


Mixture Model per la stima di densità di probabilità

Iuri Frosio
Università degli Studi di Milano
Dipartimento di Scienze dell'Informazione
AIS Lab.
frosio@dsi.unimi.it


In questa lezione…

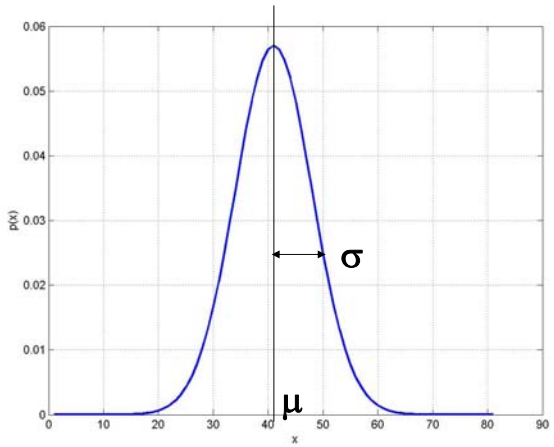
- Richiami di statistica;
- Metodi non parametrici per la stima di densità di probabilità;
- Mixture models per la stima di densità di probabilità.

Richiami di statistica

- Sia X una variabile aleatoria (v.a.). La densità di probabilità (d.d.p.) p(x) descrive la probabilità che X assuma valore uguale a x.
- Se, ad esempio, X ha una distribuzione gaussiana, con media μ, deviazione standard σ, si avrà:

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

Richiami di statistica



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

Richiami di statistica

- Proprietà di p(x):

$$p(x) > 0, \forall x; \quad \int_{-\infty}^{+\infty} p(x)dx = 1$$

- Probabilità che x sia minore di y:

$$P(x \le y) = \int_{-\infty}^{y} p(x)dx$$

Richiami di statistica

- Nella realtà, media e varianza possono non essere note, ma misurate su un numero finito (N) di dati X_i.
- Media campionaria:

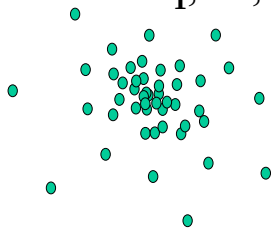
$$E[\mathbf{X}] = \mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X_i}$$

- Varianza campionaria:

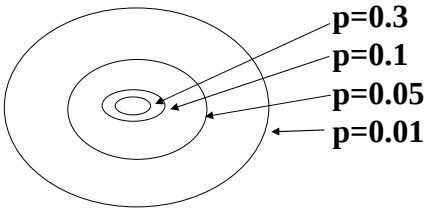
$$Var[\mathbf{X}] = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{X_i} - E[\mathbf{X}])^2$$

Stima della d.d.p.

•Data una serie di dati $\mathbf{X}_1$, …, $\mathbf{X}_D$, …



• …Stimarne la d.d.p., p(**x**).



7/49

Applicazioni della stima di d.d.p.

- analisi dei dati;
- Machine learning (apprendimento intelligente);
- Classificazione (clustering di dati e immagini);
- Ricostruzione superfici;
- ...

8/49

Metodi di stima delle d.d.p.

1) Non parametrici:

   Non viene effettuata nessuna ipotesi a priori sulla forma della d.d.p. da stimare, $p = p(x)$;

2) Semi-parametrici:

   La forma della d.d.p. da stimare è data a priori, ma il vettore dei parametri (θ) non ha alcun significato statistico, $p = p(x, \theta)$;

3) Parametrici:

   La forma della d.d.p. da stimare è data a priori, il vettore dei parametri (θ) è composto da elementi con significato statistico, $p = p(x, \theta)$.

Metodi non parametrici

Problema

- Data una serie di dati $\mathbf{x_i}$, determinare la densità di probabilità che ha generato gli $\mathbf{x_i}$.

Esempio

- Distribuzione dell'età degli studenti.

Ipotesi

- Nessuna.

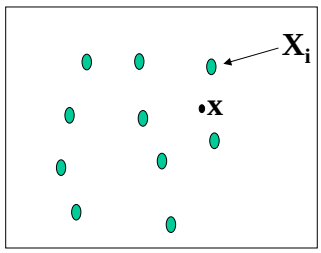
Soluzione

- Metodo di Parzen, K-nearest neighboors.

Osservazioni

- Le densità di probabilità, $p(\mathbf{x})$, viene stimata in ogni punto $\mathbf{x}$ di interesse.

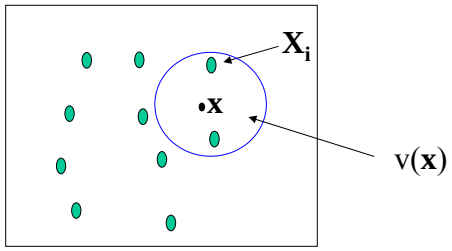
Metodi non parametrici



In generale: $p(\mathbf{x}) = \dfrac{n(\mathbf{x})}{N \cdot v(\mathbf{x})}$

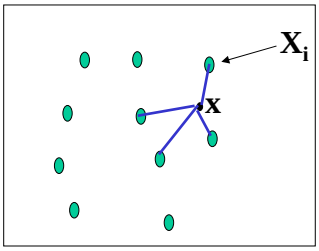
Dove: p(**x**) è la stima della densità di probabilità in **x**, n(**x**) è il numero di campioni nell'intorno di **x**, v(**x**) è il volume centrato in x contenente gli n(**x**) campioni, N è il numero totale di campioni.

Metodo di Parzen



Si assegna il colume v(**x**), costante e indipendente dalla posizione **x**. Si determina poi il numero di campioni n(**x**) che cadono all'interno di v(**x**).

K-nearest neighboors



Si assegna n(**x**)=K, costante e indipendente dalla posizione **x**. Si determina poi il volume v(**x**) all'interno del quale cadono gli n(**x**) campioni. La stima di p(**x**) è infine ottenuta come:

$$p(\mathbf{x}) = \frac{n(\mathbf{x}) - 1}{N \cdot v(\mathbf{x})}$$

Problema: definizione del volume v(**x**).

Metodi semi-parametrici

- La d.d.p. viene descritta da una curva (polinomi, RBF, …);
- i parametri della curva, θ, non hanno alcun significato statistico (es. coefficienti del polinomio, …);
- normalizzazione della curva per avere:

$$p(x) > 0, \forall x; \quad \int_{-\infty}^{+\infty} p(x)dx = 1$$

Metodi parametrici

Problema

- Data una serie di dati X_i, determinare la densità di probabilità che ha generato gli X_i.

Esempio

- Distribuzione dell'età degli studenti.

Ipotesi

- Distribuzione gaussiana.

Soluzione

- Dai campioni, $\mu=E[X]$, $\sigma^2=Var[X]$.

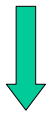
Osservazioni

- Sono stati calcolati i *parametri* di una distribuzione nota (gaussiana); per tale motivo si parla di *stima parametrica*.
- Può essere misurata *a posteriori* l'affidabilità della stima.

Riassunto

- Dati $\mathbf{x_1}$, …, $\mathbf{x_D}$;
- Stima della d.d.p.:
  - Metodi non parametrici;
  - Metodi semiparametrici;
  - Metodi parametrici.

Descrizione della realtà



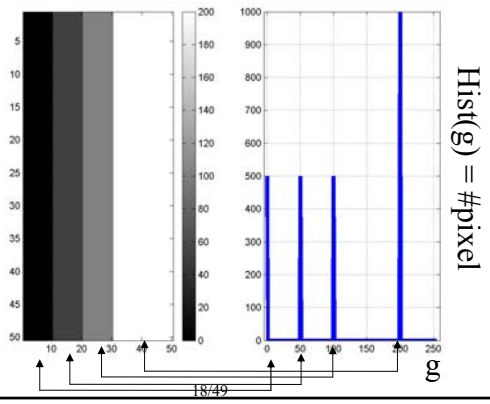
Descrizione ed interpretazione della realtà

Mixture di distribuzioni

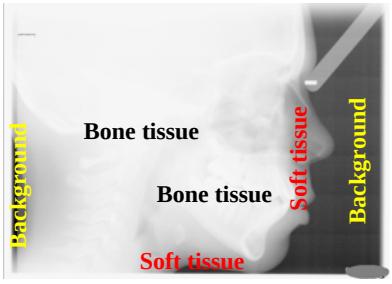
- Non sempre i dati $\mathbf{x}_i$ provengono da… Una sola distribuzione!
- I metodi non parametrici per la stima di p(**x**) restano validi ma…
- I metodi parametrici vanno rivisti: deve essere utilizzato un gruppo (mixture) di d.d.p.!
- Esempio → clusterizzazione immagini radiografica cefalometrica…

Istogramma

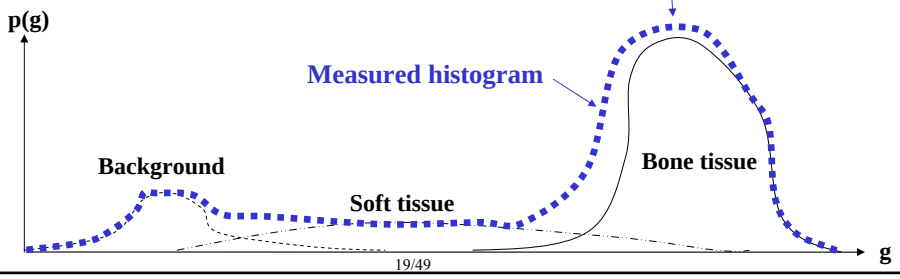
I(x,y) →immagine NRow x NCol, 8 bit (256 livelli di grigio);
Hist(·) → istogramma, 256 componenti;
Hist(g) → # pixel t.c. I(x,y)=g;
Hist(g) / (NRow * NCol) = p(g). → **ddp**



Istogramma radiografia cefalometrica

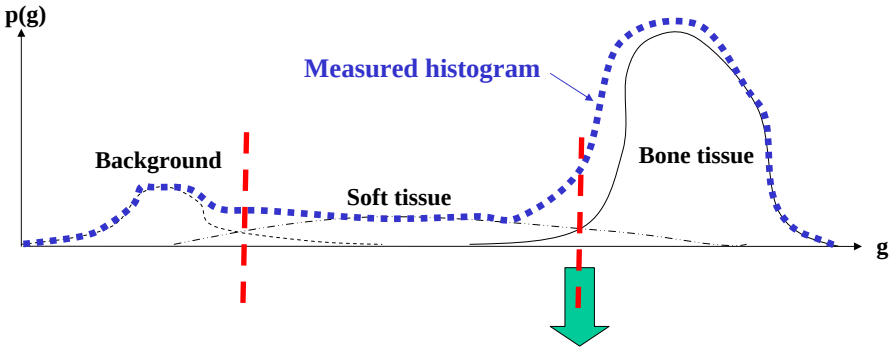


Istogramma tipico immagine radiografica cefalometrica

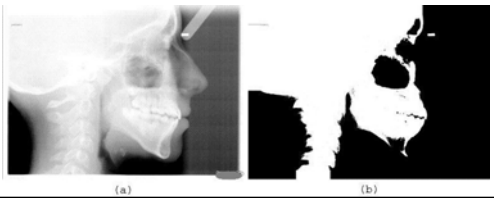


19/49

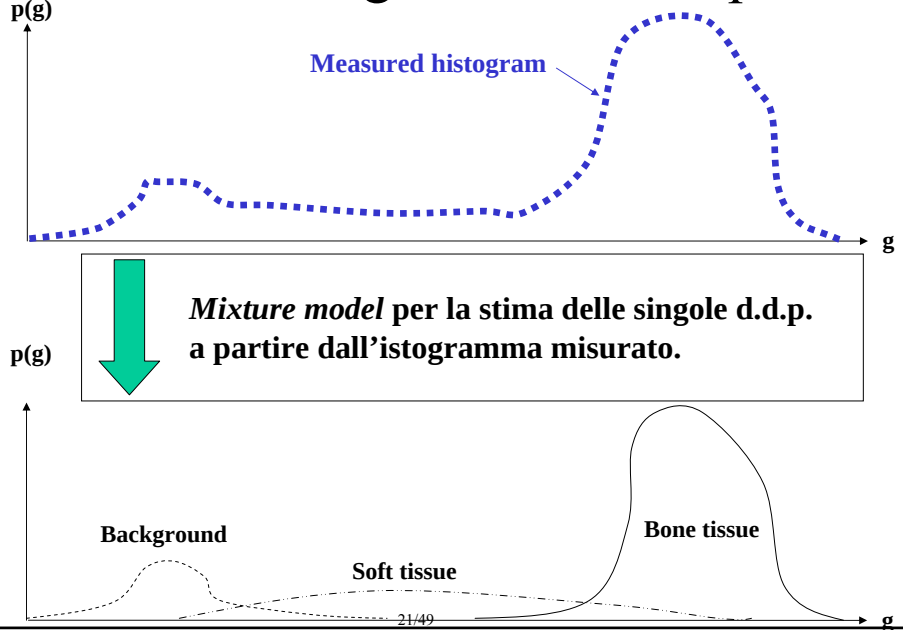
Clusterizzazione basata sull'istogramma



Livelli di grigio diversi identificano tessuti diversi (clutering).



Dall'istogramma alle ddp



21/49

Mixture models

- La d.d.p. complessiva è la combinazione lineare di M d.d.p. di base.

Probabilità del dato x nella j-esima distribuzione

ddp complessiva

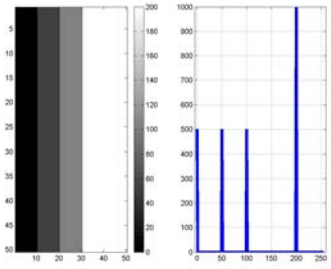
$$p(\mathbf{x}) = \sum_{j=1}^{M} p_\theta(\mathbf{x} \mid j) P(j)$$

Probabilità che il dato x provenga dalla j-esima distribuzione

$$\sum_{j=1}^{M} P(j) = 1; \quad 0 <= P(j) <= 1; \quad \int p_\theta(\mathbf{x} \mid j) = 1$$

22/49

Mixture models



4 distribuzioni

j=0 (nero)	P(0) = 20%;
j=1 (grigio scuro)	P(1) = 20%;
j=2 (grigio chiaro)	P(2) = 20%;
j=3 (bianco)	P(3) = 40%

$$p(\mathbf{x}) = \sum_{j=1}^{M} p_\theta(\mathbf{x} \mid j) P(j)$$

p(100 | j=2) = 100%,

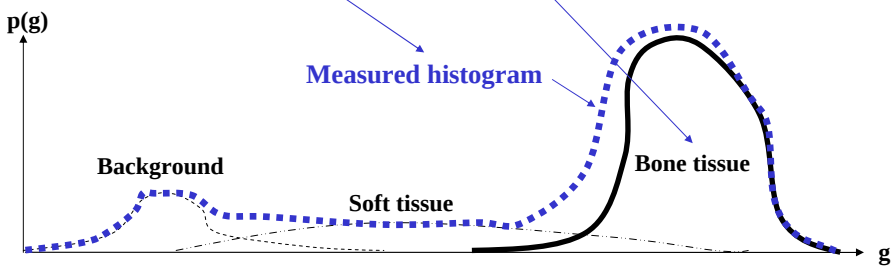
p(99 | j=2) = 0%

p(200 | j=3) = 100%

p(100 | j=3) = 0%

Mixture models

$$p(\mathbf{x}) = \sum_{j=1}^{M} p_\theta(\mathbf{x} \mid j) P(j)$$



P(j) → Es. i pixel "Bone" sono 1/3 dei pixel dell'immagine → P(j) = 0.33;

p(x | j=3) viene determinato sulla base della d.d.p. relativa a j=3 (Bone tissue);

Ad esempio si avrà p(x=0 | j=3) → 0.

Mixture models: incognite

$$p(\mathbf{x}) = \sum_{j=1}^{M} p_\theta(\mathbf{x} \mid j) P(j)$$

Dati

La d.d.p. complessiva, p(x), viene misurata.

La forma delle d.d.p. di base, $p_\theta(x|j)$, viene scelta *a priori* (es. gaussiana).

Incognite

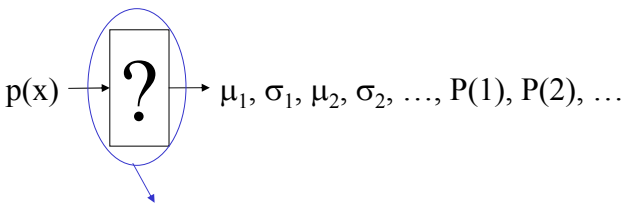
I parametri θ di ogni d.d.p. di base, $p_\theta(x|j)$, devono essere stimati (*stima parametrica* – es. media e varianza di una gaussiana).

Le probabilità per ogni d.d.p. di base, P(j), devono essere stimate.

25/49

Mixture models: framework

- $p(x) = \Sigma_{i=1..M}\, p_\theta(x|j)\, P(j)$, nota;
- $p_\theta(x|j)$ distribuzioni note;
- Parametri θ delle distribuzioni → da stimare;
- Probabilità P(j) di ogni distribuzione → da stimare.



UN ALGORITMO PER LA STIMA DI μ_1, σ_1, μ_2, σ_2, …, P(1), P(2), … A PARTIRE DA p(x) → MASSIMA VEROSIMIGLIANZA + METODO DEL GRADIENTE O ALGORITMO EM.

26/49

Likelihood function

- Campioni:
  $x_1, \ldots, x_D$;
- Parametri stimati della distribuzione:
  θ, es. $\theta = [\mu 1, \sigma 1, \ldots, \mu_M, \sigma_M]$ per distribuzioni gaussiane.
- Funzione di verosimiglianza, L:
  $L = p(x_1|\theta) \cdot p(x_2|\theta) \cdot \ldots p(x_D|\theta)$;

PROBABILITA' DI GENERARE IL DATO X_D AVENDO ASSEGNATO I PARAMETRI θ.

- Logaritmo negativo di L($\log(a \cdot b) = \log(a) \cdot \log(b)$):
  $E = -\log(L) = -\Sigma_j \log[p(x_j|\theta)]$

Il vettore di parametri θ tale per cui L è massima (ovvero -log(L) è minima – E è massima) corrisponde alla massima probabilità di registrare i dati $x_1, \ldots, x_D$ (massima verosimiglianza).

Stima parametri del mixture model

Stima dei parametri del mixture model

$\theta = [\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, P(1), P(2), \ldots]$

Massimizzazione verosimiglianza

$L = p(x_1|\theta) \cdot p(x_2|\theta) \cdot \ldots p(x_D|\theta)$;

Minimizzazione logaritmo negativo di L

$E = -\log(L) = -\Sigma_j \log[p(x_j|\theta)]$

Likelihood function & mixture model

Mixture model:

$p(x) = \Sigma_{j=1..M}\, p_\theta(x \mid j)$

Likelihood function:

$L = L(\theta) = p(x_1 \mid \theta) \cdot p(x_2 \mid \theta) \cdot \ldots\ p(x_D \mid \theta);$

Negative log likelihood function:

$E = E(\theta) = -\log(L) = -\Sigma_{i=1..D} \log [p(x_i \mid \theta)] =$

$= -\Sigma_{i=1..D} \log [\Sigma_{j=1..M}\, p_\theta(x_i \mid j)]$

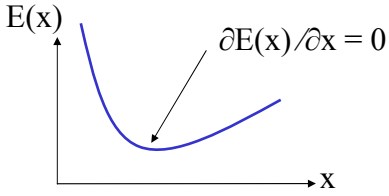
29/49

Ottimizzazione – forma chiusa

- E' necessario minimizzare:

  $E = E(\theta) =$

  $= E(\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, P(1), P(2), \ldots)$
- In forma chiusa:
  - $\partial E/\partial \mu_1 = 0;$
  - $\partial E/\partial \mu_2 = 0$
  - …
  - $\partial E/\partial \sigma_1 = 0$
  - …
  - $\partial E/\partial P(1) = 0$
  - …



Il sistema non è risolvibile in maniera diretta (sistema di equazioni non lineari)…

30/49

Ottimizzazione – metodi iterativi

- Se un sistema di equazioni non è risolvibile in modo diretto…
- … METODI ITERATIVI!

- Per i mixture models:
  - Metodo del gradiente (lento, poco usato);
  - EM (Expectation Maximization, robusto e veloce, molto usato).

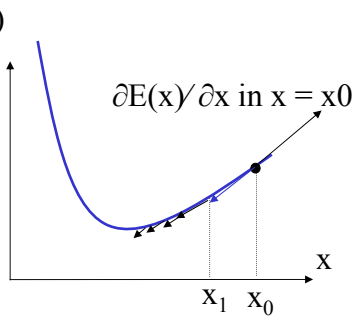
Metodo del gradiente

1) Inizializzo: $x = x_0$;

2) Calcolo $\partial E(x) / \partial x$ in $x = x_0$;

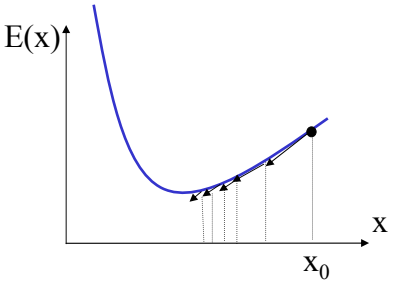
3) Aggiorno $x_1 = x_0 - \eta \cdot \partial E(x) / \partial x$, $0<h<1$

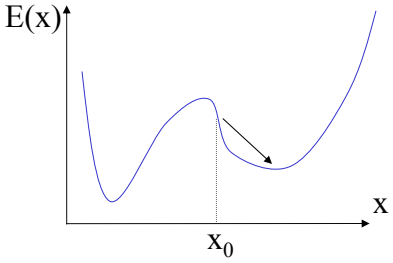
4) Itero 2) e 3)



Problemi metodo del gradiente

- Problemi:
  - Scelta di η;
  - il metodo è lento all'approssimarsi della soluzione;
  - possibili minimi locali.





33/49

metodo del gradiente

- C'è inoltre un constraint:

$$\sum_{j=1}^{M} P(j) = 1$$

$$P(j) = \frac{\exp(\gamma_j)}{\sum_{k=1}^{M} \exp(\gamma_j)} \qquad \text{(softmax function)}$$

$$\frac{\partial E}{\partial \gamma_j} = -\sum_{n=1}^{N} \left\{ P(j \mid \mathbf{x}^{\mathbf{n}}) - P(j) \right\}$$

34/49

Algoritmo EM

- Funzione da minimizzare (incognita θ):

$$E = -\ln(L) = -\ln \prod_{n=1}^{N} p_{\vartheta}(x^n) = -\sum_{n=1}^{N} \ln p_{\vartheta}(x^n)$$

- Per ogni iterazione, i parametri vengono aggiornati (old → new, indice θ omesso):

$$E^{new} - E^{old} = -\sum_{n=1}^{N} \ln p^{new}(x^n) - \left[-\sum_{n=1}^{N} \ln p^{old}(x^n) \right] = -\sum_{n=1}^{N} \ln \left[\frac{p^{new}(x^n)}{p^{old}(x^n)} \right]$$

Algoritmo EM

$$E^{new} - E^{old} = -\sum_{n=1}^{N} \ln p^{new}(x^n) - \left[-\sum_{n=1}^{N} \ln p^{old}(x^n) \right] = -\sum_{n=1}^{N} \ln \left[\frac{p^{new}(x^n)}{p^{old}(x^n)} \right]$$

Ricordando che:

$$p(x) = \sum_{j=1}^{M} P(j) \cdot p(x \mid j)$$

Si ottiene:

$$E^{new} - E^{old} = \sum_{n=1}^{N} -\ln \left[\frac{\sum_{j=1}^{M} P^{new}(j) p^{new}(x^n \mid j)}{p^{old}(x^n)} \cdot \frac{P^{old}(j \mid x^n)}{P^{old}(j \mid x^n)} \right]$$

1

Algoritmo EM

- La disuguaglianza di Jensen dice che:

$$dati \quad \lambda_j^2 \quad t.c. \quad \sum_{j=1}^{M} \lambda_j^2 = 1$$

$$\ln\left(\sum_{j=1}^{M} \lambda_j^2 K_j\right) \geq \sum_{j=1}^{M} \lambda_j^2 \ln\left(K_j\right) \Rightarrow$$

$$\Rightarrow -\ln\left(\sum_{j=1}^{M} \lambda_j^2 K_j\right) \leq -\sum_{j=1}^{M} \lambda_j^2 \ln\left(K_j\right)$$

Algoritmo EM

$$\sum_{j=1}^{M} \lambda_j^2 = 1 \quad \longleftrightarrow \quad \sum_{j=1}^{M} P^{old}\left(j \mid x^n\right) = 1, \forall n$$

Jensen — Mixture model

E' possibile applicare la disuguaglianza di Jensen ai mixture model, ove i vari $P^{old}(j \mid x^n)$ giocano il ruolo dei λ_j^2.

Algoritmo EM

- Applicando Jensen:

$$-\ln\left(\sum_{j=1}^{M}\lambda_j^2 K_j\right) \le -\sum_{j=1}^{M}\lambda_j^2 \ln\left(K_j\right)$$

$$E^{new} - E^{old} = \sum_{n=1}^{N} -\ln\left[\frac{\sum_{j=1}^{M} P^{new}(j)p^{new}\left(x^n \mid j\right)}{p^{old}\left(x^n\right)} \cdot \frac{P^{old}\left(j \mid x^n\right)}{P^{old}\left(j \mid x^n\right)}\right]$$

Algoritmo EM

$$E^{new} - E^{old} = \sum_{n=1}^{N} -\ln\left[\frac{\sum_{j=1}^{M} P^{new}(j)p^{new}\left(x^n \mid j\right)}{p^{old}\left(x^n\right)} \cdot \frac{P^{old}\left(j \mid x^n\right)}{P^{old}\left(j \mid x^n\right)}\right]$$

$$\le -\sum_{j=1}^{M} P^{old}\left(j \mid x^n\right) \cdot \log\left[\frac{P^{new}(j) \cdot p^{new}\left(x^n \mid j\right)}{p^{old}\left(x^n\right) \cdot P^{old}\left(j \mid x^n\right)}\right]$$

Algoritmo EM

$$E^{new} - E^{old} \leq -\sum_{n=1}^{N}\sum_{j=1}^{M} P^{old}\left(j \mid x^n\right)\ln\left\{\frac{P^{new}(j)p^{new}\left(x^n \mid j\right)}{p^{old}\left(x^n\right)\cdot P^{old}\left(j \mid x^n\right)}\right\}$$

$$E^{new} \leq E^{old} + Q$$

Minimizzando Q si minimizza E!

$P^{old}(x^n)$, $P^{old}(j \mid x^n)$ → costanti

$Q = Q(\theta^{new})$

Algoritmo EM

Eliminando i termini costanti nella somma (!), è sufficiente minimizzare ad ogni iterazione:

$$\tilde{Q} = -\sum_{n=1}^{N}\sum_{j=1}^{M} P^{old}\left(j \mid x^n\right)\ln\left\{P^{new}(j)p^{new}\left(x^n \mid j\right)\right\}$$

Tenendo inoltre conto del fatto che:

$$\sum_{j=1}^{M} P^{new}\left(j \mid x^n\right) = 1, \forall n$$

Algoritmo EM

Si minimizza allora (metodo dei moltiplicatori di Lagrange):

$$f = \widetilde{Q} + \psi\left(\sum_{j=1}^{M} P^{new}(j) - 1\right)$$

Cioè:

$$\begin{cases} \dfrac{\partial f}{\partial \mu_j^{new}} = 0 \\ \dfrac{\partial f}{\partial \sigma_j^{new}} = 0 \\ \dfrac{\partial f}{\partial P^{new}(j)} = 0 \end{cases} \longrightarrow$$

Da questo sistema possono essere ricavate le equazioni per l'aggiornamento dei parametri del mixture model ad ogni iterazione.

Algoritmo EM

- Aggiornamento P(j):

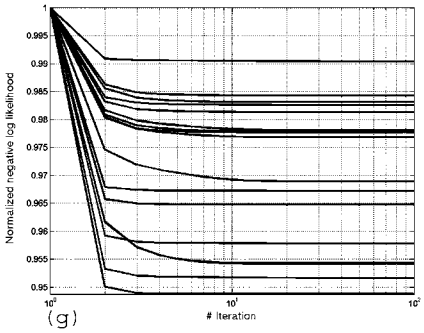
$$P^{new}(j) = \frac{1}{N}\sum_{n=1}^{N} P^{old}(j \mid x)$$

- Nel caso di ddp gaussiane:

$$\mu_j^{new} = \frac{\sum_{n=1}^{N} P^{old}(j \mid x^n) x^n}{\sum_{n=1}^{N} P^{old}(j \mid x^n)} \qquad (\sigma_j^{new})^2 = \frac{\sum_{n=1}^{N} P^{old}(j \mid x^n)(x^n - \mu_j)^2}{\sum_{n=1}^{N} P^{old}(j \mid x^n)}$$

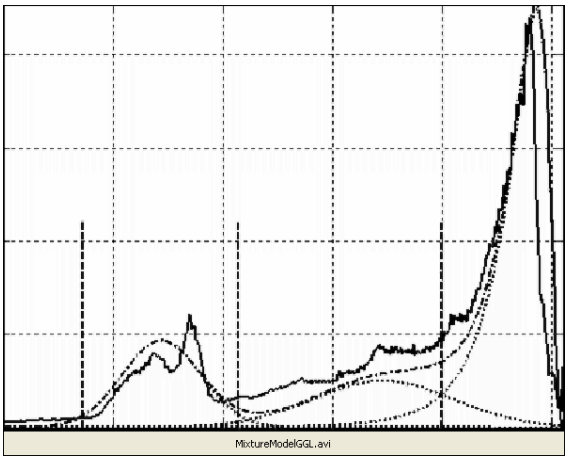
Algoritmo EM

- Durante l'addestramento del mixture model:
  - L cresce (E diminuisce);
  - I parametri θ si stabilizzano;
  - Il mixture model si adatta alla p(x).



45/49

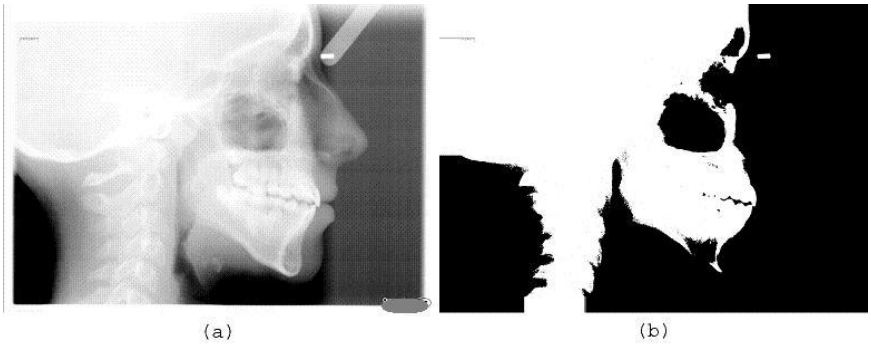
Clusterizzazione basata sull'istogramma



46/49

Risultato

- Clusterizzazione immagine in 3 zone (background, soft tissue, bone tissue).



Applicazioni mixture models

- Analisi statistica;
- Machine learning;
- Feature matching;
- Ricostruzione di superfici;
- …

Bibliografia

- Christopher M. Bishop, Neural networks for pattern recognition. Clarendon Press, Oxford, capitolo 2.
  - 2.1 Metodi parametrici per la stima di ddp
  - 2.2 Massima verosimiglianza
  - 2.5 Metodi non parametrici per la stima di ddp
  - 2.6 Mixture models e algoritmo EM