Tecniche di riconoscimento statistico

Teoria e applicazioni industriali

Parte 2 – Teoria della decisione

Ennio Ottaviani
On AIR srl
ennio.ottaviani@onairweb.com
http://www.onairweb.com/corsoPR

A.A. 2018-2019

Impostazione del problema decisionale

- Consideriamo un problema con C classi ω_1 ... ω_C e siano α_1 ... α_C le corrispondenti decisioni su un oggetto (*pattern*) incognito X
- Supponiamo note le probabilità a priori $P(\omega_j)$ delle singole classi e le funzioni di costo $\lambda(\alpha_i,\omega_j)$, cioè quanto costa decidere per la classe i quando la classe vera è la j
- L'oggetto X è descritto da un vettore di *features* ad N componenti, distribuite in base ad una certa distribuzione di probabilità condizionata alla classe di appartenenza $p(X|\ \omega_j)$
- Supponiamo che queste probabilità condizionate (dette anche verosimiglianze) siano note o comunque stimabili sulla base di un *training set*

Principio MAP

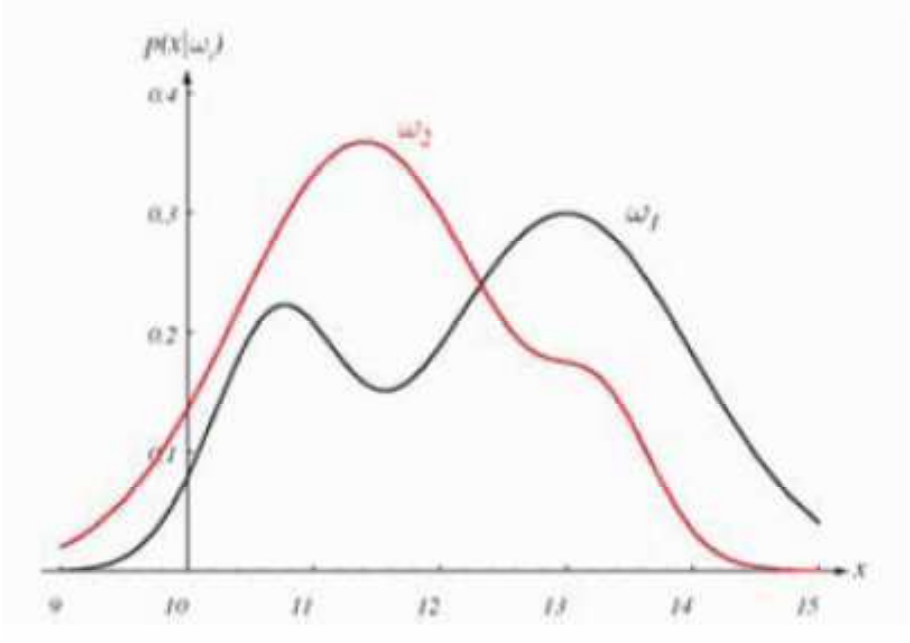
- Da un punto di vista puramente probabilistico, dato un pattern X, la decisione ottimale αj sarà quella che lo associa alla massima probabilità a posteriori $p(\omega_j|X)$
- La probabilità a posteriori $p(\omega_j|X)$ si calcola con il teorema di Bayes

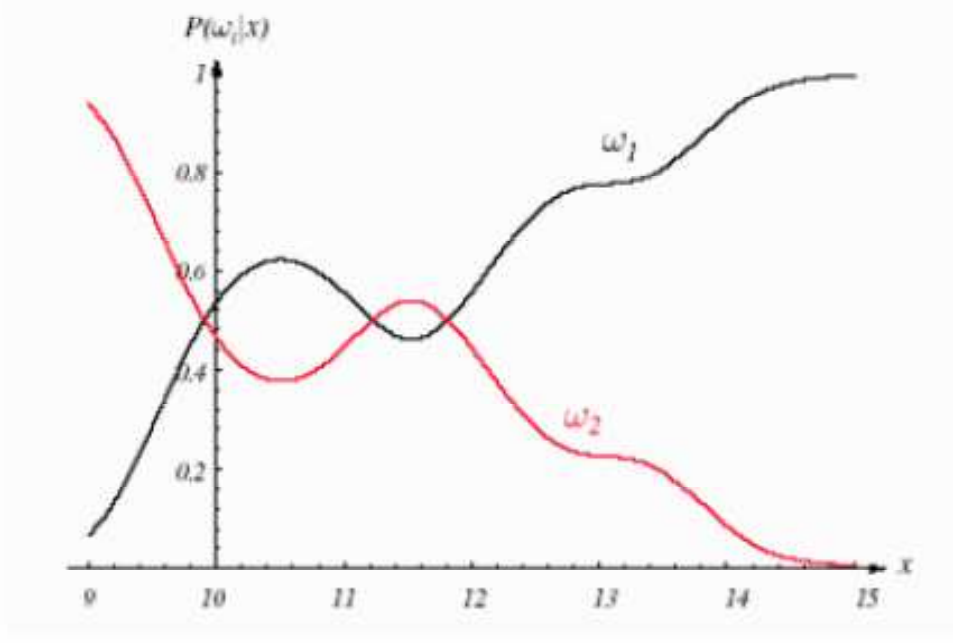
$$P(\omega_j|x) = \frac{p(x|\omega_j)\cdot P(\omega_j)}{p(x)} \qquad p(x) = \sum_{j=1}^{C} p(x|\omega_j)\cdot P(\omega_j)$$

- Le probabilità a posteriori è il prodotto (normalizzato) della probabilità a priori e della verosimiglianza
- Il principio MAP minimizza di fatto la probabilità di sbagliare (fatto dimostrabile se si assumono note tutte le d.d.p)

Esempio

- C= 2, P(1)=2/3, P(2)=1/3





Osservazioni

- Se le probabilità a priori sono uguali (classi equiprobabili), conta solo la verosimiglianza. Questo accade anche quando non sono note e neppure stimabili (assenza di informazioni a priori)

- Se le probabilità a posteriori sono uguali (per almeno 2 classi), significa che X non fornisce informazioni sufficienti per decidere in modo univoco

- Il termine di normalizzazione non ha alcun impatto sulla decisione e non viene quindi mai considerato

- La regola di decisione induce nel *feature space* un insieme di regioni di decisione

Errore bayesiano

- La probabilità di errore può essere scritta come

$$p(\text{error}) = \sum_{i=1}^{C} p(\text{error}|\omega_i)\,p(\omega_i) \qquad p(\text{error}|\omega_i) = \int_{\mathcal{C}[\Omega_i]} p(\boldsymbol{x}|\omega_i)\,d\boldsymbol{x}$$

$$p(\text{error}) = \sum_{i=1}^{C} \int_{\mathcal{C}[\Omega_i]} p(\boldsymbol{x}|\omega_i)\,p(\omega_i)\,d\boldsymbol{x}$$

$$= \sum_{i=1}^{C} p(\omega_i) \left(1 - \int_{\Omega_i} p(\boldsymbol{x}|\omega_i)\,d\boldsymbol{x}\right)$$

$$= 1 - \sum_{i=1}^{C} p(\omega_i) \int_{\Omega_i} p(\boldsymbol{x}|\omega_i)\,d\boldsymbol{x}$$

Errore bayesiano

- Applicando il principio MAP, per ogni x viene scelta la classe che massimizza $p(\omega_i)\,p(\boldsymbol{x}|\omega_i)$ in modo da minimizzare *p(error)*

- Ne segue che la minima probabilità di errore (errore bayesiano) vale

$$e_B = 1 - \int \max_i p(\omega_i)\,p(\boldsymbol{x}|\omega_i)\,d\boldsymbol{x}$$

- Se le distribuzioni di probabilità sono note, si potrebbe quindi calcolare il limite inferiore ottenibile per l'errore di classificazione (ma è un calcolo proibitivo)
- Esistono tecniche matematiche efficienti per ottenere buone stime del vero errore bayesiano

Caso binario

- Nel caso binario si possono fare due tipi di errore

$$\epsilon_1 = \int_{\Omega_2} p(\boldsymbol{x}|\omega_1)\, d\boldsymbol{x} \qquad \epsilon_2 = \int_{\Omega_1} p(\boldsymbol{x}|\omega_2)\, d\boldsymbol{x}$$

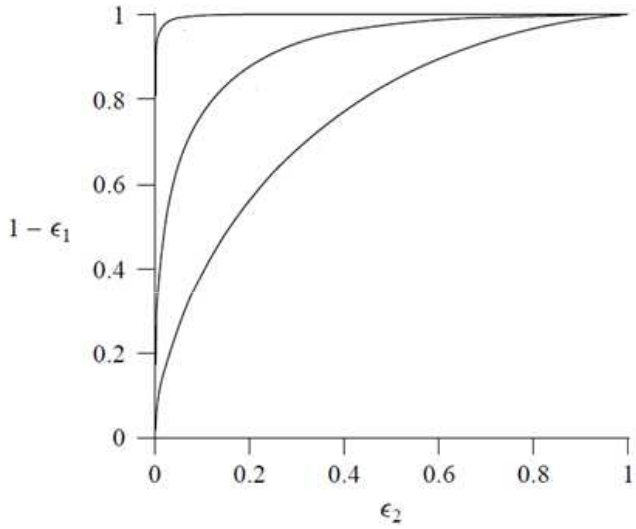
- L'errore bayesiano è una media pesata dei due

- Nel linguaggio tecnico, si intende di solito la classe 1 come classe di "segnale" e la classe 2 come classe di "rumore". Le relative probabilità vengono allora dette
  - probabilità di mancata detezione
  - probabilità di falso allarme

- I due tassi di errore vengono rappresentati nella curva ROC

Caso binario

- La curva ROC rappresenta tutte le possibili coppie di errori per un dato classificatore (punti di lavoro). Dipende solo dalle d.d.p.



- Il classificatore migliore sottende un'area maggiore (AUC)

- Per la scelta del punto di lavoro sono possibili diverse strategie (es. falso allarme costante, CFAR)

Caso binario

- Nella pratica, gli errori vengono valutati tramite la matrice di confusione

Stima-verità	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

- Valutazioni sintetiche si ottengono tramite i rapporti
  - Sensitività = VP / (VP+FN) → quanti casi P sono stati identificati correttamente
  - Predittività = VP / (VP+FP) → quanti casi identificati come P lo sono davvero
  - Specificità = VN / (VN+FP) → quanti casi N sono stati identificati correttamente

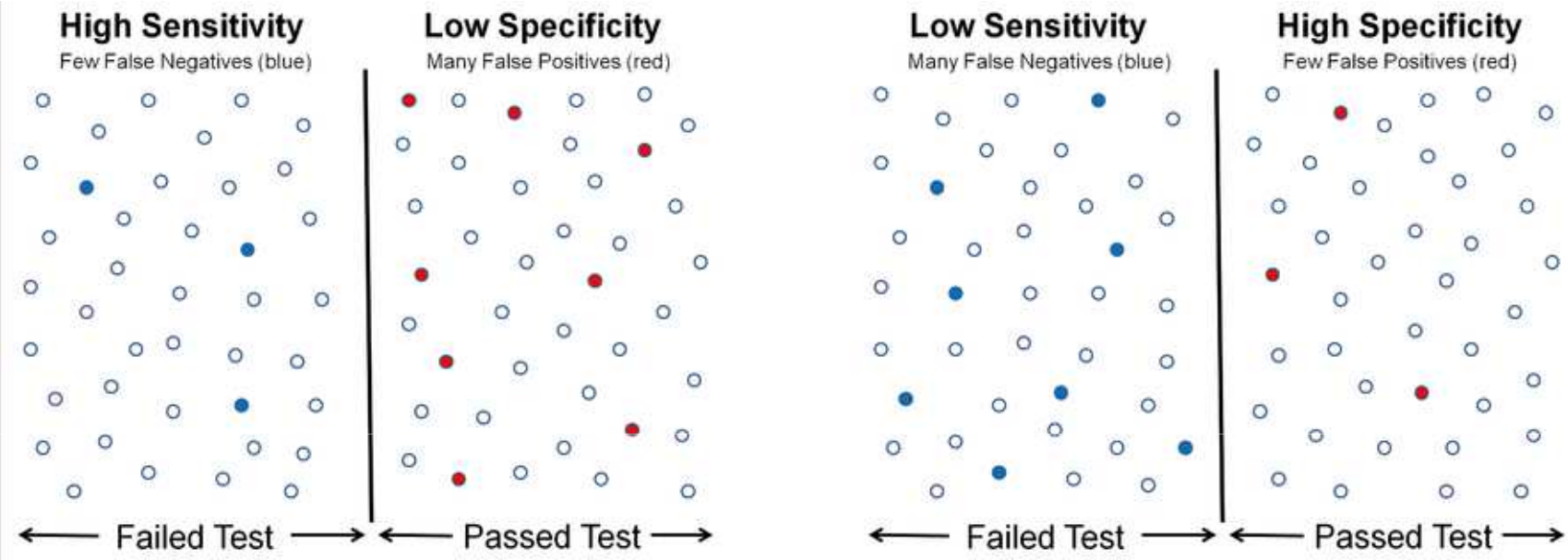
- Diversi domini applicativi usano diverse terminologie, per cui è facile fare confusione

Valutazione del test binario

- La valutazione di un test binario coinvolge un gran numero di possibili indici

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	True positive	False Negative (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	False Positive (Type I error)	True negative	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Esempio



- Un modo semplice per valutare il test è osservare da quale parte cade la maggior parte degli errori, cioè se sono più frequenti i falsi positivi (test poco specifico) o i falsi negativi (test poco sensibile)

- Questo però vale solo per popolazioni bilanciate!

Esempio

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	Condition negative	
Fecal occult blood screen test outcome	Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

F1-score

- Dovendo comparare un gran numero di classificatori in modo automatico, occorre un test che eviti le situazioni degeneri in cui precision o recall tendono a zero

$$F_1 \triangleq \frac{2}{1/P + 1/R} = \frac{2PR}{R+P}$$

- F1 è la media armonica di precision (P) e recall (R)

- Se P e R sono simili, lo è anche F1, ma se P (o R) tende a zero, anche F1 tende a zero
- Il massimo di F1 definisce un punto di lavoro bilanciato in cui i due tipi di errore hanno la stessa frequenza

Funzioni di rischio

- In presenza di costi di classificazione, la decisione ottima non può basarsi solo sul principio MAP ma deve tenere conto anche del costo degli errori

- Si deve scegliere la decisione αi che minimizza il rischio condizionale $R(\alpha_i|X)$

$$R(\alpha_i|x) = \sum_{j=1}^{C} \lambda(\alpha_i|\omega_j) \cdot P(\omega_j|x)$$

- Se i costi sono tutti noti, il criterio del minimo rischio è una ovvia conseguenza del calcolo bayesiano delle probabilità a posteriori per tutte le classi

Funzioni di rischio

- Se C=2, ci sono 4 costi λ da considerare e due quantità R

$$R(\alpha_1|x) = \lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x)$$
$$R(\alpha_2|x) = \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x)$$

- La condizione di minimo rischio diventa allora

$$\frac{P(\omega_1|x)}{P(\omega_2|x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}$$

- Ricordando il teorema di Bayes si può riscrivere ancora come

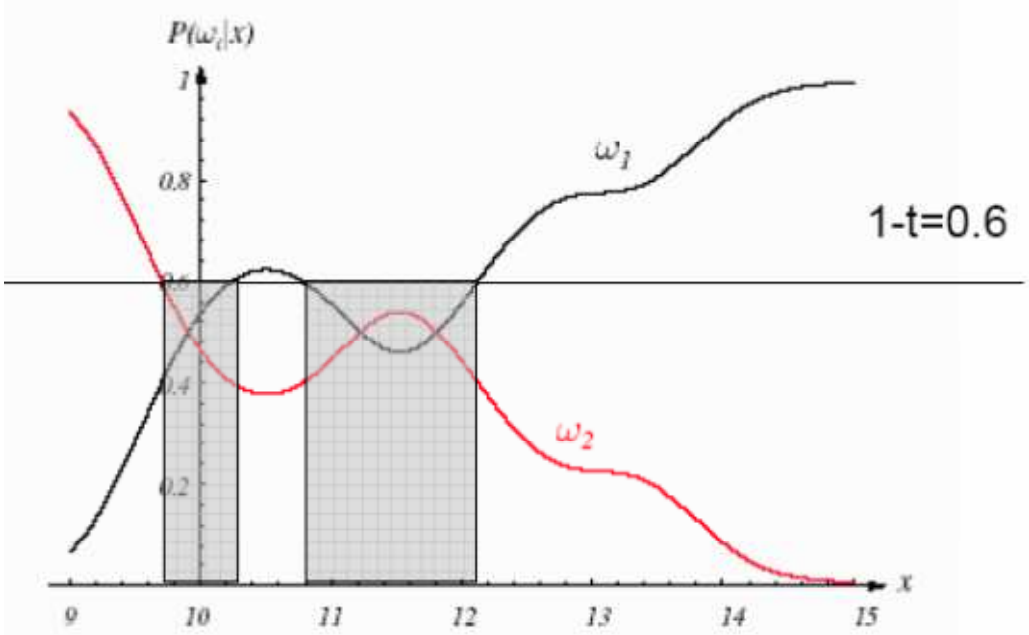
$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

- Questo test è detto *Likelyhood Ratio Test* (LRT)

Classificazione con scarto

- Nella pratica sono a volte possibili casi in cui anche la decisione ottimale produce una probabilità di errore (o un costo) inaccettabile

- Si ammette quindi la possibilità di scarto (o rigetto), cioè di decidere di non decidere. Se il costo dello scarto è inferiore al costo di un errore, la cosa può essere vantaggiosa



Funzioni discriminanti

- Un modo efficace di rappresentare un classificatore a C classi è tramite le funzioni discriminanti $g_i(X)$ i=1...C
- Un pattern X è assegnato alla classe ω_i se $g_i(X) > g_j(X)$ $j \neq i$
- Per il classificatore bayesiano $g_i(X)=p(X|\omega_i)P(\omega_i)$, ma la scelta non è unica. In generale ogni funzione monotona di $P(\omega_i|X)$ va bene (es. logaritmo)
- Date le g_i le regioni di decisione e le frontiere sono date da.

$$R_i(\mathbf{x}) = \left\{ \mathbf{x} \middle| g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall j \neq i \right\} \qquad \Gamma_{ij}(\mathbf{x}) = \left\{ \mathbf{x} \middle| g_i(\mathbf{x}) = g_j(\mathbf{x}) \; j \neq i \right\}$$

Distribuzioni gaussiane

- La teoria delle decisioni risulta particolarmente semplificata quando le distribuzioni di probabilità coinvolte sono di tipo gaussiano

- La distribuzione gaussiana è una ipotesi appropriata per $p(X|\omega_i)$ quando i vettori X generati possono essere pensati come una versione rumorosa di un unico prototipo μi

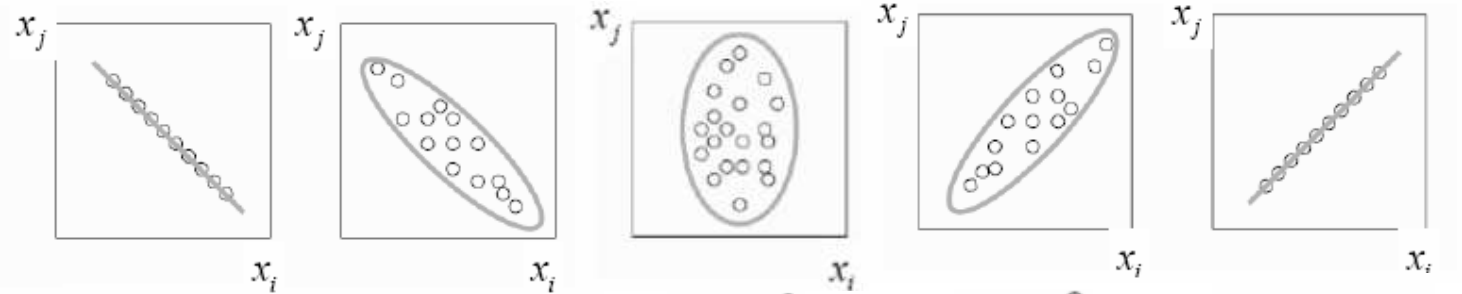
$$p(\mathbf{x}|\omega_i) = A_i \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

- Vettore medio $\boldsymbol{\mu}_i = E[\mathbf{x}|\omega_i]$ $\qquad A_i = \dfrac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}}$

- Matrice di covarianza $\boldsymbol{\Sigma}_i = E\left[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T|\omega_i\right]$

Proprietà della covarianza

- E’ una matrice simmetrica per costruzione
- Gli elementi diagonali sono le varianze delle componenti di X
- Gli elementi extradiagonali sono tale per cui $|\sigma_{ij}| \leq \sigma_i \sigma_j$
- Se due componenti i e j sono indipendenti allora $\sigma_{ij}=0$



Distribuzione dei dati

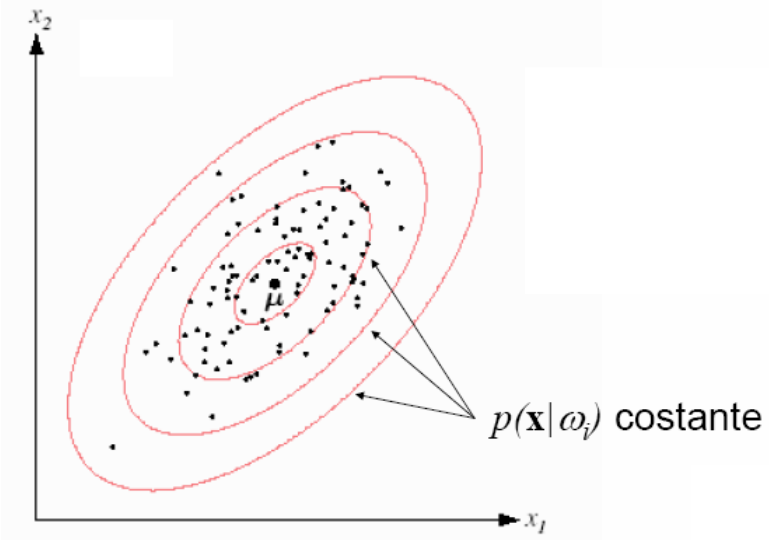
- I vettori estratti da una distribuzione gaussiana tendono a concentrasi in una “nuvola”

- Il centro della nuvola è il vettore medio, mentre la forma della nuvola è determinata dalla matrice di covarianza

- I vettori equiprobabili appartengono a curve su cui è costante la quantità

$$(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

- Essa è detta distanza di Mahalanobis



Classificatore quadratico

- Nel caso gaussiano, la forma più conveniente per le funzioni discriminanti è

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

- In termini di medie e covarianze, esse assumono la forma

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{d}{2}\ln 2\pi + \ln P(\omega_i)$$

- Le funzioni discriminanti sono quindi di tipo quadratico

- Il classificatore quadratico risulta quindi ottimale ogni volta che si può assumere che i dati siano distribuiti in modo gaussiano

- Medie e covarianza vengono di norma valutate tramite apprendimento statistico

Classificatore lineare

- Spesso si può assumere che la matrice di covarianza Σ sia unica, cioè non dipenda dalla classe. Omettendo i termini costanti si ha

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

- Sviluppando il termine quadratico e tralasciando le costanti si ha

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \qquad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

- Le funzioni discriminanti diventano semplici funzioni lineari di X (iperpiani). Questa è in assoluto la forma più semplice, ed anche la più usata, di funzione discriminante

Naive Bayes

- Un approccio intermedio tra quadratico completo e lineare è il classificatore bayesiano “naive”
- Si assume che le matrici di covarianza delle classi siano tutte diagonali, cioè si ignorano le correlazioni tra i dati
- L’approccio, per quanto semplice, fornisce spesso buoni risultati. Questo perché la decisione corretta si può spesso ottenere anche se le probabilità sono modellate in modo grezzo.
- Dal punto di vista geometrico equivale ad usare una distanza euclidea pesata al posto della distanza di Mahalanobis

Classificatore regolarizzato

- Il classificatore quadratico può essere di difficile addestramento per campioni squilibrati e/o poco popolati. Per tenere comunque conto delle correlazioni, si introduce un primo termine di regolarizzazione in base alla covarianza totale Sw

$$\boldsymbol{\Sigma}_i^{\lambda} = \frac{(1-\lambda)\boldsymbol{S}_i + \lambda \boldsymbol{S}}{(1-\lambda)n_i + \lambda n} \qquad \boldsymbol{S}_i = n_i \hat{\boldsymbol{\Sigma}}_i, \quad \boldsymbol{S} = n\boldsymbol{S}_W$$

- Un secondo termine di regolarizzazione modifica le singole covarianze “rinforzando” le componenti diagonali

$$\boldsymbol{\Sigma}_i^{\lambda,\gamma} = (1-\gamma)\boldsymbol{\Sigma}_i^{\lambda} + \gamma c_i(\lambda)\boldsymbol{I}_p$$

- Giocando su λ e γ si ottengono comportamenti intermedi tra quadratico, lineare e naive bayes

Misture gaussiane

- L'approccio bayesiano si generalizza facilmente a combinazioni convesse di probabilità generiche, del tipo

$$p(\boldsymbol{x}) = \sum_{j=1}^{g} \pi_j \, p(\boldsymbol{x}; \boldsymbol{\theta}_j)$$

- Il caso più comune è la mistura gaussiana (GMM o MOG). Ma mentre con *g=1* è facile ottenere stime dei parametri (media e covarianza), con *g>1* non esiste una soluzione diretta

- Esiste un algoritmo iterativo (Expectation-Maximization, EM) che a *g* fissato, data una partizione dei dati in *g* componenti, determina peso, media e covarianza di ogni componente (variabili latenti)

EM per misture gaussiane

- L'algoritmo EM pò essere descritto sinteticamente con una sequenza di passi, basati sulla stima di una matrice w_{ij} che misura la probabilità di appartenenza del i-esimo dato alla j-esima componente della mistura

- Inizializzazione: si assegna un valore iniziale ai parametri della mistura, ad esempio partizionando i dati e calcolando media e covarianza di ogni partizione

- Passo E: al giro m-esimo si stima w_{ij} applicando

$$w_{ij} = \frac{\pi_j^{(m)} p(\boldsymbol{x}_i | \boldsymbol{\theta}_j^{(m)})}{\sum_k \pi_k^{(m)} p(\boldsymbol{x}_i | \boldsymbol{\theta}_k^{(m)})}$$

EM per misture gaussiane

- Passo M: si aggiornano i parametri delle componenti gaussiane con

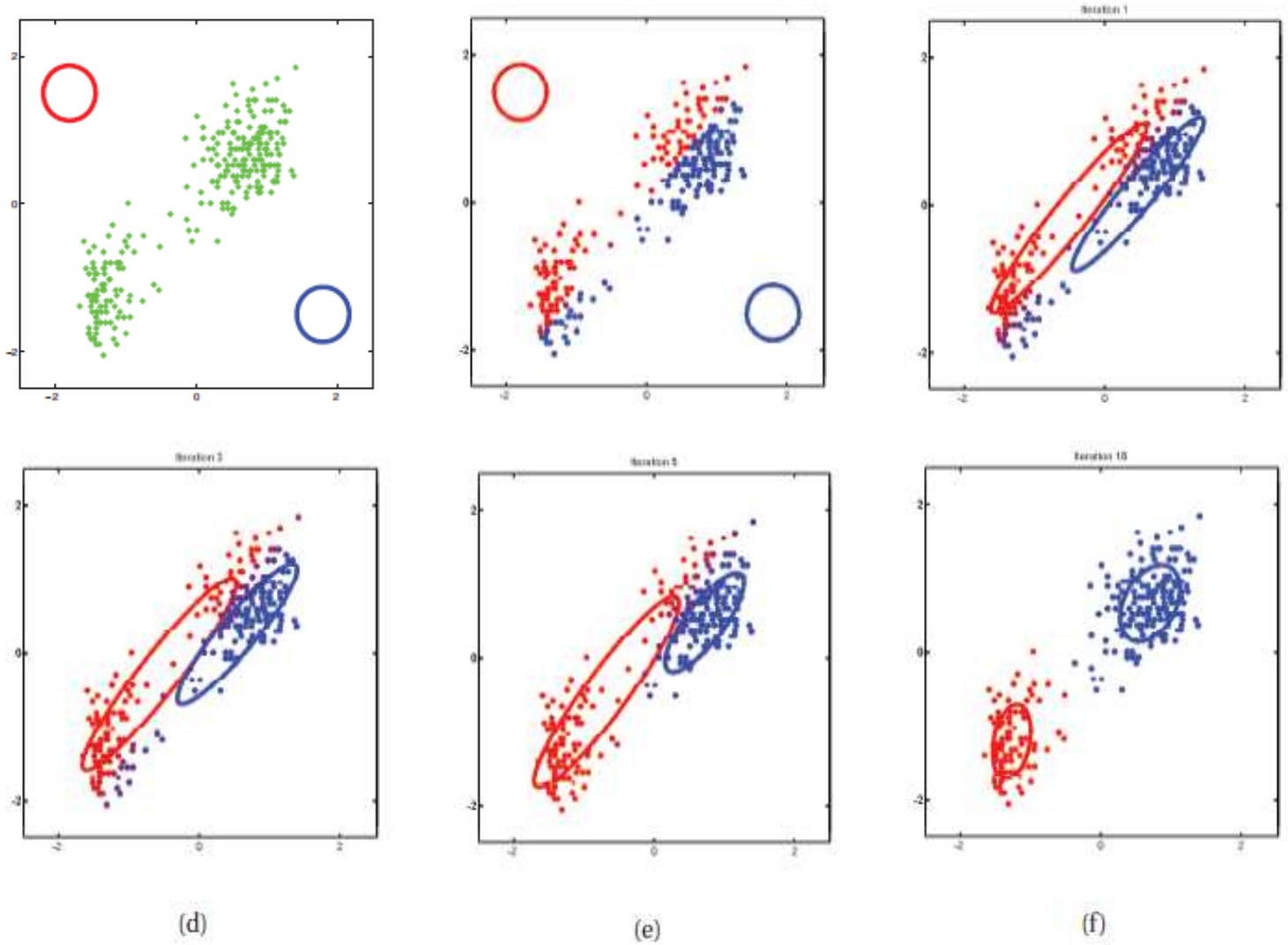
$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij} \qquad \hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{n} w_{ij}\boldsymbol{x}_i}{\sum_{i=1}^{n} w_{ij}}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^{n} w_{ij}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_j)^T}{\sum_{i=1}^{n} w_{ij}}$$

- Si iterano i passi E-M fino al raggiungimento di una condizione di stabilità della soluzione (massimo locale della likelyhood)

- La convergenza di EM è però molto lenta, ed il risultato dipende criticamente dalla scelta iniziale

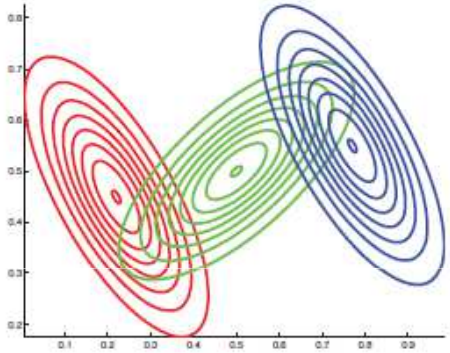
Esempio

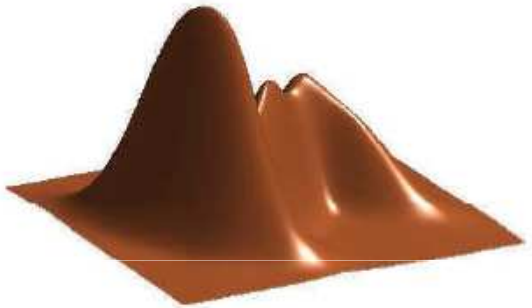


(d) (e) (f)

Vantaggi delle misture gaussiane

- L'uso di GMM consente di modellare distribuzioni di tipo qualunque





- Le componenti della mistura (e le loro variabili latenti) rappresentano un possibile *soft-clustering* dei dati (in analogia alla classificazione non supervisionata)

- L'algoritmo EM si presta ad essere modificato facilmente per generare versioni più specifiche

Problemi delle misture gaussiane

- L'uso di GMM presenta però diverse criticità pratiche:
  - E' difficile decidere a priori il numero di componenti della mistura, e si deve procedere per tentativi
  - Durante EM, alcune componenti possono degenerare, se la loro varianza tende a zero
  - Non è possibile definire per le misture una distanza analoga a quella di Mahalanobis
  - Il passo E tende ad essere computazionalmente instabile (elevato rischio di underflow)

Il trucco log-sum-exp

- La somma di quantità piccole richiesta dalla normalizzazione bayesiana ed a rischio di underflow può essere fatta con un semplice trucco numerico

$$\log \sum_c e^{b_c} = \log \left[(\sum_c e^{b_c - B}) e^B \right] = \left[\log (\sum_c e^{b_c - B}) \right] + B$$

dove $B = \max_c b_c$

- Ad esempio:

$$\log(e^{-120} + e^{-121}) = \log \left(e^{-120} (e^0 + e^{-1}) \right) = \log(e^0 + e^{-1}) - 120$$