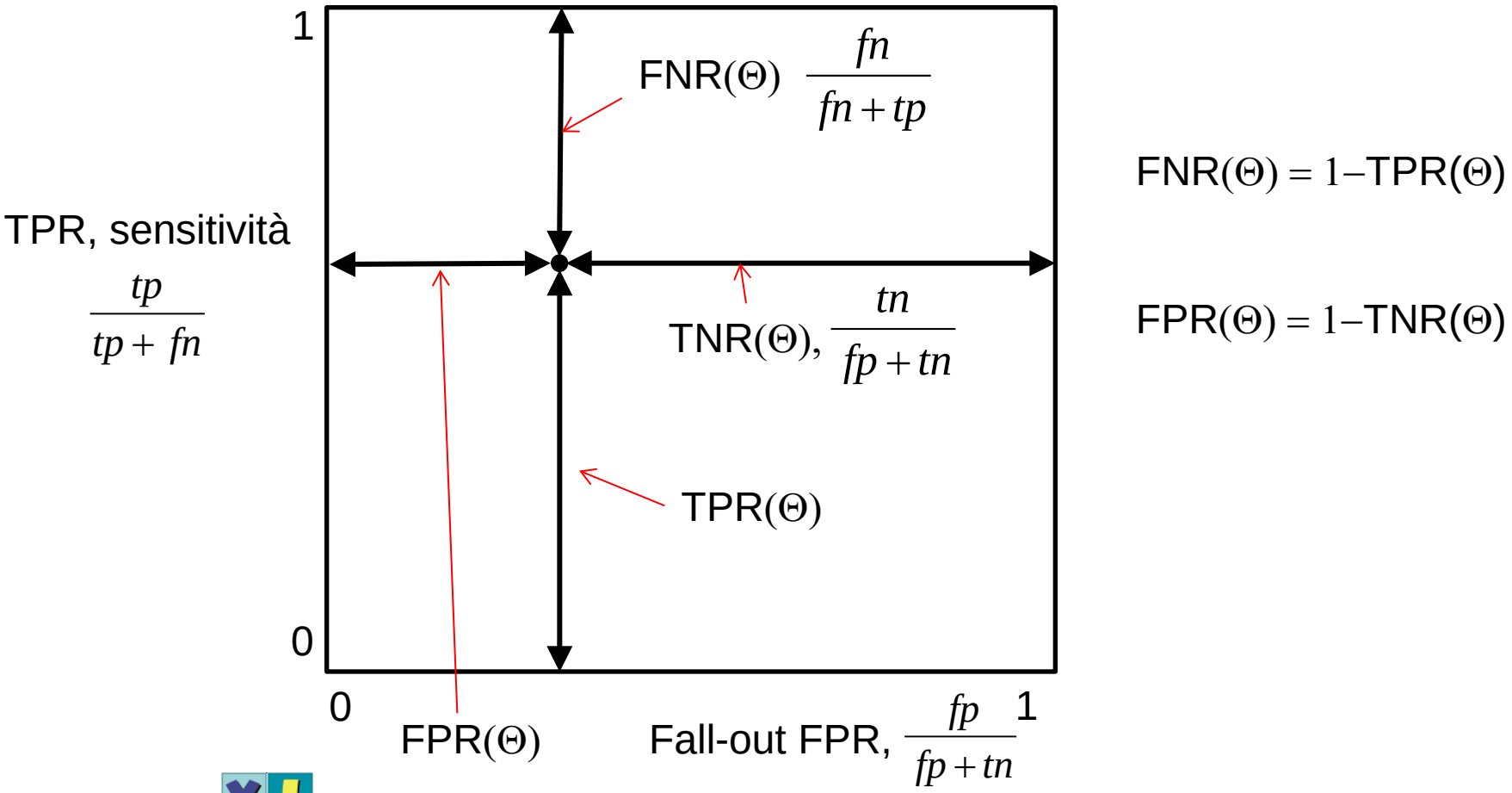
$$\frac{tn}{tn+fn}$$

- **False discovery rate FDR**
  - Proporzione di falsi positivi trovati tra tutti i campioni classificati come positivi
  - Usato spesso in bioinformatica

$$\frac{fp}{fp+tp} = 1 - PPV$$

Matrice di confusione (7)

- Come sono legate (alcune del)le misure viste?



Matrice di confusione – costi

– Includo il concetto di costo

Classifier 1

Indici reali

	P	N
P	20	10 ← fp
N	30	90

fn

Indici predetti

Classifier 2

Indici reali

	P	N
P	10	20
N	15	105

Indici predetti

Matrice di costo Co $\{c_{ij}\}$

Indici reali

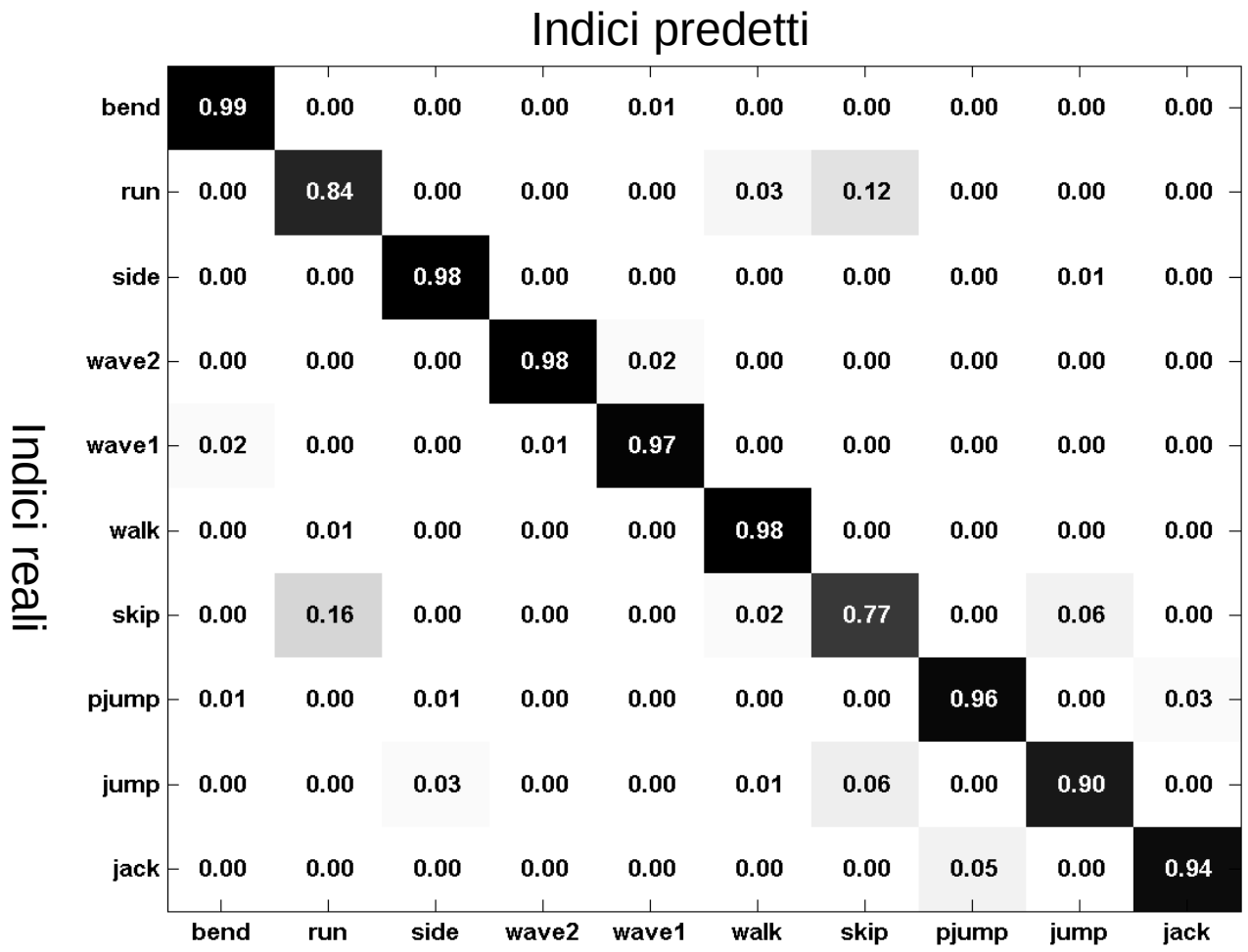
	P	N
P	0	2
N	1	0

Indici predetti

Error rate: 40/150
Costo:
30x1+10x2=50

Error rate: 35/150
Costo:
15x1+20x2=55

Matrice di confusione – C classi

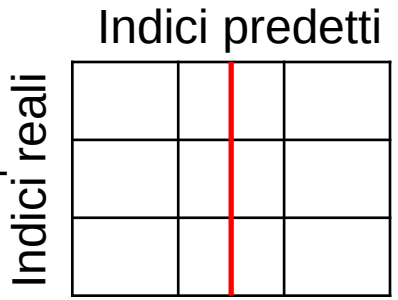


Cosa si può dire?

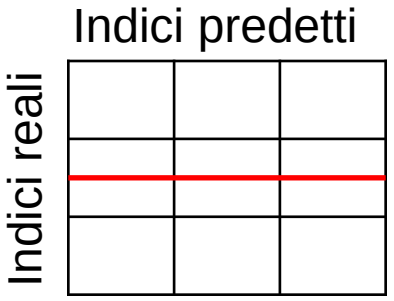
Matrice di confusione – C classi (2)

- Risulta semplice generalizzare precisione e sensitività nel caso multiclasse
- Ovviamente, si parlerà di precisione e sensitività *associata alla singola classe*
- *Conf* matrice di confusione *C x C*

$$\text{precisione}_c = \frac{tp}{tp + fp} = \frac{Conf(c,c)}{\sum_d Conf(d,c)}$$



$$\text{sensitività}_c = \frac{tp}{tp + fn} = \frac{Conf(c,c)}{\sum_d Conf(c,d)}$$

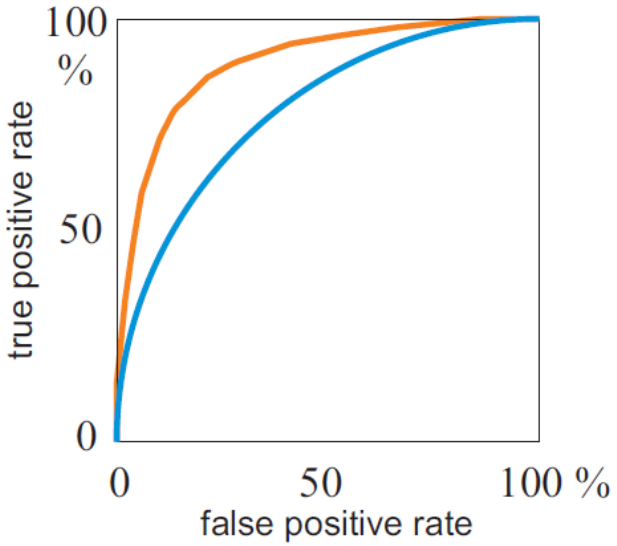


ROC – Receiver Operating Characteristic

- Criterio di valutazione di un classificatore risalente alla II guerra mondiale
- Serve per valutare *classificatori binari*
- Si basa sul concetto di *soglia* Θ
- Utilizza due delle misure già viste finora, catturabili dalla matrice di confusione (con P tutti i positivi, N tutti i negativi):
  - **False positive rate** (FPR, Fall-out, false alarms) $\frac{fp}{fp+tn} = \frac{fp}{N}$
  - **True positive rate** (TPR, *recall*, hit rate) $\frac{tp}{tp+fn} = \frac{tp}{P}$

ROC – Receiver Operating Characteristic - esempio

- E’ una curva che mostra coppie (FPR,TPR), ossia (false alarms, hit rate)



- Differenti curve ROC corrispondono a differenti classificatori
- Di norma, le curve sono non decrescenti

ROC – L'esempio storico

- Suppongo un apparecchio ricevitore che identifica una singola pulsazione (una riflessione radar da un aereoplano)

- Ho un classificatore che deve decidere due classi ω_1, ω_2
  - ω_1 : l'aereoplano non è presente (vero negativo, *tn*)
  - ω_2: l'aereoplano è presente (vero positivo, *tp*)

- Assumo un classificatore semplice a soglia Θ, operante sul voltaggio *x* che la pulsazione mi genera

ROC - assunzioni

- In particolare,
  - Se non ho l'aereo, il voltaggio *x* assumerà il valore μ_1
  - Se ho l'aereo, il voltaggio *x* assumerà il valore μ_2
- Poichè ho del rumore nel ricevitore

$$p(x \mid \omega_i) = N(\mu_i, \sigma_i^2), \quad i = 1,2$$

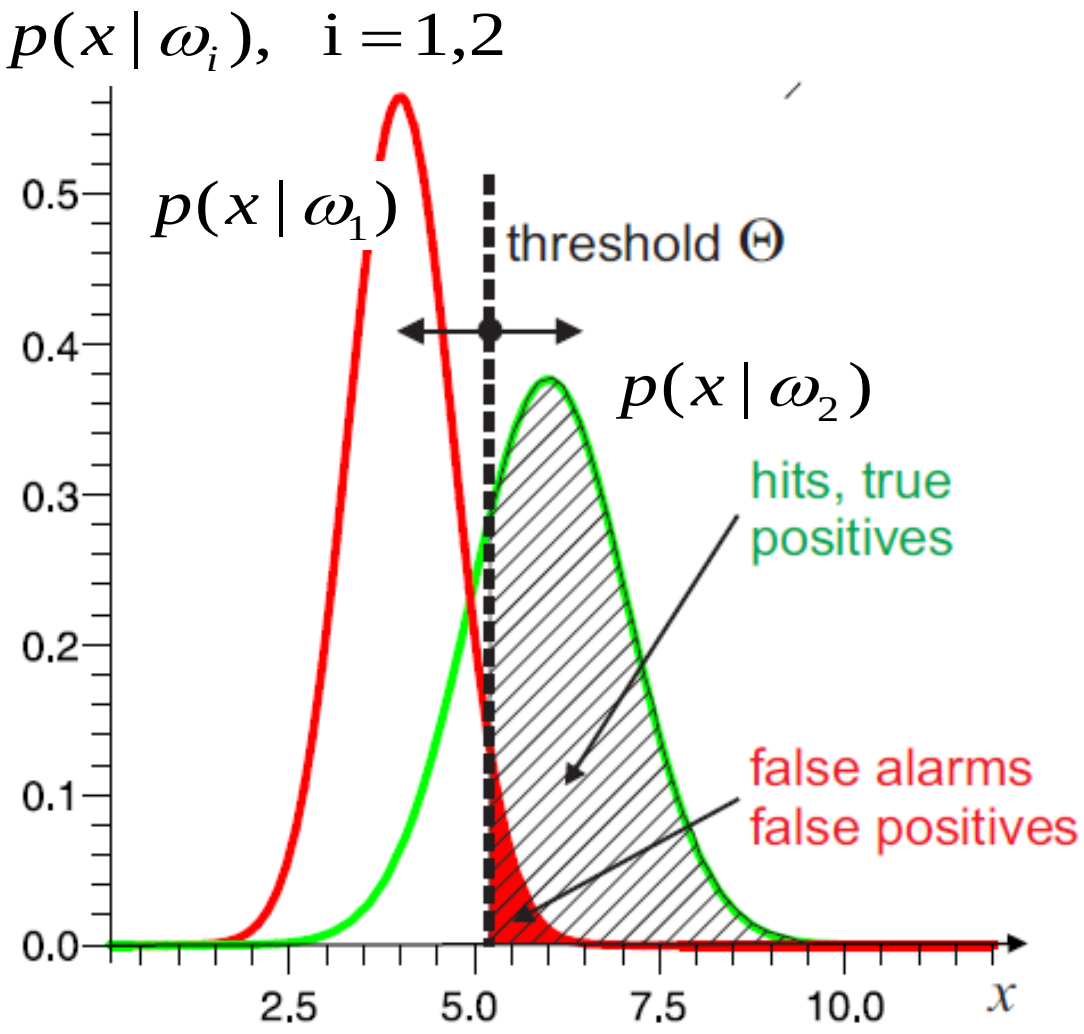
- In pratica, immagino di osservare come si distribuiscono i voltaggi per le due classi, trovando delle distribuzioni Gaussiane, o *approssimabili a Gaussiane*

ROC – assunzioni

- Il mio classificatore opererà quindi nel dover scegliere tra queste due classi, ipotizzando inoltre che
  - I valori medi $\mu 1, \mu 2$ e le deviazioni standard $\sigma 1, \sigma 2$ siano sconosciute
  - Possa modulare il valore della soglia Θ
  - Abbia a disposizione dei campioni di test etichettati (ossia conosco il ground truth, la vera etichetta)

ROC – quantità in gioco

- A questo punto, trovandomi in un caso di classificazione binaria, posso usare la notazione della matrice di confusione, e trattare le seguenti quantita
  - Veri positivi (hit)
  - Falsi positivi (False alarms)
  - Falsi negativi (Miss)
  - Veri negativi (correct rejection)



- Veri positivi (hit)

$$p(x > \Theta \mid x \in \omega_2)$$

- Falsi positivi (False alarms)

$$p(x > \Theta \mid x \in \omega_1)$$

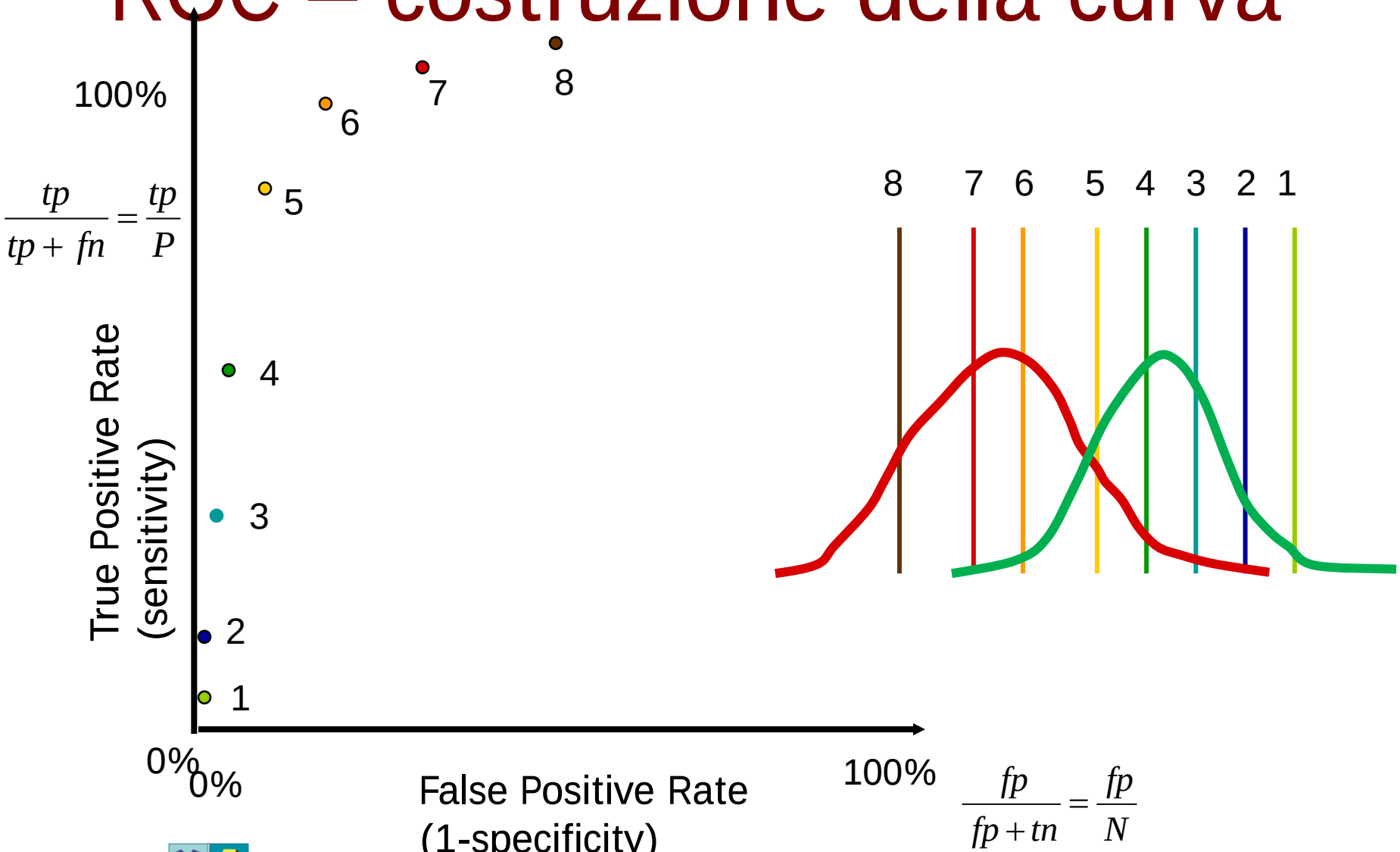
- Falsi negativi (Miss)

$$p(x < \Theta \mid x \in \omega_2)$$

- Veri negativi (correct rejection)

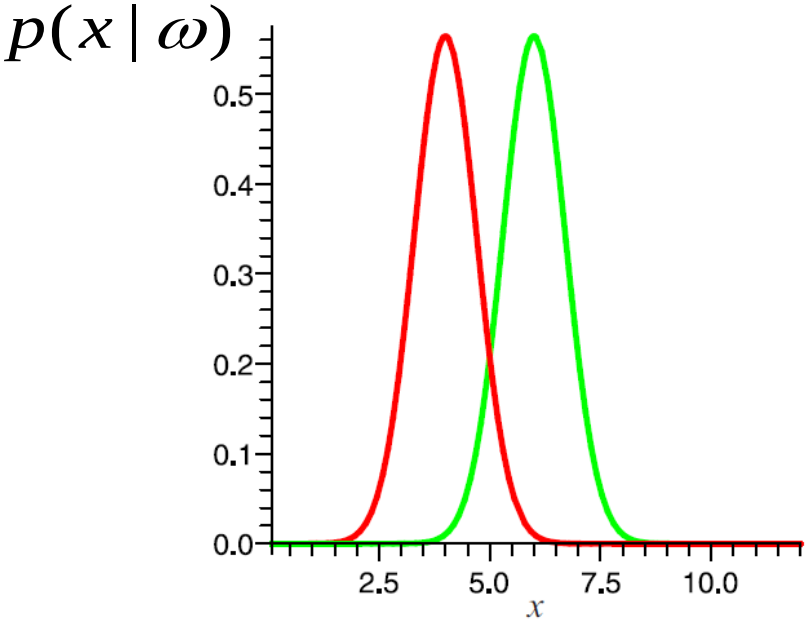
$$p(x < \Theta \mid x \in \omega_1)$$

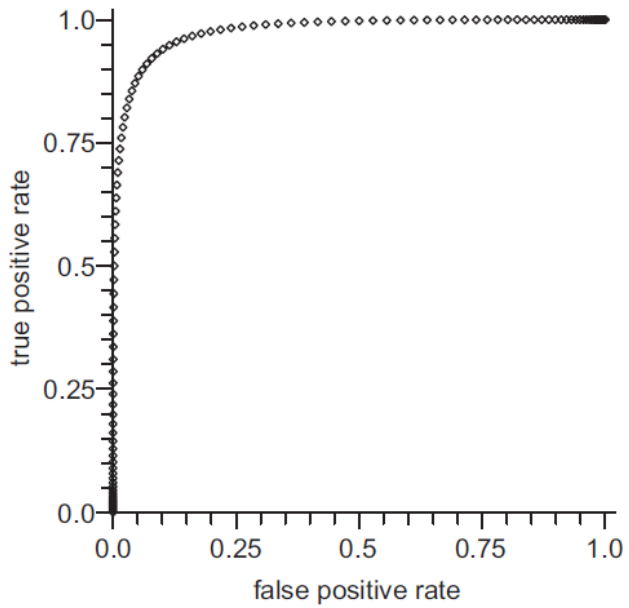
ROC – costruzione della curva



ROC - Esempi

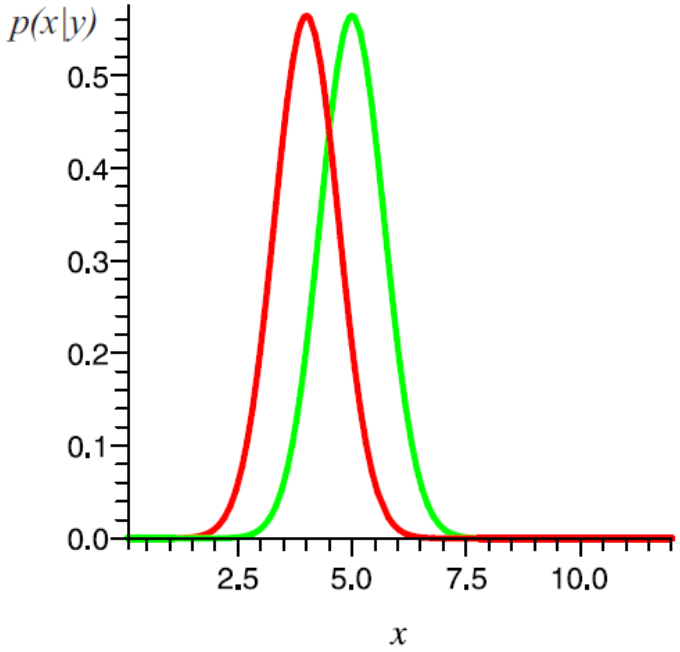
- Due classi Gaussiane, $\mu_1=4, \mu_2=6, \sigma_1=\sigma_2=1$
- Poco overlap, buona discriminabilità
- In questo caso speciale, ROC è convessa

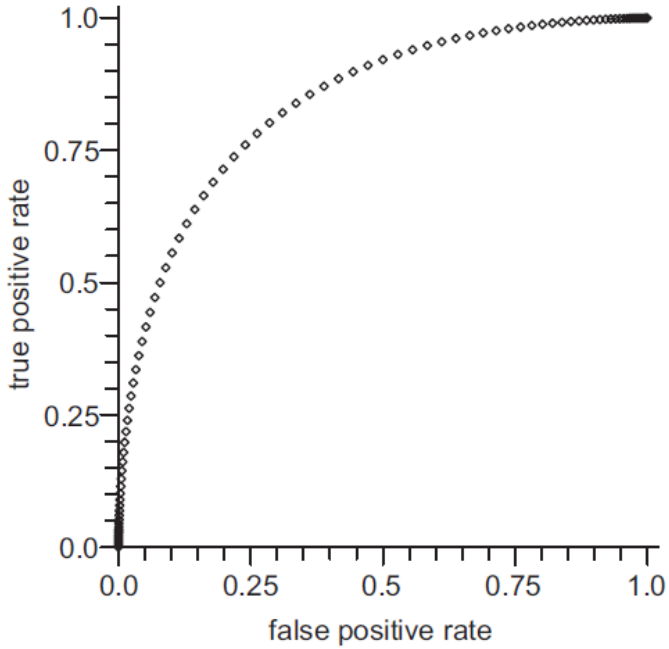




ROC - Esempi

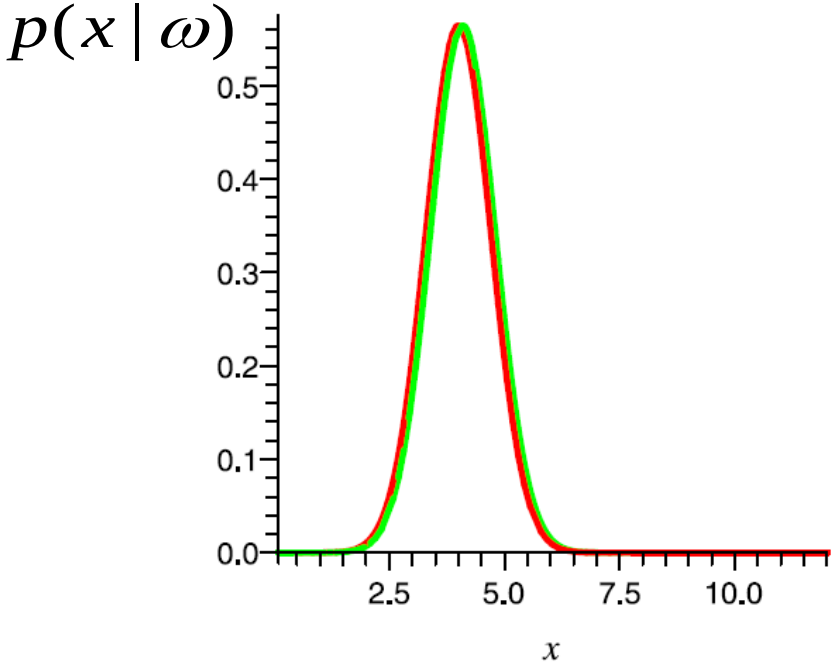
- Due classi Gaussiane, $\mu_1=4, \mu_2=5, \sigma_1=\sigma_2=1$
- Più overlap, minore discriminabilità
- In questo caso speciale, ROC è convessa

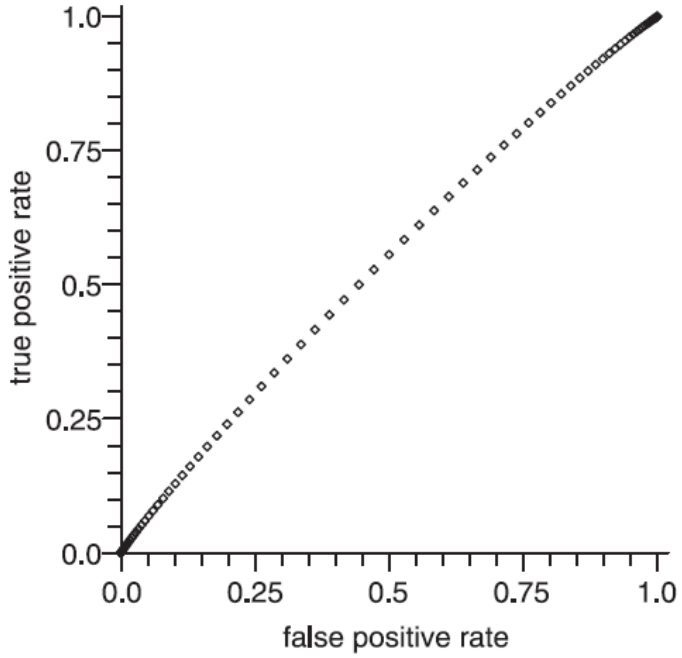




ROC - Esempi

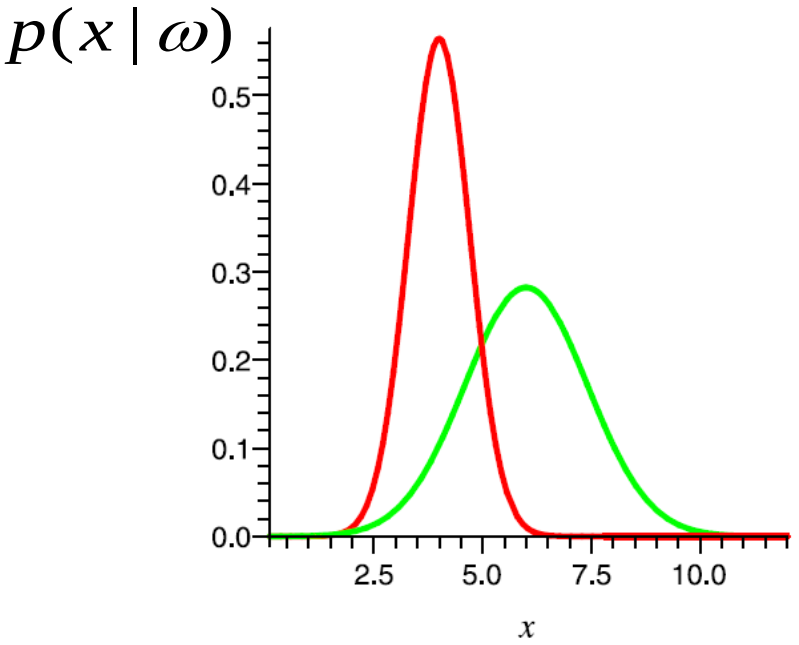
- Due classi Gaussiane, $\mu_1=4, \mu_2=4.1, \sigma_1=\sigma_2=1$
- Overlap quasi totale, no discriminabilità
- In questo caso speciale, ROC è ancora convessa

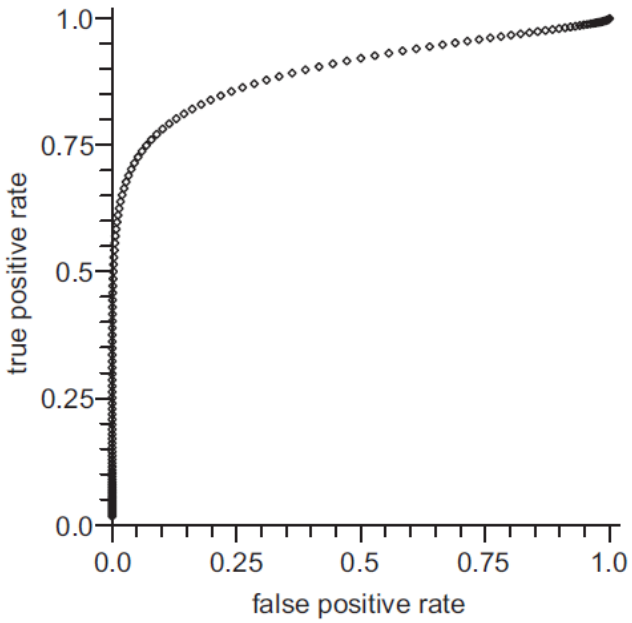




ROC - Esempi

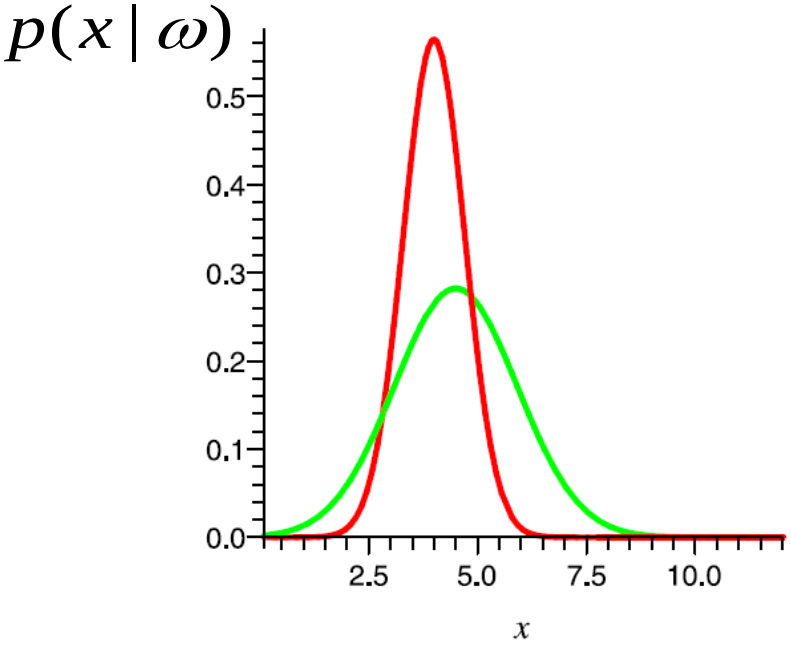
- Due classi Gaussiane, $\mu_1=4$, $\mu_2=6$, $\sigma_1=1$, $\sigma_2=2$
- Poco overlap, buona discriminabilità
- In generale, ROC non è convessa

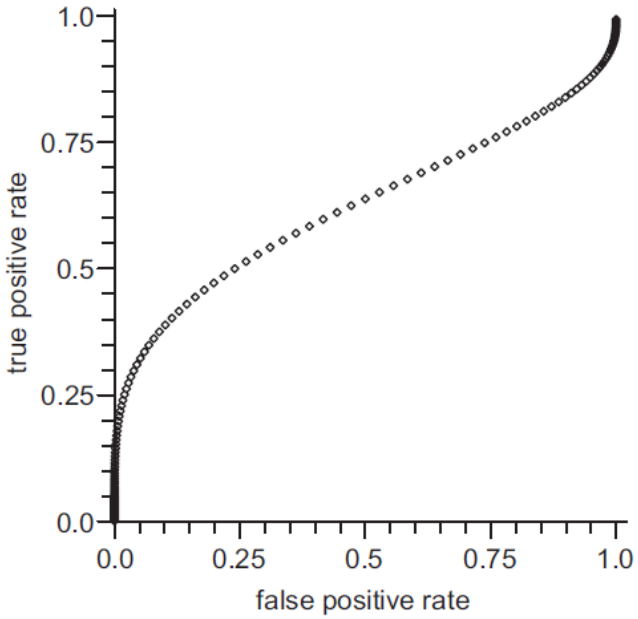




ROC - Esempi

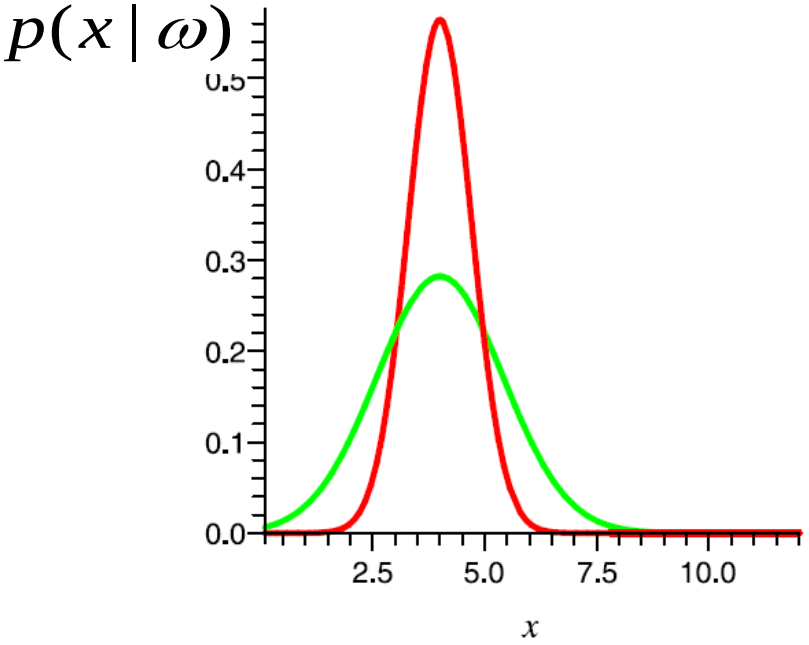
- Due classi Gaussiane, $\mu_1=4$, $\mu_2=4.5$, $\sigma_1=1$, $\sigma_2=2$
- Più overlap, meno discriminabilità
- In generale, ROC non è convessa

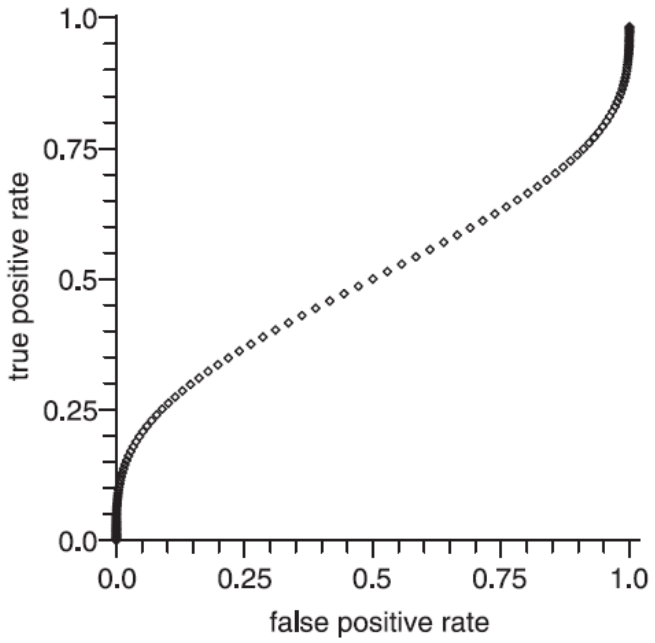




ROC - Esempi

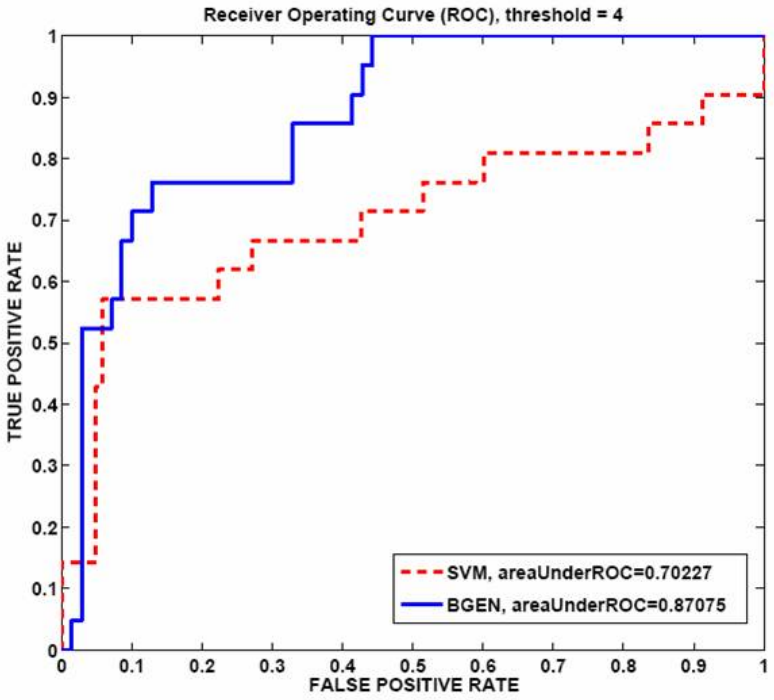
- Due classi Gaussiane, $\mu_1=4$, $\mu_2=4$, $\sigma_1=1$, $\sigma_2=2$
- Massimo overlap, cattiva discriminabilità
- In generale, ROC non è convessa





ROC – Curve di classificatori reali

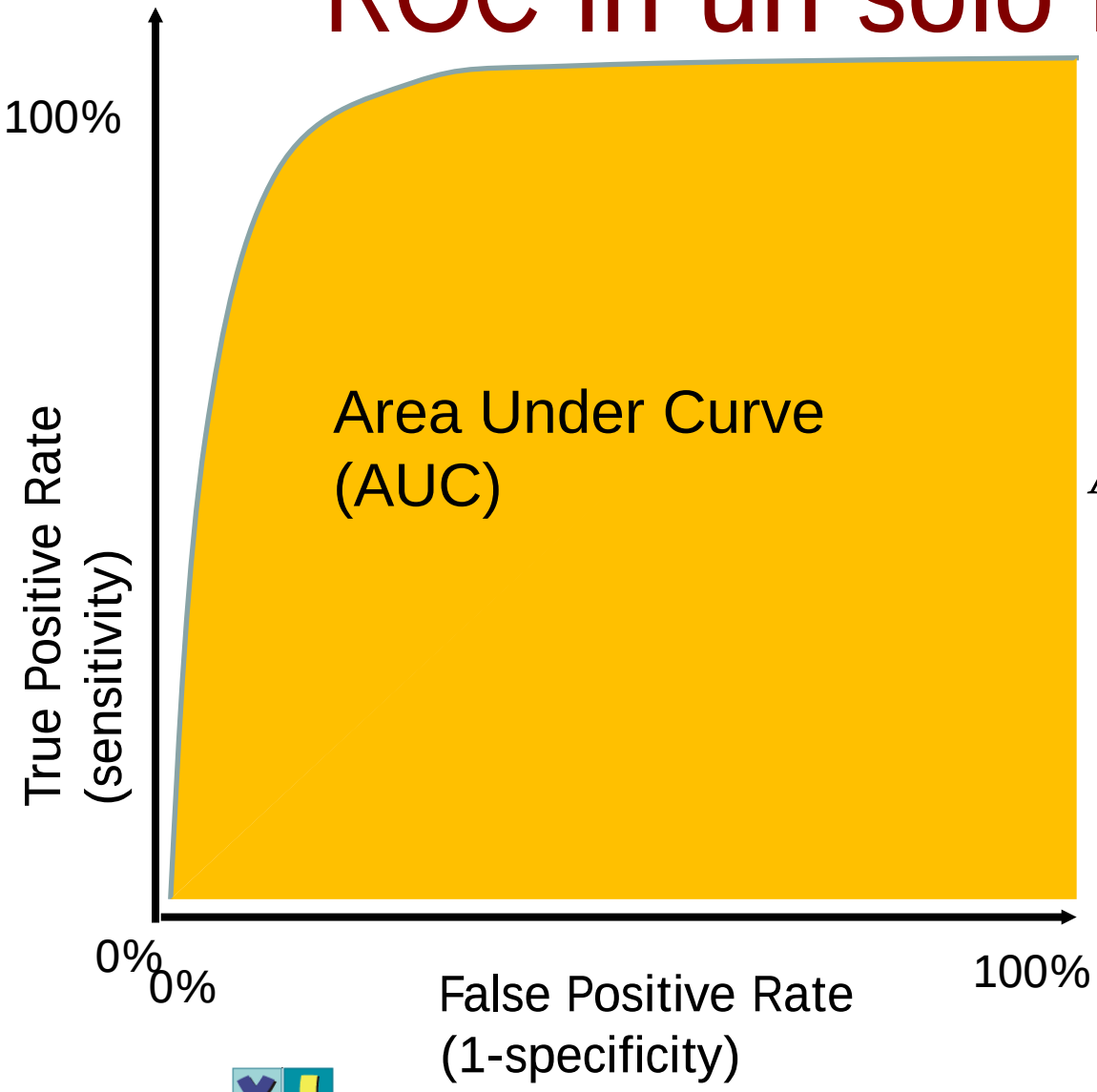
- Nel caso in cui le distribuzioni sottostanti non siano perfettamente Gaussiane, o ci siano pochi punti di valutazione, le ROC appariranno seghettate



ROC - Considerazioni

- Dato un classificatore binario, una ROC è un ottimo descrittore della sua qualità
- E' necessario pero' avere a disposizione il concetto di soglia Θ, ossia un valore che
  - identifichi una superficie di separazione (come nel caso precedente)
  - rappresenti una confidenza di appartenza ad una classe
- Come fare per riassumere una curva ROC in un numero?

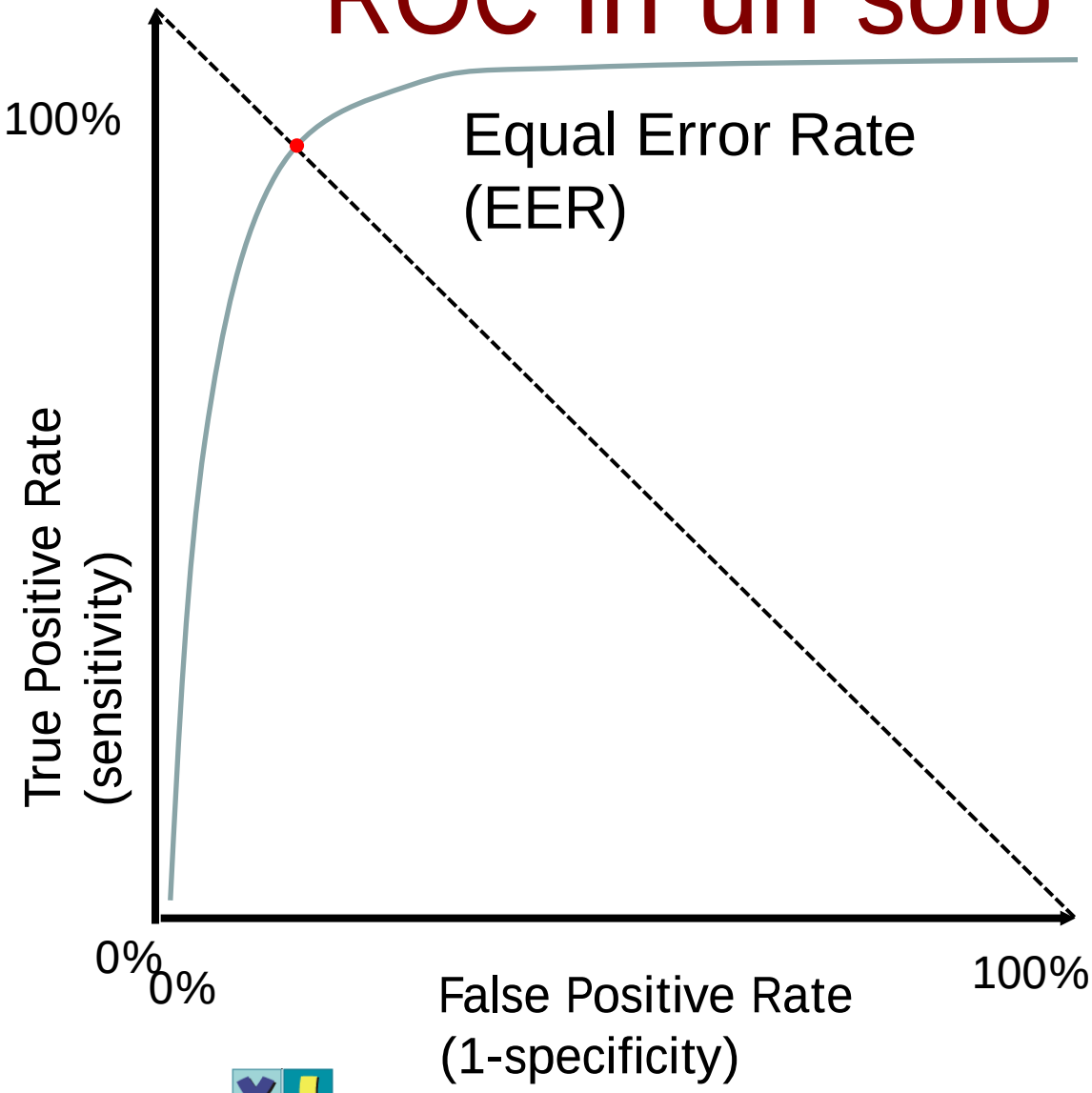
ROC in un solo numero



- Rule of thumbs:

$$AUC = \begin{cases} 0.5 \text{ no discrimin.} \\ 0.6 - 0.7 \text{ povera} \\ 0.7 - 0.8 \text{ accettabile} \\ 0.8 - 0.9 \text{ eccellente} \\ > 0.9 \text{ outstanding} \end{cases}$$

ROC in un solo numero



- EER: intersezione con la retta tratteggiata, ossia dove

FPR = 1-TPR = FNR

- *La proporzione di false accettazioni = proporzione di falsi rigetti*
- *Minore è migliore*

ROC – Best Operating Point

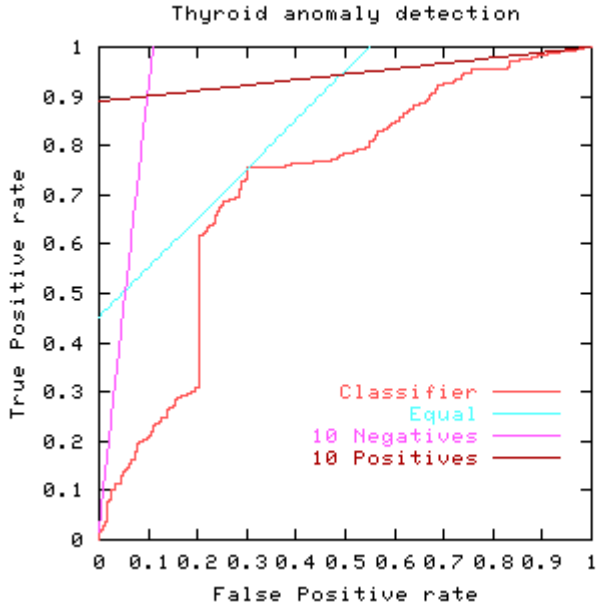
- Ogni punto della ROC rispecchia le sue performance di funzionamento data una particolare soglia Θ
- E' possibile quindi decidere quale sia la soglia Θ_{best} che mi determina le migliori performance?
- Si, ed è la soglia corrispondente al *Best Operating Point (BOP)*
  - Il BOP è un punto particolare della ROC, che rappresenta il miglior tradeoff tra sbagliare nel rivelare positivi VS il costo di generare falsi positivi

ROC – Best Operating Point (2)

- Se i costi di classificazione sono una semplice somma dei costi di misclassificare i positivi ed i positivi, allora *tutti i punti che giacciono su una retta* il cui gradiente è dato dall'importanza degli esempi positivi e negativi *hanno costo uguale*
- Se il costo per i positivi ed i negativi è lo stesso, ed abbiamo un egual numero di positivi e negativi, allora la retta ha pendenza 1, cioè 45 gradi

ROC – Best Operating Point (3)

- In questo caso, il miglior punto sulla ROC corrispondente alla migliore soglia è il punto che interseca la retta a 45 gradi più vicina al punto (0,1) dello spazio ROC.

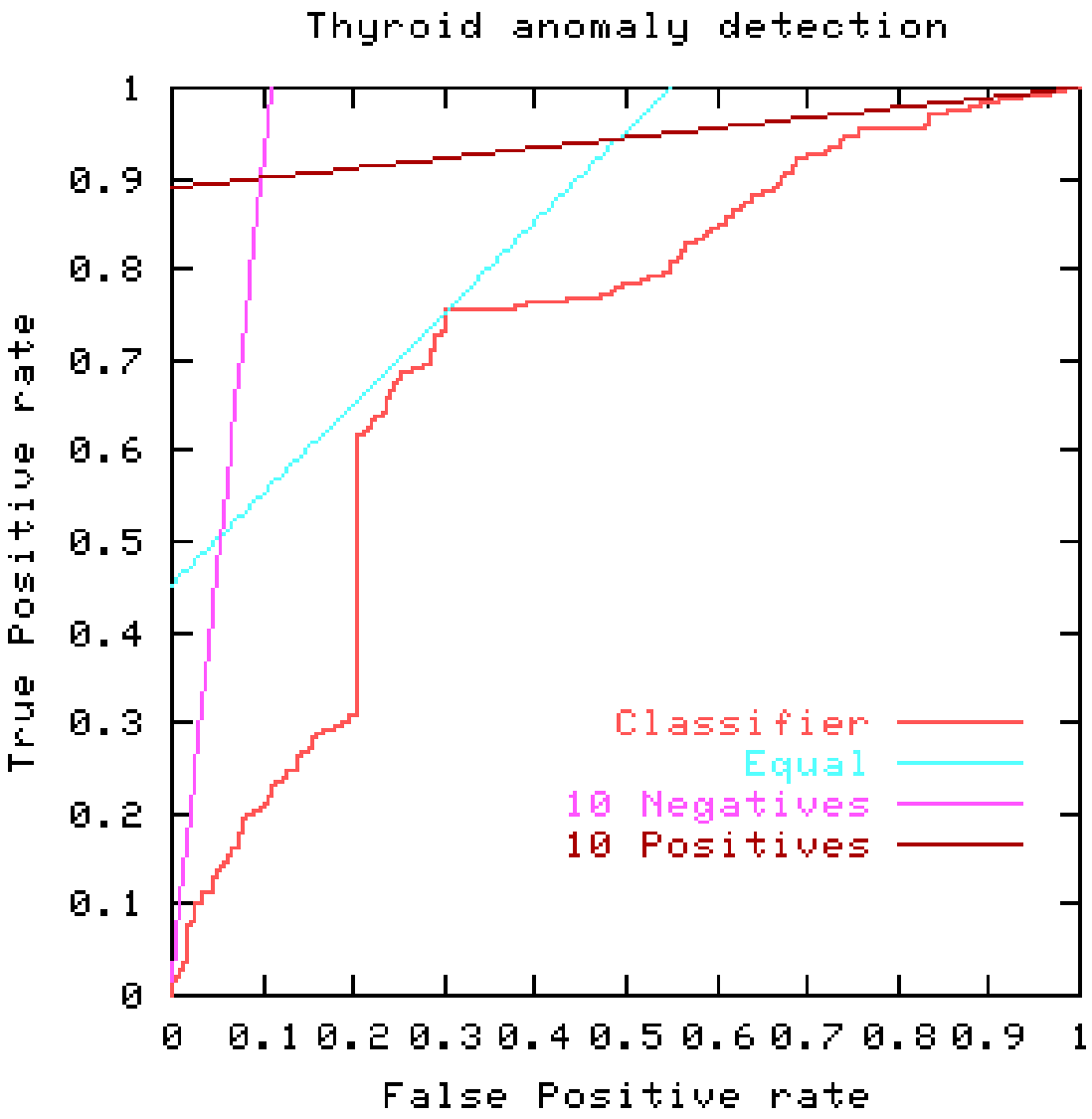


ROC – Best Operating Point (4)

- Siano
  - α = costo di un falso positivo (falso allarme)
  - β = costo di un falso negativo (miss)
  - p = proporzione dei casi positivi

  allora il costo medio di classificazione in un punto x,y della curva ROC è

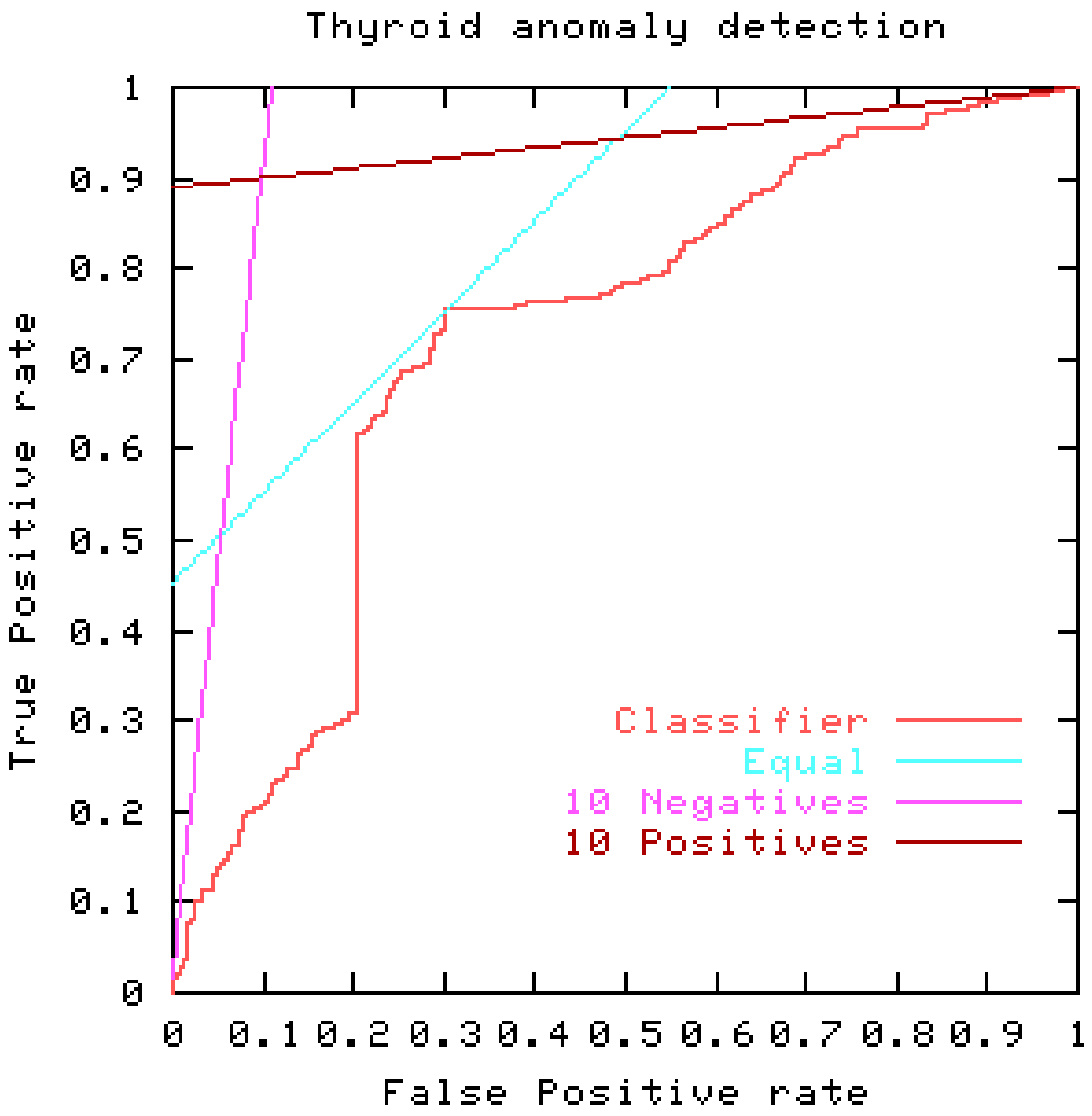
$$C = (1-p)\cdot\alpha\cdot x + p\cdot\beta\cdot(1-y)$$



- La linea ciano mostra la retta a costo minore intersecante la curva ROC quando
  - i costi di misclassificazione per i positivi e negativi sono uguali,
  - le proporzioni di negativi in termini di numero di campioni sono uguali

$$\alpha = \beta = 1$$

$$p = 0.5$$

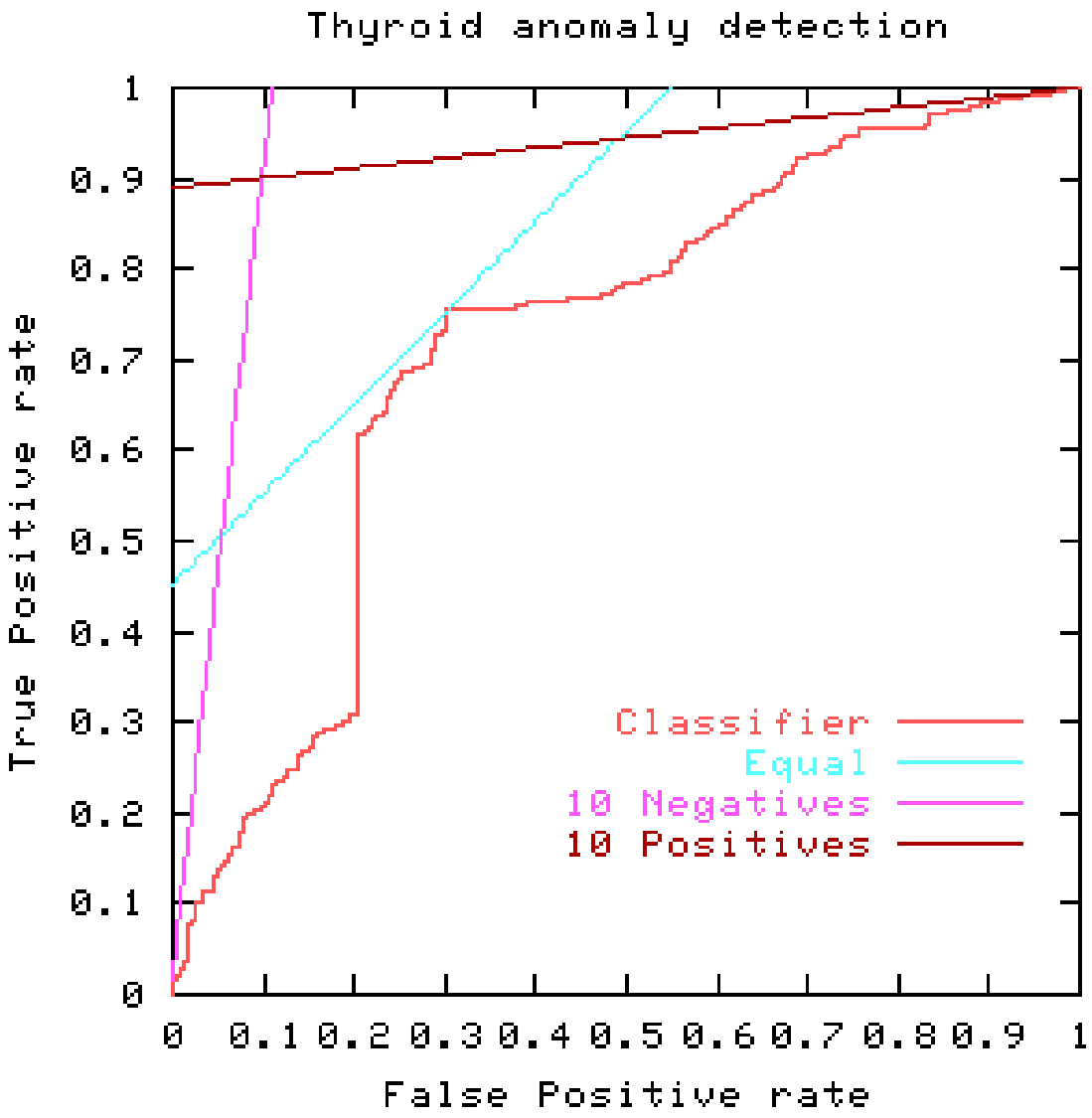


La linea viola corrisponde a quando il costo di perdere i negativi è 10 volte il costo di perdere i positivi

$$\alpha = 1$$

$$\beta = 10$$

$$p = 0.5$$



La linea marrone si presenta quando i costi di perdere un positivo sono dieci volte tanto i costi di generare un falso positivo

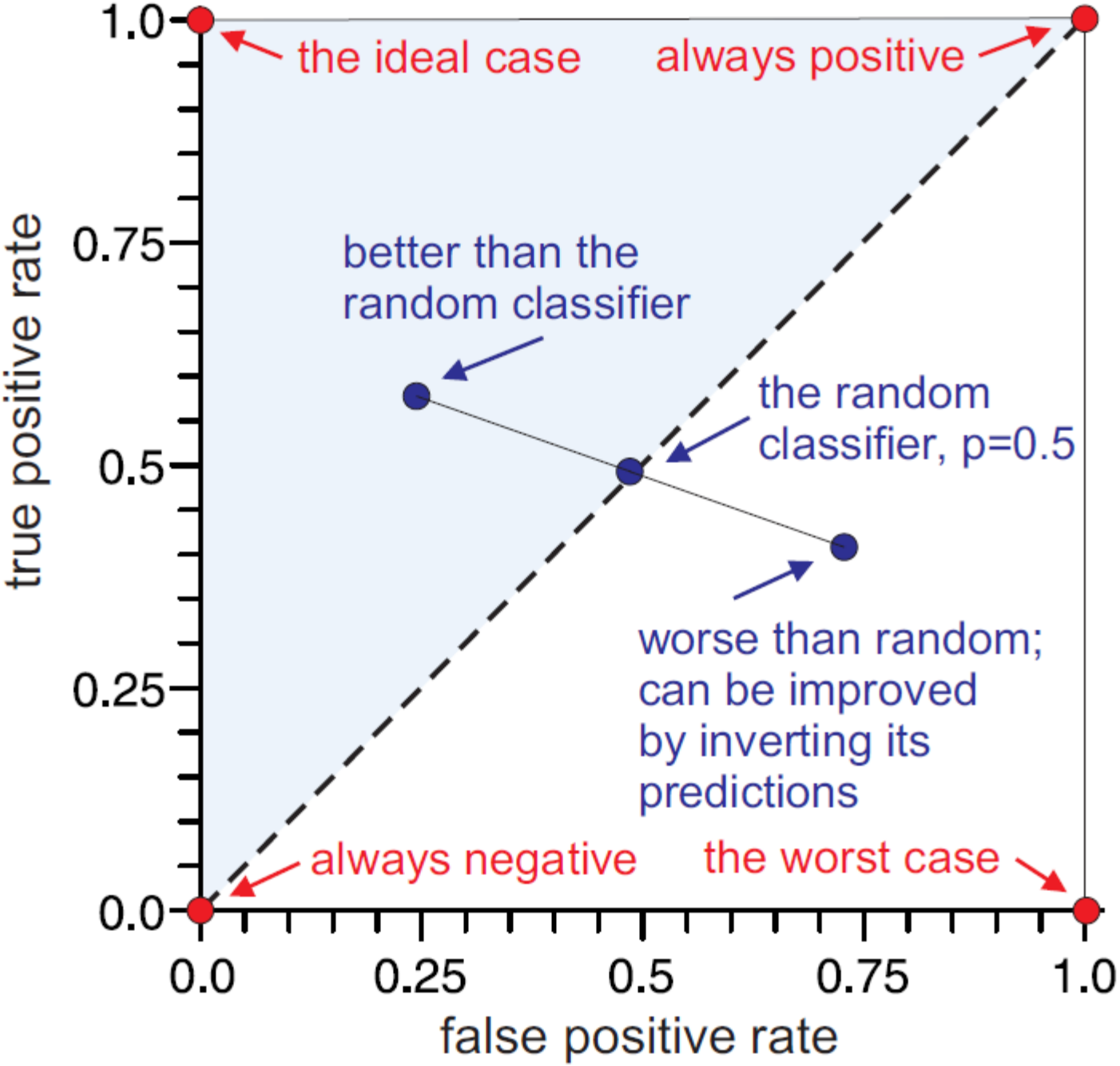
$$\alpha = 1$$

$$\beta = 10$$

$$p = 0.5$$

ROC per il confronto

- Dato un classificatore, supponiamo di non avere la possibilità di agire su una soglia
  - Perché il classificatore NON ha soglie
  - Perchè non abbiamo accesso al classificatore
- Abbiamo solo due valori puntuali, FPR e TPR
- Tale classificatore diverrà un punto nello spazio della ROC


1.0
0.75
0.5
0.25
0.0
true positive rate
the ideal case
always positive
better than the
random classifier
the random
classifier, p=0.5
worse than random;
can be improved
by inverting its
predictions
always negative
the worst case
0.0
0.25
0.5
0.75
1.0
false positive rate

ROC per il confronto - esempio

True	Predicted	
	pos	neg
pos	40	60
neg	30	70

True	Predicted	
	pos	neg
pos	70	30
neg	50	50

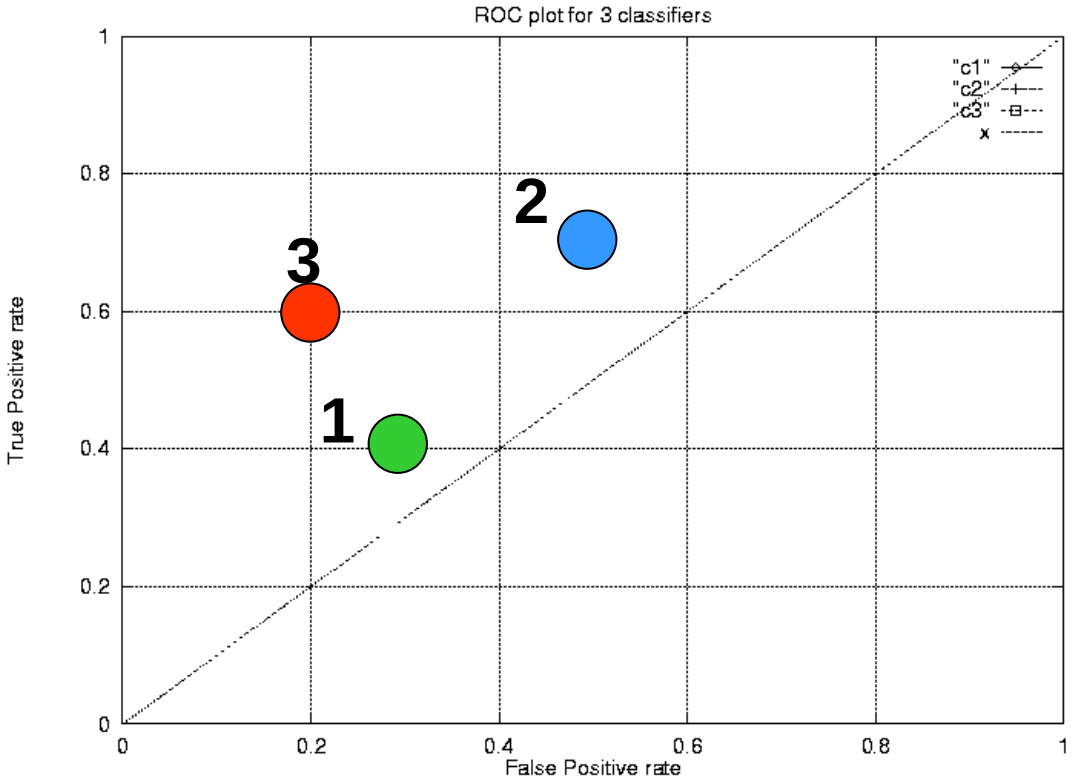
True	Predicted	
	pos	neg
pos	60	40
neg	20	80

Classifier 1
TPR = 0.4
FPR = 0.3

Classifier 2
TPR = 0.7
FPR = 0.5

Classifier 3
TPR = 0.6
FPR = 0.2

ROC per il confronto - esempio



- Quale scegliere? La dominante
  - Relazione di *dominio*

    Il classificatore A domina il classificatore B sse $TPR_A > TPR_B$ *and* $FPR_A < FPR_B$.

    - Nell'esempio, 3 domina 1
  - Altrimenti, posso scegliere a seconda delle mie esigenze (e dei costi di classificazione)

Curva Cumulative Matching Characteristic (CMC)

- In certi problemi di classificazione, l'output non è una singola classe, ma un *ordinamento* di classi
- Esempio: re-identificazione, soft-biometrics
  - Nella re-identificazione, dato un elemento di test (probe) raffigurante una persona incognita, voglio capire chi esso sia, considerando un dataset (gallery) di classi

Curva Cumulative Matching Characteristic (CMC)

- FONTE: Bolle, R.M.; Connell, J.H.; Pankanti, S.; Ratha, N.K.; Senior, A.W., "The relation between the ROC curve and the CMC," *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on* , vol., no., pp.15,20, 17-18 Oct. 2005

CMC – il problema di ranking

ω_1 ω_2 ω_3 ω_4

Gallery elements (or **classes**)

Probe **x**

- So che una delle classi è corretta, ossia effettua matching (*matcha*, è il *match corretto*) con il probe **x**
- In altre parole, all'interno della gallery c'è il soggetto che sto cercando

CMC – il problema di ranking

ω_1 ω_2 ω_3 ω_4

Gallery elements (or **classes**)

Probe **x**

- Invece che scegliere una classe vincente tramite un classificatore $p(\omega_i|\mathbf{x})$, voglio eseguire un ordinamento *decrescente* sugli *i, i=1,…,C*
- Chiaramente, se il match corretto è nelle prime posizioni del ranking, sono contento!

CMC – definizione generale

- Ho un set (grande) di campioni $x_i=B_i$, con *classe* associata $ID(B_i)$.
- Per costruire un problema di ordinamento, altrimenti detto *1:m search engine,* devo costruire due insiemi
  - Un gallery set $\mathcal{G}=\{B_1,B_2,\ldots,B_C\}$, dove *ogni* esempio ha associata una *classe* diversa
  - Un probe set $\mathcal{Q}=\{B'_1,B'_2,\ldots,B'_N\}$, in cui ogni elemento ha associata una *classe* in $\mathcal{G}$

CMC – costruzione

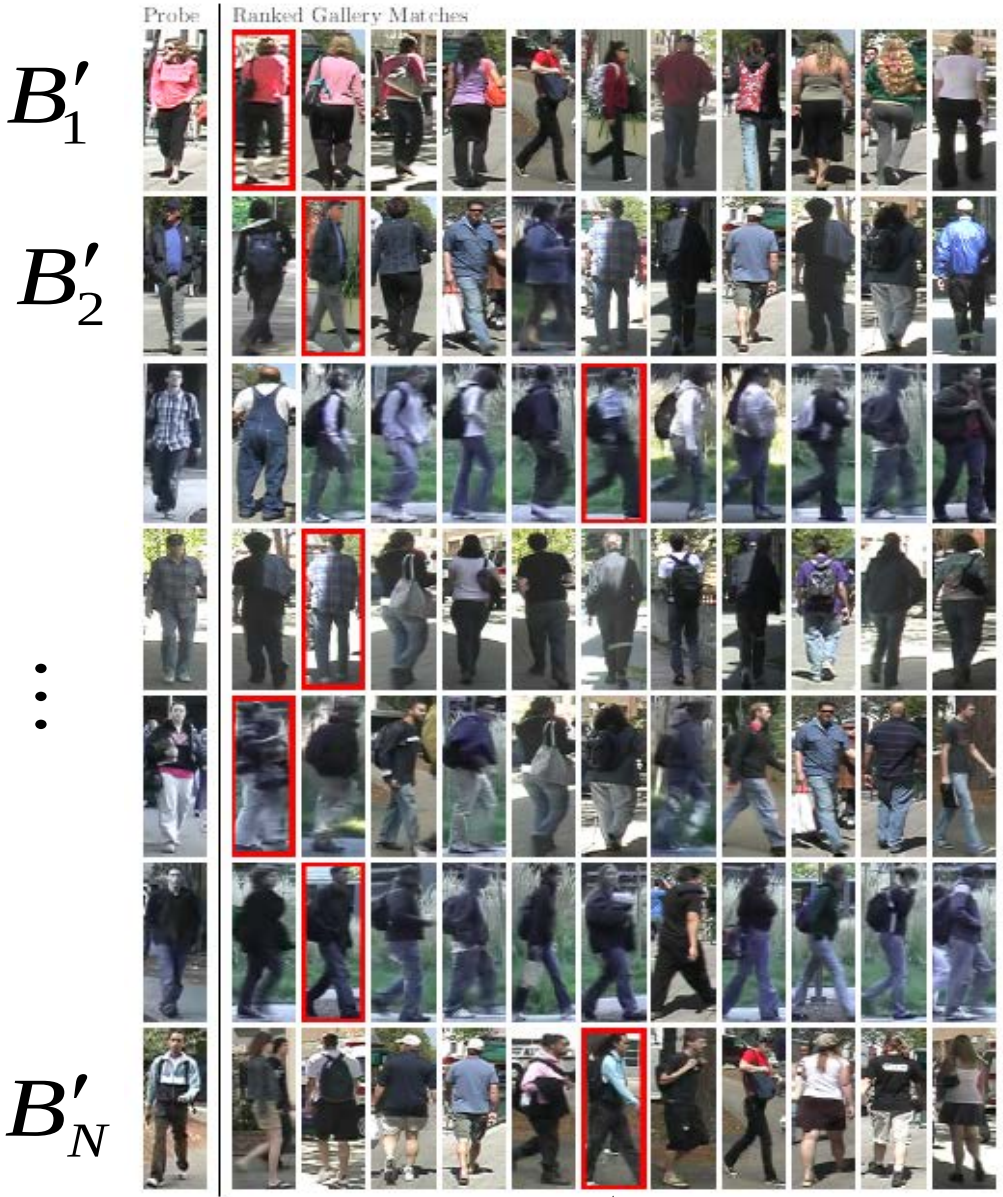
1. Dato un campione $B'_n \in Q$ ed un campione $B_m \in G$, calcolo un punteggio di similarità $s(B'_n, B_m)$ (analogo ad una probabilità a posteriori)
2. Eseguo questo per tutti i campioni $\in Q$, ottenendo una matrice di similarità S

$$S = \begin{bmatrix} s(B'_1, B_1) & s(B'_1, B_2) & \cdots & s(B'_1, B_C) \\ s(B'_2, B_1) & s(B'_2, B_2) & \cdots & s(B'_1, B_C) \\ \vdots & & \ddots & \vdots \\ s(B'_N, B_1) & s(B'_N, B_2) & \cdots & s(B'_N, B_C) \end{bmatrix}$$

CMC – costruzione (2)

3. Ordino ogni riga *n* della matrice in modo tale che
$$s\left(B'_n, B_{(1)}\right) \geq s\left(B'_n, B_{(2)}\right) \geq \cdots \geq s\left(B'_n, B_{(C)}\right)$$
dove $B_{(m)}$ indica un particolare elemento della gallery

- Al probe B'_n viene assegnato il rank $k_n=k$ se il campione di $\mathcal{G}$ è $B_{(k)}$
  - In pratica, $k_n=1$ significa che al primo posto del ranking ho la giusto campione della gallery, $k_n=2$ se il corretto campione si trovava in seconda posizione etc.


Probe
Ranked Gallery Matches
B'_1
B'_2
B'_N
$B_{(1)}$
$k_N=6$
$B_{(11)}\cdots$

CMC – costruzione (3)

- Ottengo così un set **K** di N rank $\{k_n; n=1,\ldots,N\}$ con $1 \leq k_n \leq C$

4. Definisco ora la *probabilità discreta di rank*

$$P(k) = \frac{1}{N}\left(\#k_n = k\right)$$

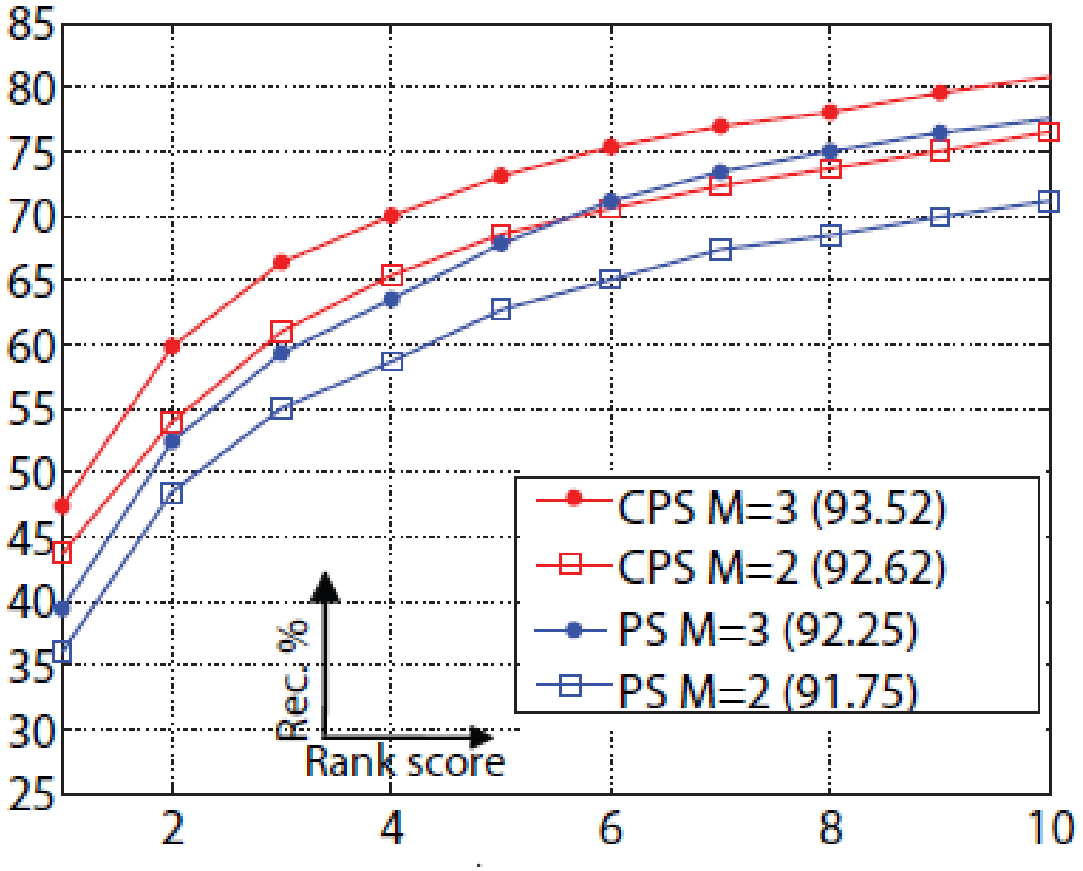
$$= \frac{1}{N}\left(\#k_n \in \mathbf{K} = k\right)$$

CMC – costruzione (4)

4. Posso costruire ora la CMC, ossia

$$\begin{aligned} \mathrm{CMC}(k) &= \frac{1}{N}\left(\#k_n \le k\right) \\ &= \frac{1}{N}\left(\#k_n \in \mathbf{K} \le k\right) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left(k_n \le k\right) \end{aligned}$$

CMC esempio



- Anche in questo caso, posso calcolare un numero riassuntivo della curva, ossia l'AUC (in figura tra parentesi)
- *Domanda sulla popolosità della gallery*

Misure di valutazione di un regressore

$$Y = f(X) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$$

$$Y \in \mathcal{R}$$

Mean Squared Error

- Dato il regressore ideale

$$Y = f(X) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$$

  vogliamo ricavare tramite training un modello $\hat{f}(X)$

- Il Mean Squared Error è

$$MSE = E\left[\left(Y - \hat{f}(X)\right)^2\right]$$

Mean Squared Error (2)

- Nel momento in cui ho N realizzazioni $\{x\}$ della v.a. X, ho

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{f}\left(x_n\right)\right)^2$$

dove $\{y\}$ sono le realizzazioni della variabile Y, o semplicemente risposte

Correlazione

- Utile nel momento in cui voglio avere una nozione di significatività statistica associata ai risultati

$\mathbf{y}$	$\hat{\mathbf{f}}(\mathbf{x})$
y_1	$\hat{f}(x_1)$
y_2	$\hat{f}(x_2)$
$\vdots$	$\vdots$
y_N	$\hat{f}(x_N)$

- Ho i miei valori ground truth e i miei risultati di regressione, ossia i vettori $\mathbf{y}$, $\hat{\mathbf{f}}(\mathbf{x})$
- Da accoppiare al MSE!!!

Correlazione - Pearson

- Calcolo il coeff ρ di Pearson tra i due vettori $\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})$

$$\rho\left(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})\right) = \frac{\mathrm{cov}\left(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})\right)}{\sigma_{\mathbf{y}} \sigma_{\hat{\mathbf{f}}(\mathbf{x})}}$$

- Il coefficiente varia tra -1 (max correlazione negativa) a 1 (max correlazione positiva). Se $\rho\left(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{x})\right) \approx 0$ non c'è correlazione
- Associato a ρ ho un p-value: se < 0.05 la correlazione è significativa, ossia non dovuta al caso

Paired t-test

- Utile nel momento in cui voglio capire invece che i risultati avuti dal regressore sono sicuramente sbagliati

$\mathbf{y}$	$\hat{\mathbf{f}}(\mathbf{x})$
y_1	$\hat{f}(x_1)$
y_2	$\hat{f}(x_2)$
$\vdots$	$\vdots$
y_N	$\hat{f}(x_N)$

- Di nuovo, ho i miei valori ground truth e i miei risultati di regressione, ossia i vettori $\mathbf{y}$, $\hat{\mathbf{f}}(\mathbf{x})$
- Da accoppiare al MSE!!!

Paired t-test (2)

- Eseguo un paired t-test tra i due vettori $\mathbf{y}$, $\hat{\mathbf{f}}(\mathbf{x})$
- L'ipotesi nulla mi dice che

$$\mathbf{y} - \hat{\mathbf{f}}(\mathbf{x}) \sim \mathcal{N}\left(0, \sigma^{unknown}\right)$$

  ossia che le due distribuzioni sono molto simili.
- Se il test restituisce 1, sono sicuro al 5% di significatività statistica (ossia p-value <0.05)che le due distribuzioni non sono simili, e che quindi la regressione non ha funzionato a dovere.

Bias e Varianza

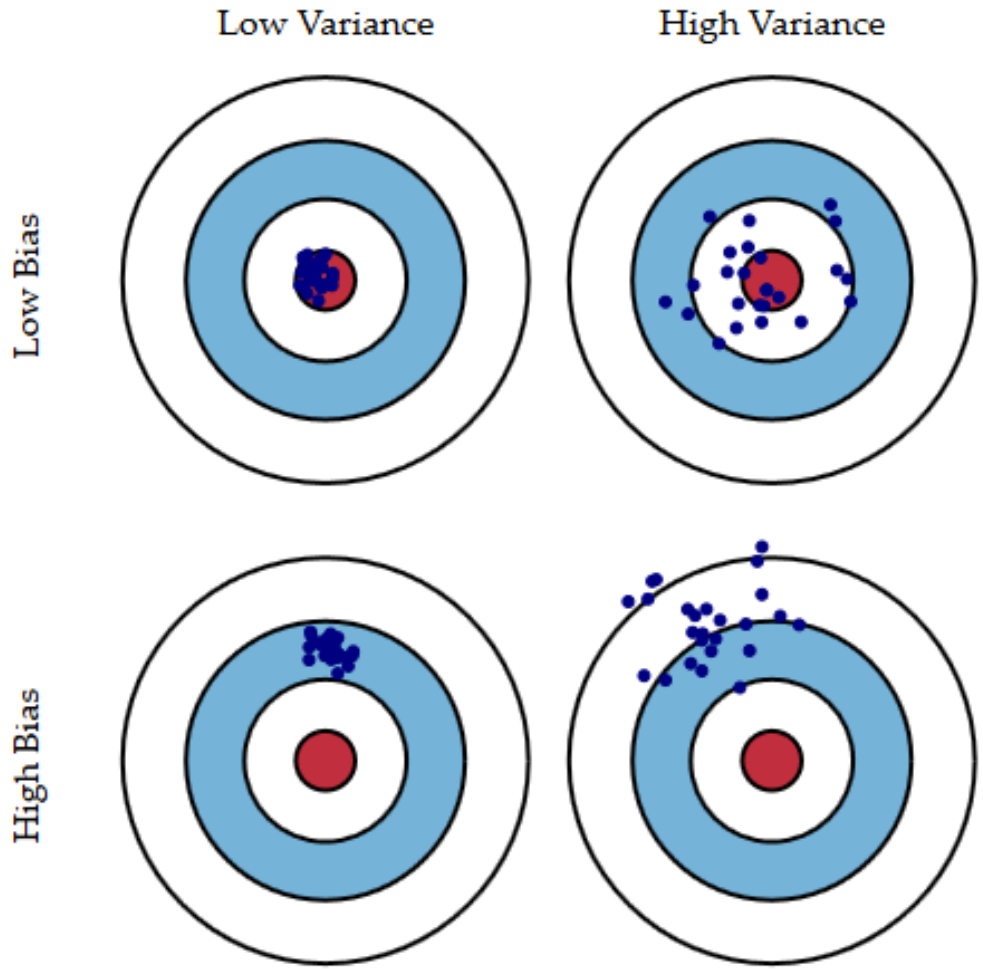
Bias e varianza

- Gli errori di predizione (cioè, l'error rate, ma più intuitivamente il MSE) possono essere decomposti in
  - Errore di bias
  - Errore di varianza
- Esiste un tradeoff per un modello di classificazione nel ridurre questi due tipi di errore
- Identificare correttamente questi errori aiuta a capire profondamente le performance di un classificatore, ed aiuta ad evitare *over* o *underfitting* (=non imparare nulla)
- Gli errori di bias e varianza possono essere definiti secondo tre punti di vista: 1)Concettuale 2)Grafico 3)Matematico

Bias e varianza – def. concettuale

- Errore di **bias**
  - La differenze tra la *predizione media* del nostro modello (ossia la predizione valutata addestrando su training set diversi il nostro classificatore) ed il *valore corretto che vogliamo predire*
- Errore di **varianza**
  - E' definito come *la variabilità nella predizione* che fa un modello su un particolare campione di test. Anche in questo caso, la variabilità si può catturare addestrando su training set diversi il nostro classificatore e dandogli lo stesso esempio da classificare

Bias e varianza – def. grafica



Bias e varianza – def. Matematica (specifica per un regressore)

- Denotiamo la variabile da predirre Y (un valore su cui fare regressione) e X la variabile visibile (le osservazioni).
- Possiamo assumere esista una relazione del tipo

$$Y = f(X) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$$

- f è pertanto il **classificatore ideale**
- L'errore irriducibile testimonia il fatto empirico che molto spesso non si incapsulare fenomeni con modelli

Bias e varianza – def. matematica

- Il nostro obiettivo è quello di stimare $\hat{f}(X)$ quanto più simile a $f(X)$ tramite le nostre tecniche di learning
- L'errore che produrro con la stime è il solito MSE

$$MSE = E\left[\left(Y - \hat{f}(X)\right)^2\right]$$

- Si può dimostrare che

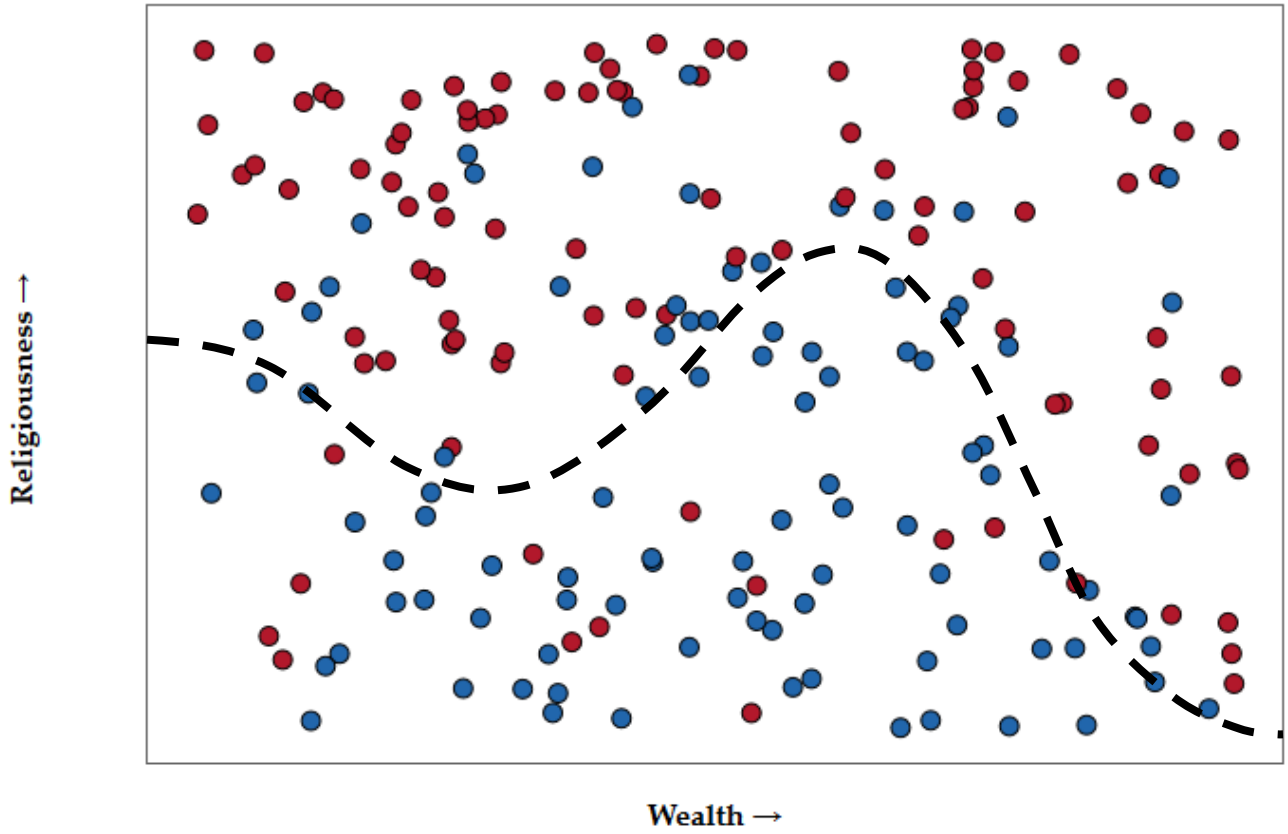
$$MSE = \left(E\left[\hat{f}(X)\right] - f(X)\right)^2 + E\left[\hat{f}(X) - E\left[\hat{f}(X)\right]\right]^2 + \varepsilon$$

BIAS^2 VARIANCE

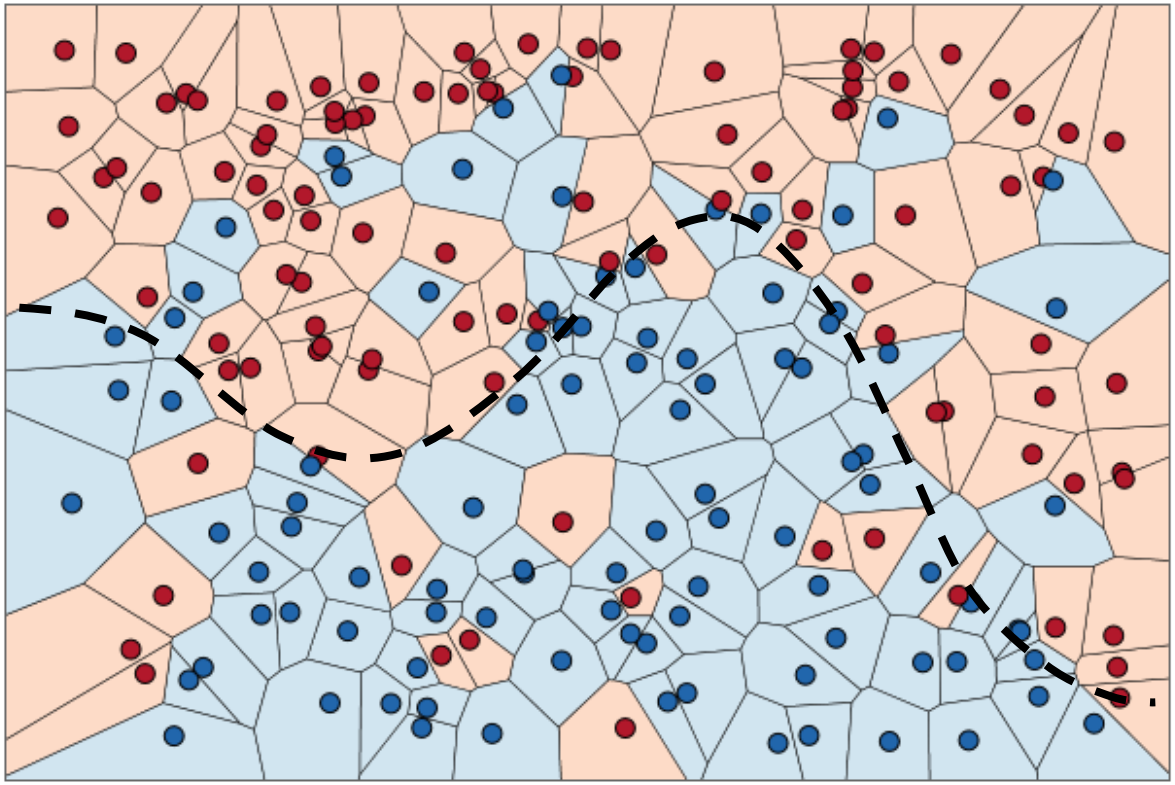
Bias e varianza – Risultato fondamentale

- **In un caso ideale, vorrei annullare i due errori; in pratica diminuendo uno aumenta l'altro (mentre ε rimane invariato)**
- Esempio: modellazione di un predittore per la percentuale di votanti un presidente USA repubblicano
  - Supponiamo di avere un training set di votanti con tre features:
    - Orientamento politico (su cui vogliamo fare classificazione)
    - Benessere
    - Religiosità

Bias e Varianza: esempio



- Come classificatore, uso knn (k=1)

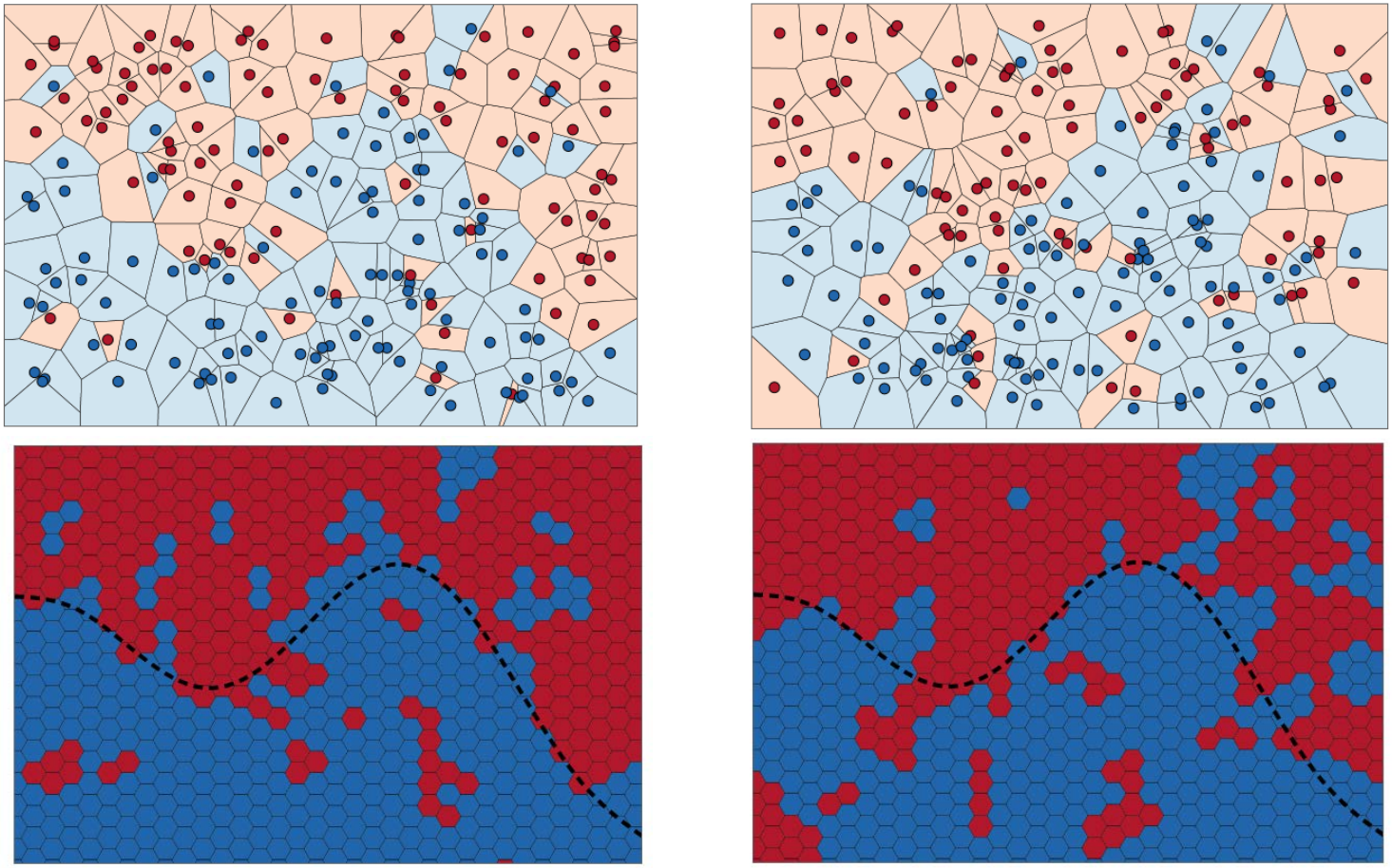


Bias e varianza

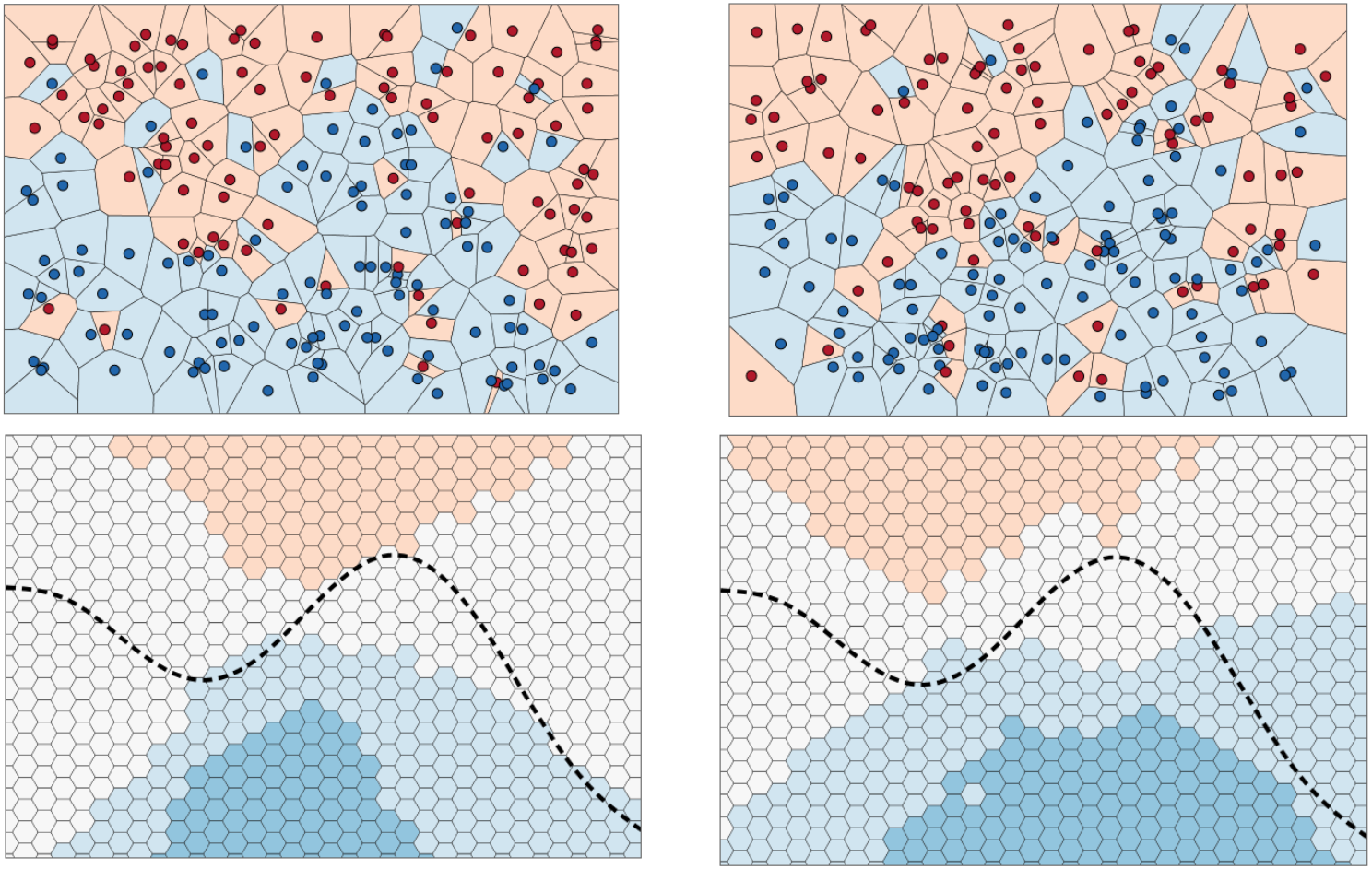
- Una scelta cruciale del classificatore è k
  - **Aumentando k**, aumento il bias, diminuisco la varianza
  - **Diminuendo k**, diminuisco il bias, aumento la varianza
  - Dimostrazione empirica:
    - Nei prox esempi, fissate una locazione…
    - demo a

http://scott.fortmann-roe.com/docs/BiasVariance.html

K=1



K=100



Bias O varianza?

- Quando mi addestro un classificatore, meglio minimizzare il bias o la varianza?
- *Innanzitutto, li devo poter osservare entrambi!!!*
- Per questo servono i principali metodi di testing
  - Hold-out
  - Cross-validation

Bias O varianza?

- In letteratura si riportano spesso i valori di accuratezza media, i valori di precisione media, etc… → ci si focalizza sul BIAS!!!
- Questo perche si assume che si abbiano molti campioni di test a disposizione (*long run assumption*) →credenza errata!
- E' necessario prestare attenzione anche alla varianza

Principi di testing

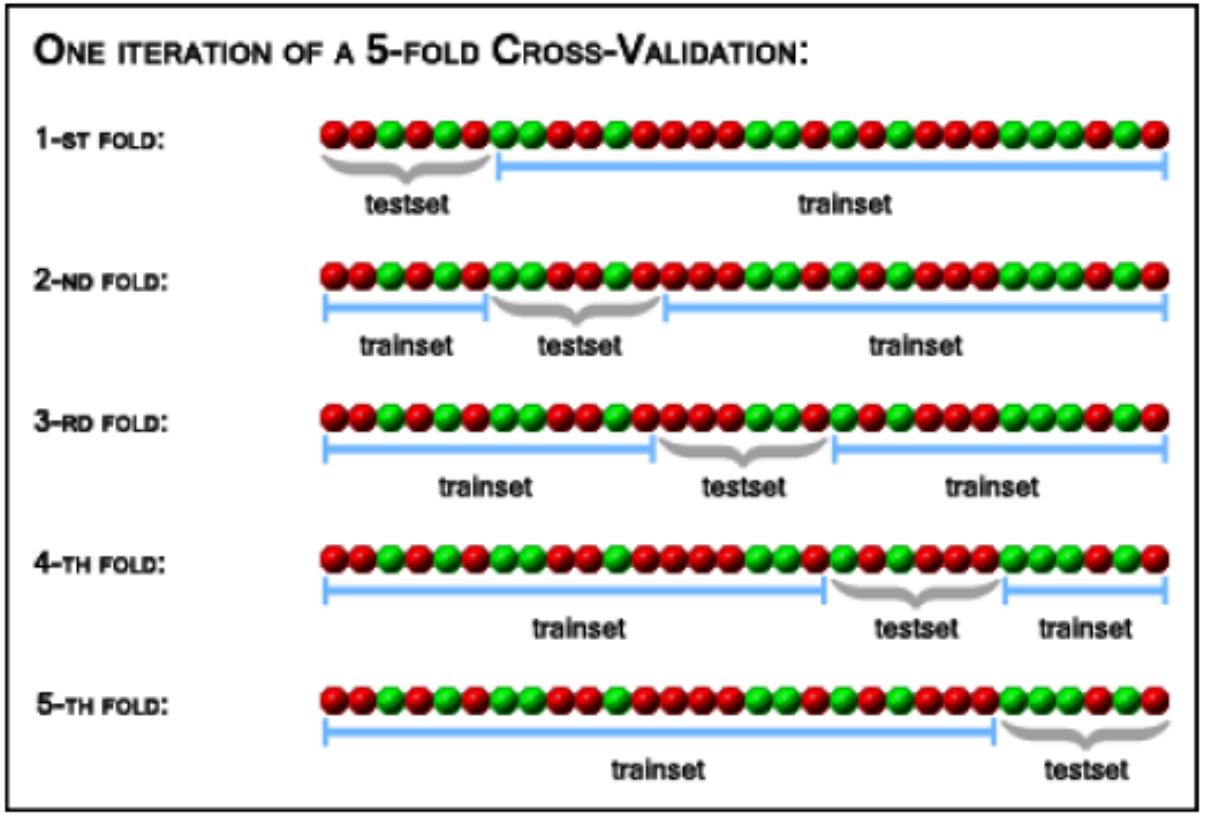
- Tutto parte da dati etichettati: li devo usare sia per fare il training che per fare testing
  - *Più dati di training* danno una maggiore generalizzazione (se essi rappresentano tutta la varietà assunta dalla classe in esame)
  - *Più dati di testing* danno una migliore stima di bontà del classificatore
- Sistemi di validazione principali:
  - Hold out
  - Cross validation

Hold out

- I dati etichettati vengono partizionati a random in training e testing, seguendo delle proporzioni note a priori
  - Training set (per esempio 2/3 dei dati totali) per addestrare il classificatore
  - Testing set (seguendo l'esempio di cui sopra, 1/3 dei dati)
- Una volta eseguito il testing, calcolo delle misure di bontà
- Ripeto k volte la procedura, medio le stime ottenute, calcolo la varianza
- (+) veloce; (-) posso usare alcuni campioni più volte (o mai) per il training che per il testing

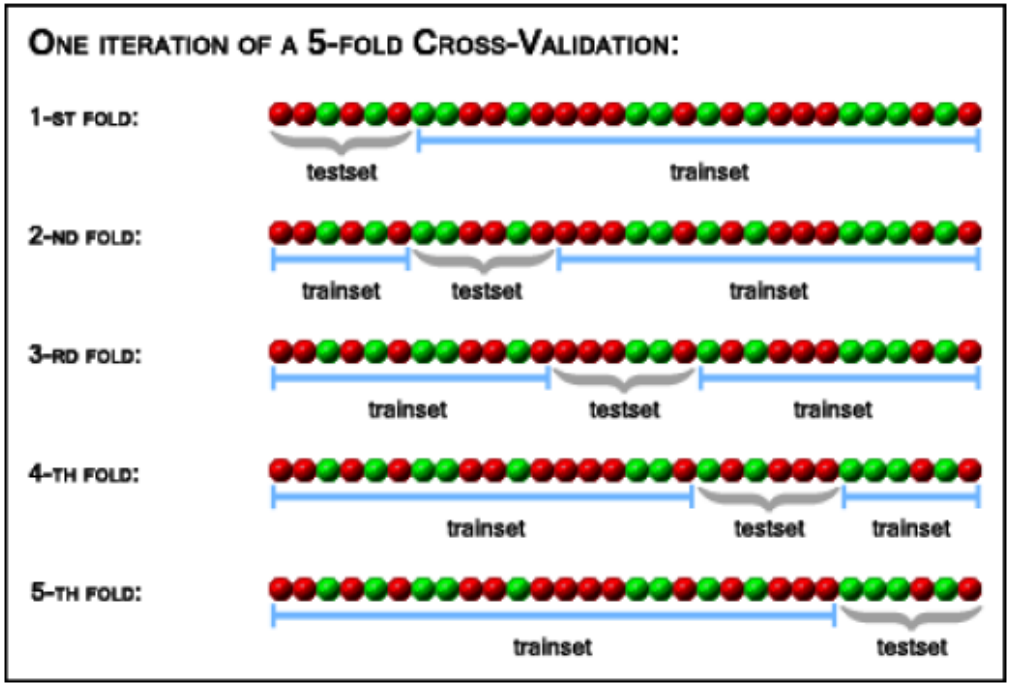
K-fold cross validation

- Il training set viene diviso *a caso* in K insiemi disgiunti di egual cardinalità, dove ogni insieme ha la stessa distribuzione di esempi per ogni classe



K-fold cross validation

- Il classificatore viene trainato K volte, ogni volta con un diverso insieme usato per il test
- Le misure finali di bontà vengono mediate sui K fold, e vengono estratte le varianze



K-fold cross validation

- (+) tutti i campioni vengono usati i maniera equilibrata per il ttraining e per il testing (-) più lento e laborioso da portare a termine

Leave One-Out (LOO)

- Un caso speciale di K-fold cross validazione in cui K=N con N numero totale di campioni presenti nel dataset etichettato
- Eseguo N esperimenti di classificazione usando n-1 campioni per il training, il rimanente per il testing
- Con LOO non posso garantire di avere nel training e ne testing set la stessa distribuzione di campioni presente nel dataset originale
  - Caso estremo: 50% classe A, 50% classe B. Il classificatore (stupido): dà come classificazione l'indice di classe maggiormente rappresentato nel training set. LOO dà come stima di errore (numero di casi mal classificati) il 100%…

Materiale aggiuntivo

- Il materiale aggiuntivo riguarda principalmente la curva ROC e la curva CMC
  - ROC analysis.pdf
  - ROC-CMC.pdf
  - rocHandout.pdf

Materiale aggiuntivo

- Per quanto riguarda bias e variance, suggerisco di guardare nei seguenti siti
  - http://scott.fortmann-roe.com/docs/BiasVariance.html
- Come approfondimento in generale su valutazione dei classificatori supervisionati
  - http://scott.fortmann-roe.com/docs/MeasuringError.html