

Università di Verona

A.A. 2020-21

# Machine Learning & Artificial Intelligence

Stima dei parametri:  
approccio Maximum Likelihood e  
approccio Bayesiano

Vittorio Murino

# Introduzione

- Per creare un classificatore ottimale che utilizzi la regola di decisione Bayesiana è necessario conoscere:
  - Le **probabilità a priori**  $P(\omega_i)$
  - Le **densità condizionali**  $p(\mathbf{x} \mid \omega_i)$
- Le performance di un classificatore dipendono fortemente dalla bontà di queste componenti
- **NON SI HANNO PRATICAMENTE MAI TUTTE QUESTE INFORMAZIONI!**

- Più spesso, si hanno unicamente:
  - Una *vaga conoscenza del problema*, da cui estrarre vaghe probabilità a priori.
  - *Alcuni pattern particolarmente rappresentativi, **training data***, usati per *addestrare* il classificatore (spesso troppo pochi!)
- La stima delle probabilità a priori di solito non risulta particolarmente difficoltosa.
- La stima delle densità condizionali è più complessa.

- Assunto che la conoscenza, benché approssimativa, delle densità a priori non presenta problemi, per quanto riguarda le densità condizionali le problematiche si possono suddividere in:
  1. **Stimare la funzione sconosciuta**  $p(\mathbf{x} \mid \omega_j)$
  2. **Stimare i parametri sconosciuti della funzione conosciuta**  $p(\mathbf{x} \mid \omega_j)$

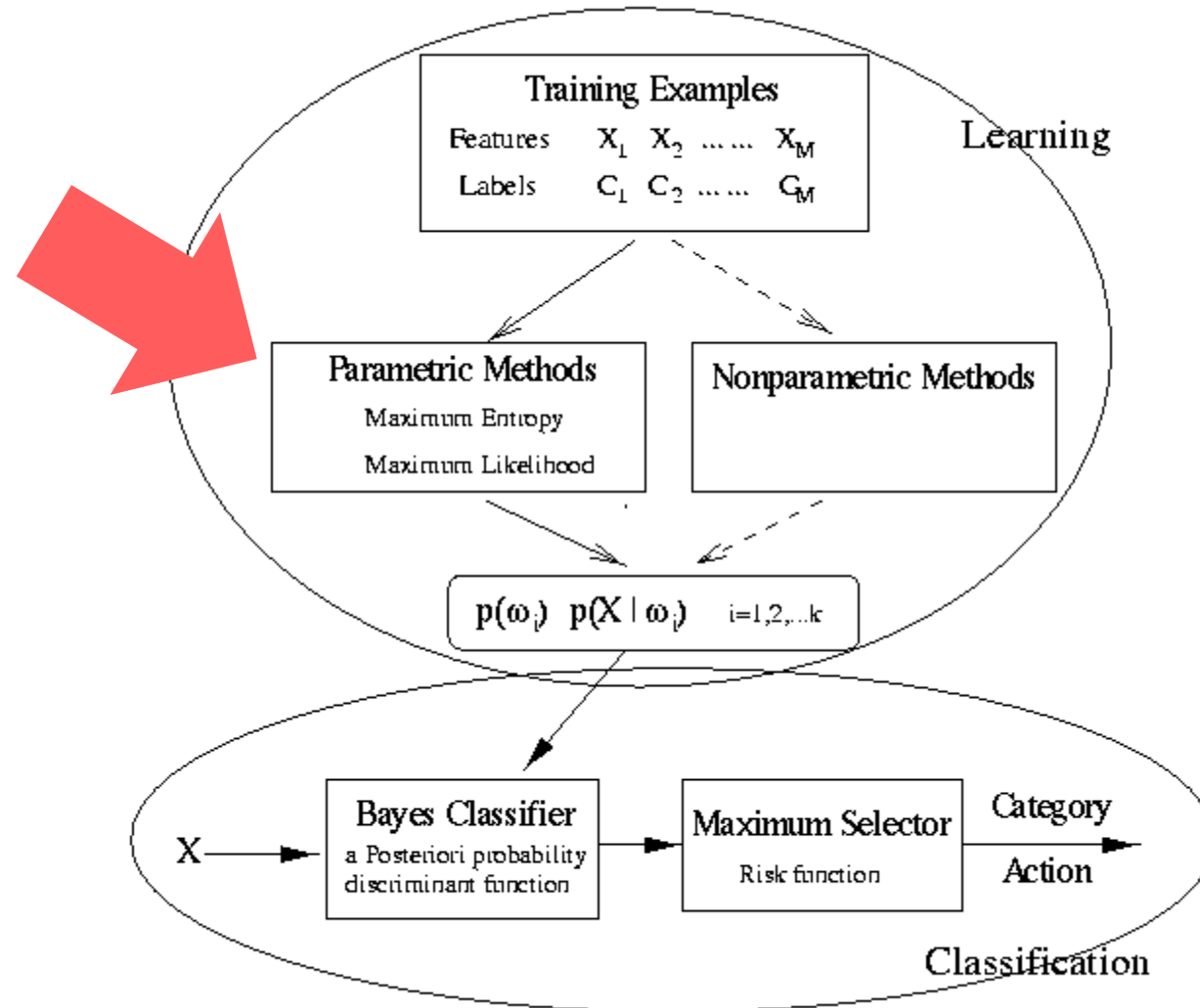
Per es., stimare il vettore  $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

quando  $p(\mathbf{x} \mid \omega_j) \approx N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

# Stima dei parametri

- Il secondo punto risulta di gran lunga più semplice (sebbene complesso!), e rappresenta un problema classico nella statistica.
- Trasferito nella *pattern recognition*, un approccio è quello di
  - 1) stimare i parametri dai dati di training
  - 2) usare le stime risultanti come se fossero valori veri
  - 3) utilizzare infine la teoria di decisione Bayesiana per costruire un classificatore

# Uno sguardo d'insieme



# Stima dei parametri – Probabilità a priori

- Supponiamo di avere un insieme di  $n$  dati di training in cui ad ogni pattern è assegnata un'etichetta d'identità (ossia conosco per certo a quale stato  $\omega_j$  appartiene il pattern  $k$ -esimo)

➔ *problema di learning dei parametri supervisionato*

- Allora 
$$P(\omega_i) = \frac{n_i}{n}$$

dove  $n_i$  è il numero di campioni con etichetta  $\omega_i$ , operazione dimostrabile formalmente

- Questa facile operazione non è di grande utilità, perchè *le probabilità a priori, in pratica, non sono così utili, se confrontate alle densità condizionali.*

# Stima dei parametri – Istanza del problema

- Supponiamo di avere  $c$  set di campioni  $D_1, D_2, \dots, D_c$  tracciati indipendentemente in accordo alla densità  $p(x/\omega_j)$ , assumendo che  $p(x/\omega_j)$  abbia forma parametrica conosciuta
- Il problema di stima dei parametri consiste nello stimare i parametri che definiscono  $p(x/\omega_j)$
- Per semplificare il problema, assumiamo inoltre che:
  - i campioni appartenenti al set  $D_i$  non danno informazioni relative ai parametri di  $p(x/\omega_j)$  se  $i \neq j$



# Stima dei parametri – Due approcci

- Specificatamente, il problema può essere formulato come:
  - Dato un set di training  $D=\{x_1, x_2, \dots, x_n\}$
  - $p(\mathbf{x}/\omega)$  è determinata da  $\theta$ , che è un vettore rappresentante i parametri necessari  
(p.e.,  $\theta = (\mu, \Sigma)$  se  $p(\mathbf{x} | \omega) \approx N(\mu, \Sigma)$  )
  - Vogliamo trovare il migliore  $\theta$  usando l'insieme di training.
- Esistono due approcci
  - Stima **Maximum-likelihood (ML)**
  - Stima di **Bayes**

# Stima dei parametri – Due approcci (2)

- Approccio *Maximum Likelihood*
  - I parametri sono *quantità fissate* ma sconosciute
  - La migliore stima dei loro valori è quella che *massimizza la probabilità di ottenere i dati di training*
- Approccio *Bayesiano*
  - I parametri sono *variabili aleatorie* aventi determinate probabilità a priori
  - Le osservazioni dei dati di training trasformano queste probabilità in probabilità a posteriori modificando la stima dei veri valori dei parametri.
  - Aggiungendo campioni di training il risultato è di rifinire meglio la forma delle densità a posteriori, causando un innalzamento di esse in corrispondenza dei veri valori dei parametri (fenomeno di *Bayesian Learning*).
- I risultati dei due approcci, benché proceduralmente diversi, sono qualitativamente simili.

# Approccio *Maximum Likelihood*

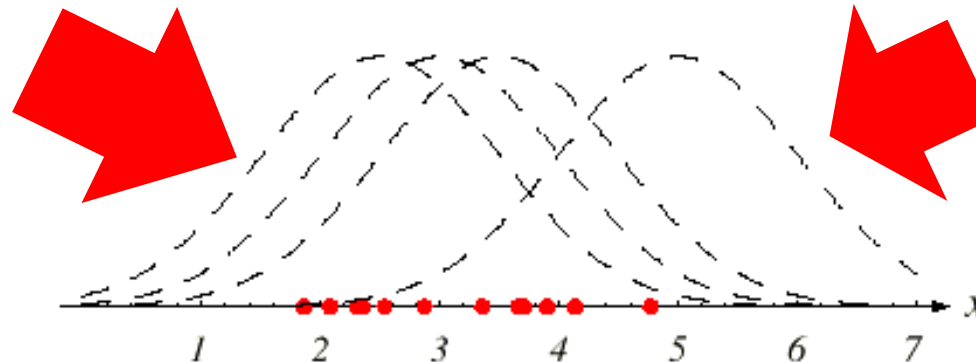
- In forza dell'ipotesi di partenza del problema, poiché i pattern del set  $\mathbf{D}$  sono i.i.d., abbiamo che:

$$p(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^n p(x_k \mid \boldsymbol{\theta})$$

- Vista come funzione di  $\boldsymbol{\theta}$ ,  $p(\mathbf{D} \mid \boldsymbol{\theta})$  viene chiamata *likelihood* di  $\boldsymbol{\theta}$  rispetto al set di campioni  $\mathbf{D}$ .
- La stima di Maximum Likelihood di  $\boldsymbol{\theta}$  è, per definizione, il valore  $\hat{\boldsymbol{\theta}}$  che massimizza  $p(\mathbf{D} \mid \boldsymbol{\theta})$ ;
- Ricordiamo l'assunzione che  $\boldsymbol{\theta}$  è fissato ma sconosciuto

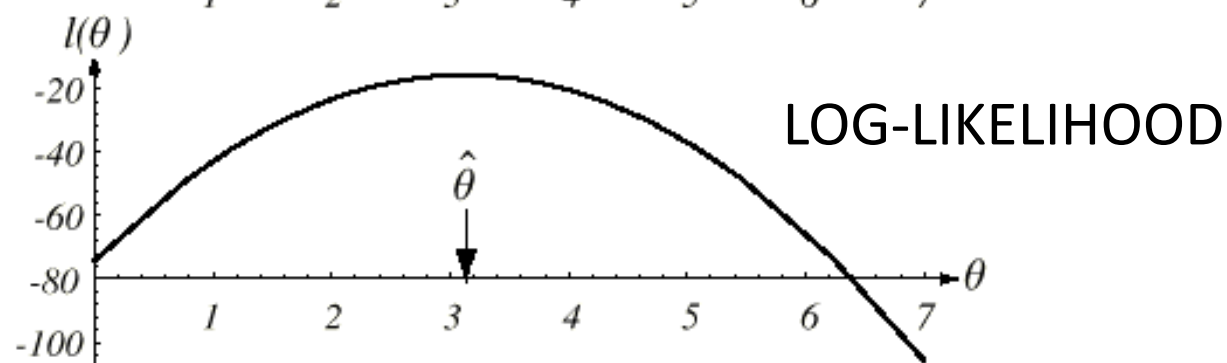
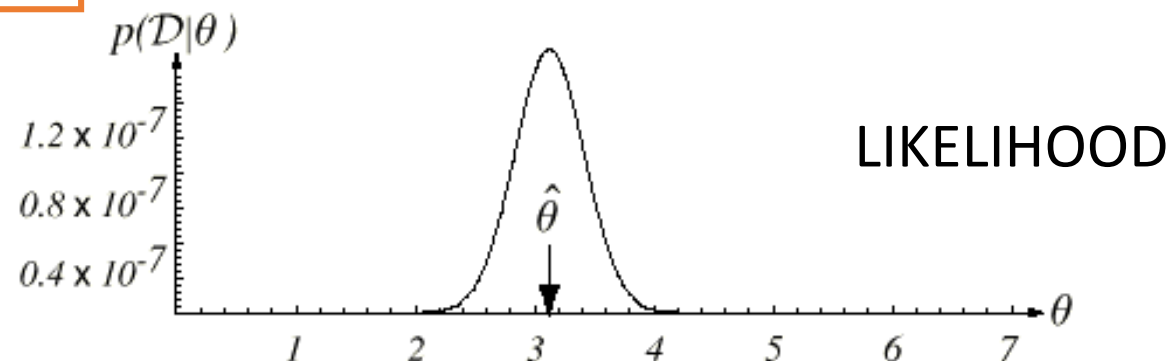
# Approccio Maximum Likelihood (2)

Punti di training 1-D  
assunti generati da una  
densità gaussiana di  
varianza fissata ma  
media sconosciuta



4 delle  
infinite  
possibili  
gaussiane

NB: La likelihood  
 $p(D|\theta)$  è funzione di  
 $\theta$ , mentre la densità  
condizionale  $p(x|\theta)$  è  
funzione di  $x$



# Approccio Maximum Likelihood (3)

- Se il numero di parametri da stimare è  $p$ , sia  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^t$  e

$$\nabla \boldsymbol{\theta} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

- Per scopi analitici risulta più semplice lavorare con il logaritmo della likelihood.
- Definiamo quindi  $l(\boldsymbol{\theta})$  come ***funzione di log-likelihood***

$$l(\boldsymbol{\theta}) \equiv \ln p(D | \boldsymbol{\theta}) = \sum_{k=1}^n \ln p(x_k | \boldsymbol{\theta})$$

# Approccio Maximum Likelihood (4)

- Lo scopo è di ottenere quindi il vettore

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

in cui la dipendenza sul data set  $D$  è implicita.

- Pertanto per ricavare il max:

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathbf{D} | \boldsymbol{\theta}) = \sum_{k=1}^n \ln p(x_k | \boldsymbol{\theta})$$



$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k=1}^n \nabla_{\boldsymbol{\theta}} \ln p(x_k | \boldsymbol{\theta})$$

da cui vogliamo ottenere  $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$

# Approccio Maximum Likelihood (5)

- Formalmente, una volta stimato il set di parametri, è necessario controllare che la soluzione trovata sia effettivamente un massimo globale, piuttosto che un massimo locale o un flesso o peggio ancora un punto di minimo.
- Bisogna anche controllare cosa accade ai bordi degli estremi dello spazio dei parametri
- Applichiamo ora l'approccio ML ad alcuni casi specifici.

# Maximum Likelihood: caso Gaussiano

- Consideriamo che i campioni siano generati da una popolazione normale multivariata di media  $\boldsymbol{\mu}$  e covarianza  $\boldsymbol{\Sigma}$ .
- Per semplicità, consideriamo il caso in cui solo la media  $\boldsymbol{\mu}$  sia sconosciuta. Consideriamo quindi il punto campione  $\mathbf{x}_k$  e troviamo:

$$\ln p(\mathbf{x}_k | \boldsymbol{\mu}) = -\frac{1}{2} \ln[(2\pi)^d |\boldsymbol{\Sigma}|] - \frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$



$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k | \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$



## Maximum Likelihood: caso Gaussiano (2)

- Identificando  $\theta$  con  $\mu$  si deduce che la stima Maximum-Likelihood di  $\mu$  deve soddisfare la relazione:

$$\sum_{k=1}^n \Sigma^{-1} (\mathbf{x}_k - \hat{\mu}) = 0$$

- Moltiplicando per  $\Sigma$  e riorganizzando la somma otteniamo

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$$

che non è altro che la semplice **media** degli esempi di training, altresì indicata con  $\hat{\mu}_n$  per indicarne la dipendenza dalla numerosità del training set.

## Maximum Likelihood: caso Gaussiano (3)

- Consideriamo ora il caso più tipico in cui la distribuzione Gaussiana abbia media e covarianza ignote.
- Consideriamo prima il caso univariato  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$
- Se si prende un singolo punto abbiamo

$$\ln p(x_k | \theta) = -\frac{1}{2} \ln[2\pi\theta_2] - \frac{1}{2\theta_2} (x_k - \theta_1)^2$$

la cui derivata è

$$\nabla_{\theta} l = \nabla_{\theta} \ln p(x_k | \theta) = \begin{bmatrix} \frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

## Maximum Likelihood: caso Gaussiano (4)

- Eguagliando a 0 e considerando tutti i punti si ottiene:

$$\sum_{k=1}^n \frac{1}{\theta_2} (x_k - \hat{\theta}_1) = 0 \quad - \sum_{k=1}^n \frac{1}{\hat{\theta}_2} + \sum_{k=1}^n \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

dove  $\hat{\theta}_1$  e  $\hat{\theta}_2$  sono le stime ML per  $\theta_1$  e  $\theta_2$ .

- Sostituendo  $\hat{\mu} = \hat{\theta}_1$  e  $\sigma^2 = \hat{\theta}_2$  si hanno le stime ML di media e varianza

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2$$

# Maximum Likelihood: caso Gaussiano (5)

- Il caso multivariato si tratta in maniera analoga con più conti. Il risultato è comunque:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

- Si noti tuttavia che la stima della covarianza è sbilanciata, i.e., il valore aspettato della varianza campione su tutti i possibili insiemi di dimensione  $n$  non è uguale alla vera varianza

$$E\left\{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right\} = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

# Maximum-Likelihood: altri casi

- Esistono, oltre alla densità Gaussiana, anche altre famiglie di densità che costituiscono altrettante famiglie di parametri:

- **Distribuzione esponenziale** 
$$p(x | \theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

- **Distribuzione uniforme** 
$$p(x | \theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{altrimenti} \end{cases}$$

- **Distribuzione di Bernoulli multivariata**

# Maximum-Likelihood – Modello d'errore

- In generale, se i modelli parametrici sono validi, il classificatore *maximum-likelihood* fornisce risultati eccellenti.
- Invece, se si usano famiglie parametriche scorrette, il classificatore produce forti errori
  - Questo accade anche se è nota la famiglia parametrica da usare, per esempio se si stima all'interno di una distribuzione gaussiana come parametro una varianza troppo larga.

## Maximum-Likelihood – Modello d'errore (2)

- Di fatto *manca un modello d'errore che dia un valore di confidenza o affidabilità alla parametrizzazione ottenuta.*
- Inoltre, per applicare la stima di Maximum-Likelihood, tutti i dati di training devono essere disponibili
  - Se vogliamo utilizzare nuovi dati di training, è necessario ricalcolare la procedura di stima Maximum-Likelihood.

# Stima di Bayes

- A differenza dell'approccio ML, in cui supponiamo  $\theta$  come fissato ma sconosciuto, *l'approccio di stima Bayesiana* dei parametri considera  $\theta$  come una **variabile aleatoria**.
- In questo caso il set di dati di training  $D$  ci permette di *convertire una distribuzione a priori  $p(\theta)$  su questa variabile in una densità di probabilità a posteriori  $p(\theta|D)$*

$$p(\theta) \rightarrow p(\theta|D)$$

- Data la difficoltà dell'argomento, è necessario un passo indietro al concetto di classificazione Bayesiana



# Approccio di stima Bayesiano – Idea centrale

- Il calcolo delle densità a posteriori  $P(\omega_i | \mathbf{x})$  sta alla base della classificazione Bayesiana
- Per creare un classificatore ottimale che utilizzi la regola di decisione Bayesiana è necessario conoscere:
  - Le **probabilità a priori**  $P(\omega_i)$
  - Le **densità condizionali**  $p(\mathbf{x} | \omega_i)$
- Quando queste quantità sono sconosciute, bisogna ricorrere a tutte le informazioni a disposizione.

## Approccio di stima Bayesiano – Idea centrale (2)

- Parte di queste informazioni può essere derivante da:
  1. **Conoscenza a priori**
    - *Forma funzionale delle densità sconosciute*
    - *Intervallo dei valori dei parametri sconosciuti*
  2. **Training set**
    - Sia  **$D$**  il *set totale di campioni*: il nostro compito si trasforma così nella stima di  $P(\omega_i | x, D)$
- Da queste probabilità possiamo ottenere il classificatore Bayesiano.

## Approccio di stima Bayesiano – Idea centrale (3)

- Dato il set di training  $D$ , la formula di Bayes diventa:

$$P(\omega_i | \mathbf{x}, D) = \frac{p(\mathbf{x} | \omega_i, D)P(\omega_i | D)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j, D)P(\omega_j | D)}$$

- Assunzioni:
  - Ragionevolmente,  $P(\omega_i | D) \Rightarrow P(\omega_i)$
  - Dato il caso di learning supervisionato il set  $D$  è partizionato in  $c$  set di campioni  $D_1, D_2, \dots, D_c$  con i campioni in  $D_i$  appartenenti a  $\omega_i$
  - I campioni appartenenti al set  $D_i$  non danno informazioni sui parametri di  $p(\mathbf{x} | \omega_j, D)$  se  $i \neq j$ .

## Approccio di stima Bayesiano – Idea centrale (4)

- Queste assunzioni portano a due conseguenze:
  1. Possiamo lavorare con ogni classe indipendentemente, ossia

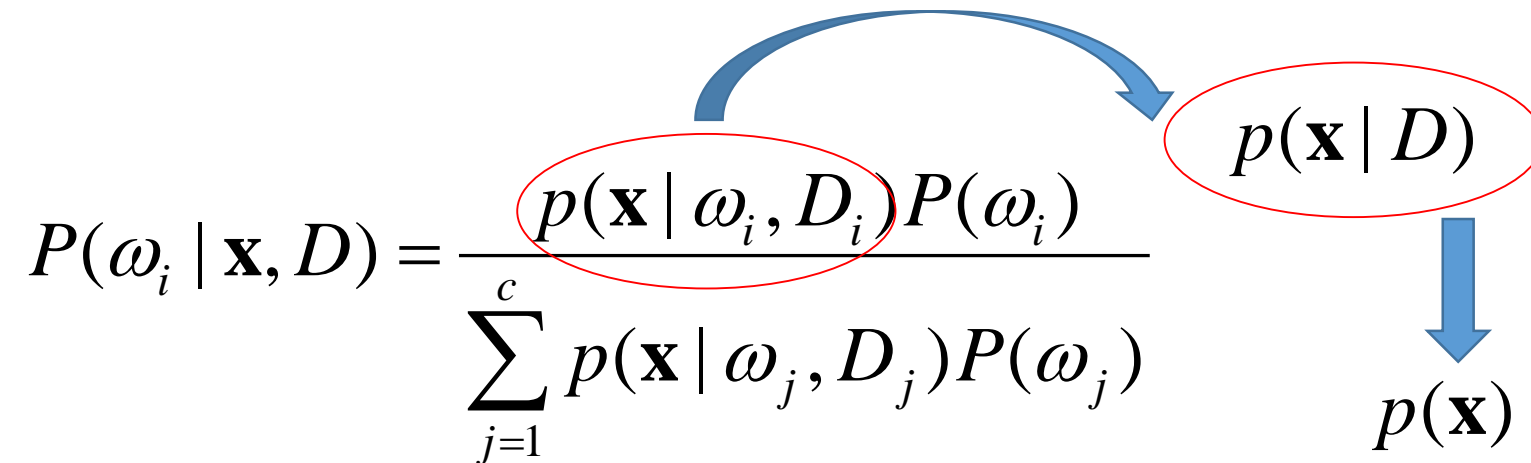
$$P(\omega_i | \mathbf{x}, D) = \frac{p(\mathbf{x} | \omega_i, D) P(\omega_i | D)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j, D) P(\omega_j | D)}$$



$$P(\omega_i | \mathbf{x}, D) = \frac{p(\mathbf{x} | \omega_i, D_i) P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j, D_j) P(\omega_j)}$$

## Approccio di stima Bayesiano – Idea centrale (5)

2. Poiché ogni classe può essere trattata indipendentemente, si possono evitare le distinzioni tra le classi e semplificare la notazione **riducendola a  $c$  diverse istanze dello stesso problema**, ossia:


$$P(\omega_i | \mathbf{x}, D) = \frac{p(\mathbf{x} | \omega_i, D_i) P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j, D_j) P(\omega_j)}$$

$p(\mathbf{x})$

*Si usa un set di campioni  $D$ , estratti secondo la distribuzione sconosciuta  $p(\mathbf{x})$ , per determinare  $p(\mathbf{x}/D)$*

## Approccio di stima Bayesiano – Idea centrale (6)

- In pratica il processo di learning Bayesiano *stima un modello implicitamente*, ossia non restituisce un vettore di parametri  $\theta$  visibile, ma *una distribuzione su di esso*, data dal training set disponibile.
- Il fatto che  $p(\mathbf{x})$  sia ignoto ma con forma parametrica nota si esprime dicendo che  $p(\mathbf{x} | \theta)$  è completamente noto.
- Si preferisce quindi scrivere  $p(\mathbf{x} | D)$  anzichè  $p(\mathbf{x} | \theta)$  perché è più significativo, benchè un modello sottostante esista (difatti il termine  $p(\mathbf{x} | \theta)$  comparirà più avanti).
- Ogni informazione si abbia prima di osservare i campioni si assume sia contenuta nella densità a priori  $p(\theta)$  nota.
- Le osservazioni convertono il prior  $p(\theta)$  in una distribuzione a posteriori  $p(\theta | D)$  che sperabilmente assume un massimo in corrispondenza del valore vero di  $\theta$ .

# Distribuzione dei parametri

- Ingredienti:

- $p(\mathbf{x})$  : sconosciuta, ma di forma parametrica nota;
- $\theta$  : *vettore dei parametri*, sconosciuto;
- $p(\mathbf{x}|\theta)$  : completamente conosciuta (essendo la forma parametrica  $p(\mathbf{x})$  );
- $p(\theta)$  : *ogni informazione a priori di osservare determinati campioni.*

L'osservazione dei campioni converte questa distribuzione in una ...

- $p(\theta|D)$  : ... *probabilità a posteriori*, presumibilmente centrata attorno ai veri valori di  $\theta$ .

## Distribuzione dei parametri (2)

- Quello che stiamo facendo è effettivamente osservare come effettivamente viene ottenuta  $p(\mathbf{x}|D)$  tramite l'ausilio di un modello di parametri implicito  $\theta$ .
- Stiamo cioè *esplicitando* il calcolo di  $p(\mathbf{x}|D)$ , per stimare  $p(\mathbf{x})$ , *convertendo il problema di stima di una densità di probabilità a quello di stima di un vettore di parametri*.
  - Ragionevolmente, abbiamo

$$p(\mathbf{x} | D) = \int p(\mathbf{x}, \theta | D) d\theta$$

dove l'integrazione si estende su tutto lo spazio dei parametri



## Distribuzione dei parametri (3)

- Quindi 
$$p(\mathbf{x} | D) = \int p(\mathbf{x}, \boldsymbol{\theta} | D) d\boldsymbol{\theta}$$
$$= \int p(\mathbf{x} | \boldsymbol{\theta}, D) p(\boldsymbol{\theta} | D) d\boldsymbol{\theta}$$

- Poichè, per ipotesi, la selezione di  $\mathbf{x}$  è indipendente dai campioni di training  $D$ , dato  $\boldsymbol{\theta}$ ,

$$p(\mathbf{x} | D) = \int p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D) d\boldsymbol{\theta}$$

- Pertanto la distribuzione  $p(\mathbf{x})$  è completamente conosciuta quando conosco il vettore dei parametri  $\boldsymbol{\theta}$

## Distribuzione dei parametri (4)

- L'equazione precedente lega esplicitamente la densità condizionale  $p(\mathbf{x}|\mathbf{D})$  alla densità a posteriori  $p(\boldsymbol{\theta}|\mathbf{D})$  tramite il vettore sconosciuto di parametri  $\boldsymbol{\theta}$ .
- Se  $p(\boldsymbol{\theta}|\mathbf{D})$  si concentra fortemente su un valore, otteniamo una stima  $\hat{\boldsymbol{\theta}}$  del vettore più probabile, quindi

$$p(\mathbf{x}|\mathbf{D}) \approx p(\mathbf{x} | \hat{\boldsymbol{\theta}})$$

- Ma questo approccio ***permette di tenere conto dell'effetto di tutti gli altri modelli***, descritti dal valore della funzione integrale, ***per tutti i possibili modelli***.

$$p(\mathbf{x} | D) = \int p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D) d\boldsymbol{\theta}$$

# Esempio: caso Gaussiano

- Utilizziamo le tecniche di stima Bayesiana per calcolare la densità a posteriori  $p(\boldsymbol{\theta}|\mathbf{D})$  e la densità  $p(\mathbf{x}|\mathbf{D})$  per il caso in cui

$$p(\mathbf{x} | \boldsymbol{\theta}) \equiv p(\mathbf{x} | \boldsymbol{\mu}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **CASO UNIVARIATO:**


$$p(\mathbf{x} | \boldsymbol{\mu}) \equiv p(x | \mu) \approx N(\mu, \sigma^2)$$

*L'unica quantità sconosciuta è la media  $\mu$*

$$p(\mu) \approx N(\mu_0, \sigma_0^2)$$

*La conoscenza a priori su  $\mu$ , espressa da una densità di cui media e varianza sono noti*



**Prior coniugato**

In pratica  $\mu_0$  rappresenta la migliore scelta iniziale per il parametro  $\mu$ , con  $\sigma_0^2$  che ne misura l'incertezza.

## Esempio: caso Gaussiano (2)

- A questo punto estraiamo  $\mu$  da  $N(\mu_0, \sigma_0^2)$
- Esso diventa il vero valore del parametro e determina completamente la densità per  $x$ .
- Supponiamo di avere  $n$  campioni di training  $D = \{x_1, x_2, \dots, x_n\}$  e calcoliamo

Densità  
riprodotta



$$p(\mu | D) = \frac{p(D | \mu) p(\mu)}{\int p(D | \mu) p(\mu) d\mu}$$
$$= \alpha \prod_{k=1}^n p(x_k | \mu) p(\mu)$$

dove  $\alpha$  è un fattore di normalizzazione dipendente da  $D$ .

## Esempio: caso Gaussiano (3)

- L'equazione mostra come l'osservazione del set di esempi di training influenzi la nostra idea sul vero valore di  $\mu$ ; essa relaciona la densità a priori  $p(\mu)$  con la densità a posteriori  $p(\mu/D)$ .
- Svolgendo i calcoli, ci si accorge che, grazie al prior normale,  $p(\mu/D)$  risulta anch'esso normale, modificandosi in dipendenza del numero di campioni che formano il training set, evolvendosi in impulso di Dirac per  $n \rightarrow \infty$  (fenomeno di Learning Bayesiano).
- Formalmente si giunge alle seguenti formule:



## Esempio: caso Gaussiano (4)

$$p(\mu | D) = \frac{p(D | \mu) p(\mu)}{\int p(D | \mu) p(\mu) d\mu} = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{-\frac{(\mu - \mu_n)^2}{2\sigma_n^2}\right\}$$

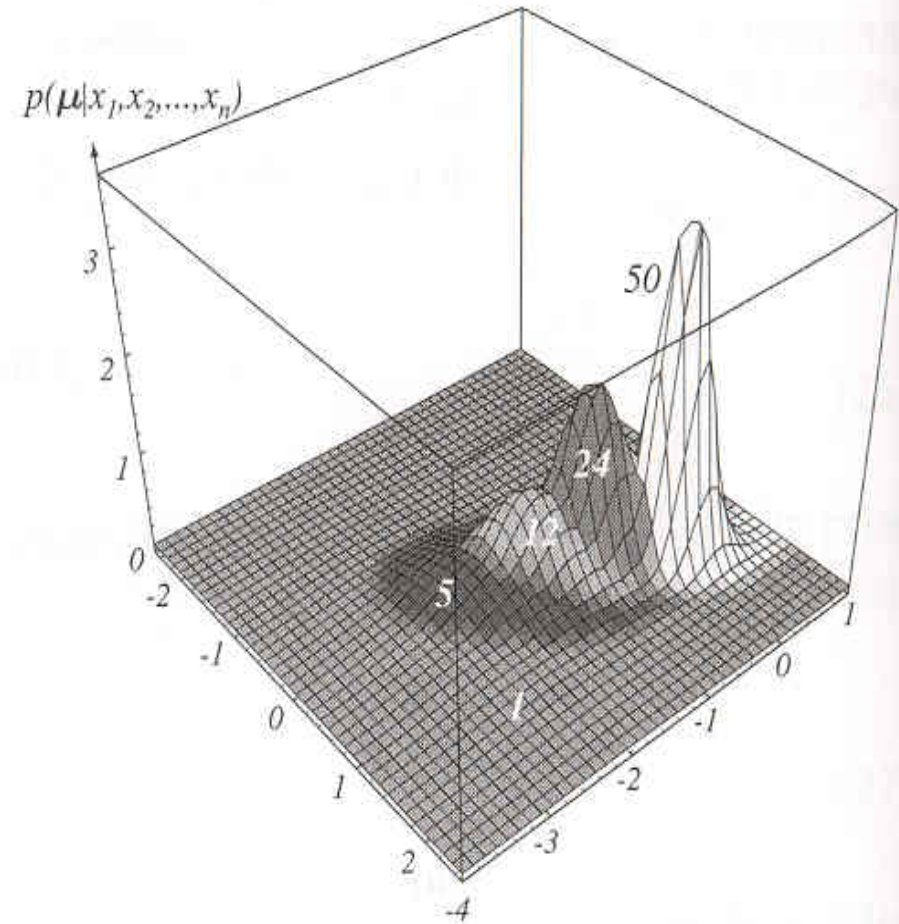
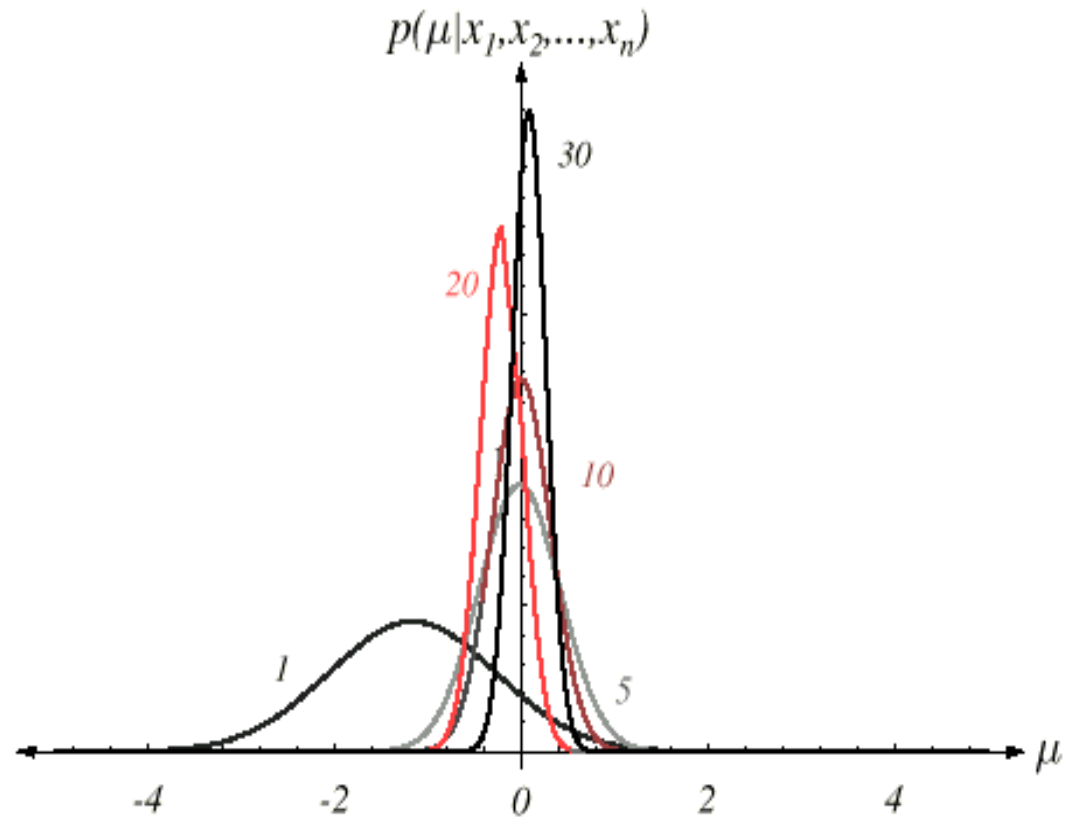
$$\text{dove } \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \left( \frac{1}{n} \sum_{k=1}^n x_k \right) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

$\mu_n$  rappresenta la nostra migliore scelta per  $\mu$  dopo aver osservato  $n$  campioni.

$\sigma_n^2$  misura l'incertezza della nostra scelta.

## Esempio: caso Gaussiano (5)



## Esempio: caso Gaussiano (6)

- A questo punto, avendo ottenuto una densità a posteriori per la media,  $p(\mu/D)$ , *quello che rimane è ottenere la densità condizionale*  $p(x|D)$ , che in notazione esatta, ricordiamo, è  $p(x/\omega_i, D_i)$ . Quindi:

$$\begin{aligned} p(x|D) &= \int p(x|\mu) p(\mu|D) d\mu \\ &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2 + \sigma_n^2}\right] f(\sigma, \sigma_n), \end{aligned} \tag{36}$$

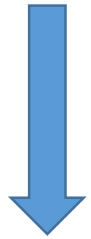


## Esempio: caso Gaussiano (7)

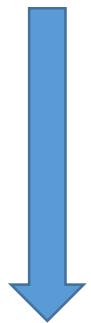
$$f(\sigma, \sigma_n) = \int \exp \left[ -\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} \left( \mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2 + \sigma_n^2} \right)^2 \right] d\mu.$$

- Osservando l'equazione 36, si nota che

$$p(x | D) \approx N(\mu_n, \sigma^2 + \sigma_n^2)$$




- Se confrontiamo la densità condizionale  $p(x|D)$ , con la sua forma parametrica  $p(x | \mu) \approx N(\mu, \sigma^2)$  osserviamo che la media condizionale è trattata come se fosse la vera media, e la varianza nota è proporzionale al grado di incertezza corrente.



## Esempio: caso Gaussiano (8)

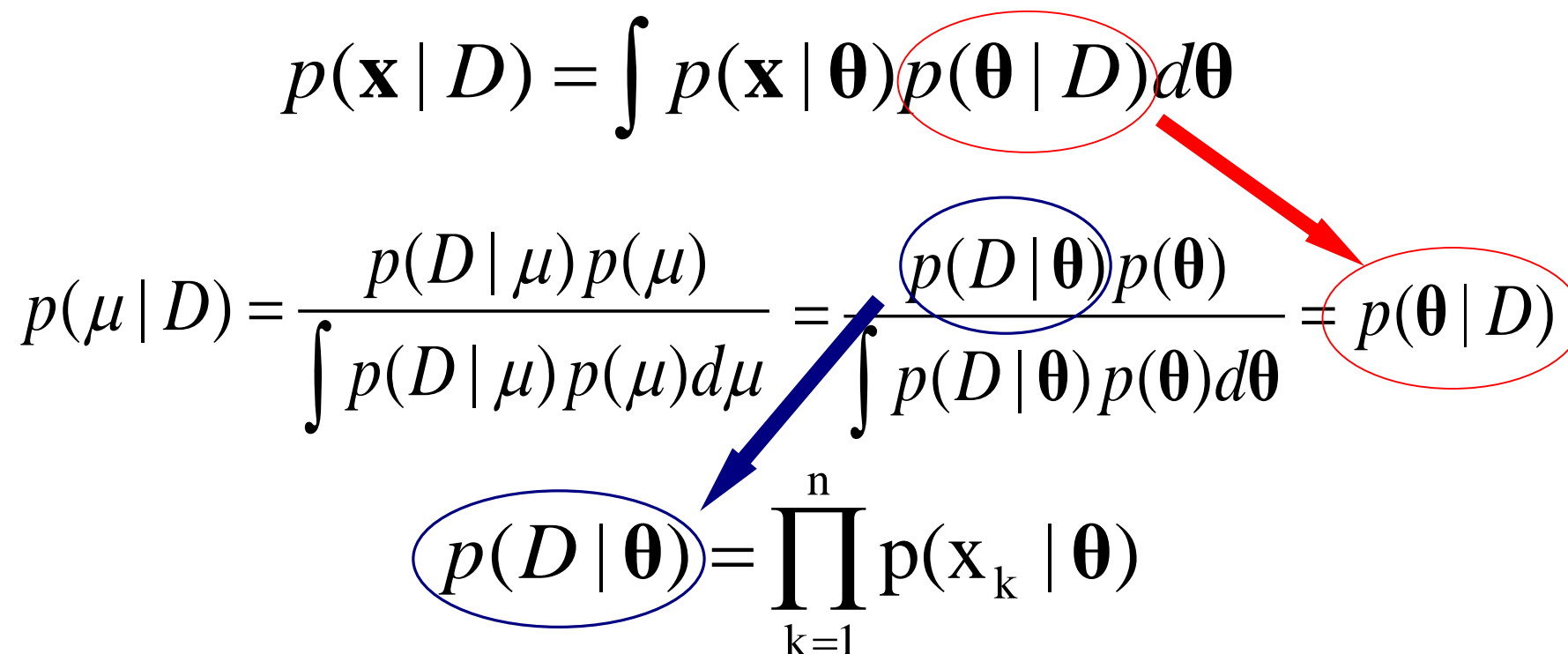
- Concludendo, la densità  $p(x/D)$  ottenuta è la densità condizionale desiderata


$$P(\omega_i | \mathbf{x}, D) = \frac{p(\mathbf{x} | \omega_i, D)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j, D)P(\omega_j)}$$

che assieme ai prior  $P(\omega_i)$  produce le informazioni desiderate per il design del classificatore, al contrario dell'approccio ML che restituisce solo le stime puntuali  $\hat{\mu}$  e  $\hat{\sigma}^2$

# Stima dei parametri Bayesiana: teoria generale

- Riassumendo ed estendendole al caso generale, le formule principali viste sono:

$$p(\mathbf{x} | D) = \int p(\mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D) d\boldsymbol{\theta}$$
$$p(\mu | D) = \frac{p(D | \mu) p(\mu)}{\int p(D | \mu) p(\mu) d\mu} = \frac{p(D | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = p(\boldsymbol{\theta} | D)$$
$$p(D | \boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k | \boldsymbol{\theta})$$


- Si noti la somiglianza con l'approccio ML, con la differenza che qui non si cerca il max puntuale  $\hat{\boldsymbol{\theta}}$

# Stima dei parametri Bayesiana: teoria generale (2)

- Vi sono ancora questioni da chiarire:
  - *Difficoltà di esplicitare le formule viste*
  - *Convergenza di  $p(\mathbf{x}|D)$  a  $p(\mathbf{x})$ ;*
- Convergenza: supponiamo  $D^n = \{x_1, \dots, x_n\}$ ,  $n > 1$ :

$$p(D^n | \theta) = p(\mathbf{x}_n | \theta) p(D^{n-1} | \theta)$$
$$p(\theta | D) = \frac{p(D | \theta) p(\theta)}{\int p(D | \theta) p(\theta) d\theta}$$
$$p(\theta | D^n) = \frac{p(\mathbf{x}_n | \theta) p(\theta | D^{n-1})}{\int p(\mathbf{x}_n | \theta) p(\theta | D^{n-1}) d\theta}$$

Metodo on line  
di Bayesian learning

Assumendo che  
 $p(\theta | D^0) = p(\theta)$

## Approccio Bayesiano – Conclusioni

- Per concludere, estendendo la notazione alle varie classi  $\omega_i$  e corrispondenti training set  $D_i$ , il *design di un classificatore Bayesiano tramite stima dei parametri con approccio Bayesiano* risulta sottostare alle seguenti formule:

$$\begin{aligned} p(\theta | D_i, \omega_i) &= \frac{p(D_i | \theta, \omega_i) p(\theta | \omega_i)}{\int p(D_i | \theta, \omega_i) p(\theta | \omega_i) d\theta} \\ &= \frac{\prod_{k=1}^{n_i} p(x_{i,k} | \theta) p(\theta | \omega_i)}{\int \prod_{k=1}^{n_i} p(x_{i,k} | \theta) p(\theta | \omega_i) d\theta} \end{aligned}$$

## Approccio Bayesiano – Conclusioni (2)

- Sia  $D_i^n = \{x_{i,1}, \dots, x_{i,n}\}$

$$\begin{aligned} p(\theta \mid D_i^n, \omega_i) &= \frac{\prod_{k=1}^{n_i} p(x_{i,k} \mid \theta, \omega_i) p(\theta \mid \omega_i)}{\int \prod_{k=1}^{n_i} p(x_{i,k} \mid \theta, \omega_i) p(\theta \mid \omega_i) d\theta} \\ &= \frac{p(x_{i,n_i} \mid \theta) p(\theta \mid D_i^{n-1}, \omega_i)}{\int p(x_{i,n_i} \mid \theta) p(\theta \mid D_i^{n-1}, \omega_i) d\theta} \end{aligned}$$

## Approccio Bayesiano – Conclusioni (3)

- Il classificatore **minimum error rate** risulta
  - Decidi  $\omega_i$  se  $P(\omega_i | x) \geq P(\omega_j | x)$ , per  $j=1, \dots, c$

$$P(\omega_i | x, D_i) = \frac{p(x | \omega_i, D_i) P(\omega_i)}{p(x | D_i)}$$

$$\begin{aligned} p(x | \omega_i, D_i) &= \int p(x, \theta | \omega_i, D_i) d\theta \\ &= \int p(x | \theta) p(\theta | \omega_i, D_i) d\theta \end{aligned}$$

# Confronto stime ML – Bayesiana

- ML restituisce una stima puntuale  $\hat{\theta}$ , l'approccio Bayesiano una distribuzione su  $\theta$ .
- Le stime risultano equivalenti per training set di cardinalità infinita
  - Al limite,  $p(\theta/D)$  converge ad una funzione delta di Dirac
- Praticamente, gli approcci sono differenti per vari motivi:
  - *Complessità computazionale*
  - *Interpretabilità*
  - *Affidabilità delle informazioni a priori*
  - *Compromesso tra accuratezza della stima e varianza*