Intelligenza Artificiale II

Ragionamento probabilistico

Apprendimento

Marco Piastra

Parte 3

Osservazioni

Maximum Likelihood Estimation

Dati e probabilità per i modelli grafici

Maximum a Posteriori Estimation

Apprendimento di modelli grafici

Eventi ed osservazioni

- Eventi

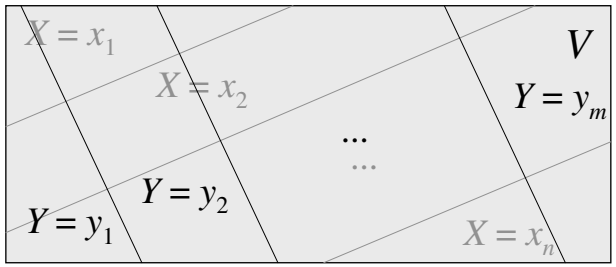
  Un **evento** è un sottoinsieme di *mondi possibili*

  Un evento si **verifica** quando il *mondo attuale* appartiene al corrispondente sottoinsieme

- Variabili aleatorie multiple

  In generale, in una rappresentazione probabilistica si hanno più variabili aleatorie

  Ciascuna combinazione di valori delle variabili aleatorie è un *evento*

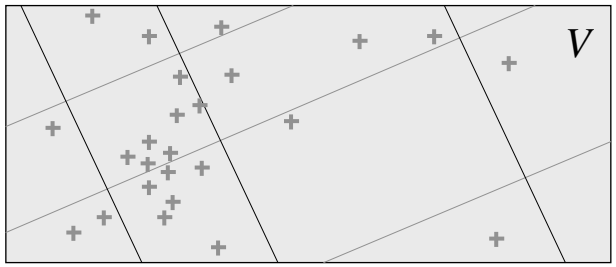


- Osservazioni (dati)

  Un'osservazione completa (dato) è un mondo possibile completamente osservato

  La combinazione dei valori della variabile aleatoria è completamente nota

  In che modo un'insieme di osservazioni (complete o meno) influenza la misura di probabilità?



Indipendenza ed osservazioni

Preludio: sulla notazione

Un'osservazione può essere l'esito di un esperimento o un test

Per convenzione, si preferisce definire un set di variabili aleatorie separato per ciascun esperimento o test

Ad esempio, se il modello probabilistico utilizza le variabili aleatorie $\{X, Y\}$, si indicano con $\{X_1, Y_1\}, \{X_2, Y_2\}, ..., \{X_n, Y_n\}$ i possibili esiti di n esperimenti o test distinti

- *Osservazioni indipendenti, stessa distribuzione di probabilità*

  Independent, Identically Distributed (IID) random variables

Definizione

Una sequenza o insieme di variabili aleatorie $\{X_1, X_2, \ldots, X_n\}$ è IID se:

1) $<X_i \perp X_j>,\ i \neq j$ (mutua indipendenza)
2) $P(X_i) = P(X_j),\ i \neq j$ (identica distribuzione)

La definizione si estende facilmente a sequenze formate da v.a. multiple

Cautela

Una sorta di idealizzazione: spesso le v. a. che descrivono risultati di esperimenti non sono IID

Esempio: misurazioni diverse su pazienti diversi possono essere viste come IID, misurazioni sullo stesso paziente in tempi diversi non sono IID

Maximum Likelihood Estimation (MLE)

Elementi di base

Un modello probabilistico parametrico $P(X)$, con parametri θ

θ sono i parametri caratterizzanti del modello, p.es. la probabilità di avere testa o croce

Un insieme di osservazioni $D = \{X_1, X_2, \ldots, X_n\}$ che si assumono essere IID

- *Likelihood function* (funzione di verosimiglianza)

  Una funzione della probabilità condizionale $P(D \mid \theta)$ derivata dal modello $P(X)$

  $$L(\theta \mid D) \;=\; P(D \mid \theta) = P(X_1, X_2, \ldots, X_n \mid \theta)$$

  Dove $P(D \mid \theta)$ è la probabilità che i parametri θ, visti come v.a., generino le osservazioni D
  Utilizzando l'ipotesi che $\{X_1, X_2, \ldots, X_n\}$ sia IID:

  $$L(\theta \mid D) \;=\; P(X_1 \mid \theta) P(X_2 \mid \theta) \ldots P(X_n \mid \theta) = \prod_i P(X_i \mid \theta)$$

- *Maximum Likelihood Estimation*

  $$\theta^*_{ML} \;=\; \arg\max_\theta L(\theta \mid D)$$

  Più comoda per i calcoli la forma logaritmica (*Log-Likelihood*)

  $$\ell(\theta \mid D) \;=\; \log L(\theta \mid D) = \log \prod_i P(X_i \mid \theta) = \sum_i \log P(X_i \mid \theta)$$

  $$\theta^*_{ML} \;=\; \arg\max_\theta \ell(\theta \mid D)$$

Esempio: lanci di una moneta (*Bernoulli Trials*)

- Descrizione

  Test: lancio di una moneta X, di caratteristiche non note. ($X = 1$ testa, $X = 0$ croce)

  Modello: $P(X = 1) = \theta,\ P(X = 0) = 1 - \theta$

  Osservazioni: una sequenza $\{X_1, X_2, \ldots, X_n\}$ (p.es. $D = \{X_1 = 1, X_2 = 1, X_3 = 0 \ldots\}$)

- (*Log-*)*Likelihood Function*

$$\ell(\theta \mid D) = \log P(D \mid \theta) = \log P(\{X_i\} \mid \theta) = \log \prod_i P(X_i \mid \theta) = \sum_i \log P(X_i \mid \theta)$$

Likelihood per P*:*

$$P(X \mid \theta) = \theta^{[X=1]}(1-\theta)^{[X=0]} \quad \text{dove:} \quad [X_i = v] = \begin{cases} 1 & se \quad X_i = v \\ 0 & se \quad X_i \neq v \end{cases}$$

$$\ell(\theta \mid D) = \sum_i \log\left(\theta^{[X_i=1]}(1-\theta)^{[X_i=0]}\right) = \log\theta \sum_i [X_i = 1] + \log(1-\theta) \sum_i [X_i = 0]$$

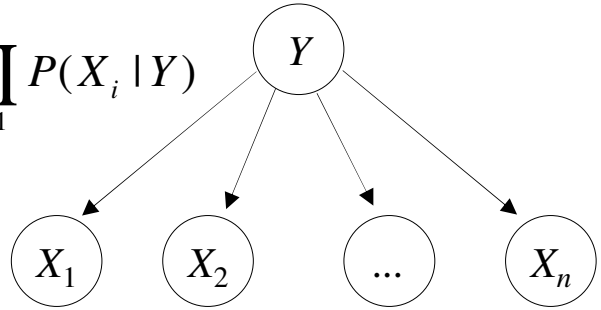
$$= \log\theta\, N_{X=1} + \log(1-\theta) N_{X=0}$$

(dove p.es. $N_{X=1}$ è il numero di $X_i = 1$ nella sequenza D)

- *Maximum Likelihood Estimation*

$$\frac{\partial \ell}{\partial \theta} = \frac{N_{X=1}}{\theta} + \frac{N_{X=0}}{(1-\theta)} \qquad \frac{\partial \ell}{\partial \theta} = 0 \quad \Rightarrow \quad \theta^*_{ML} = \frac{N_{X=1}}{N_{X=1} + N_{X=0}}$$

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Apprendimento del filtro *anti-spam*

- Descrizione

  Modello parametrico: $P(Y=k) = \pi_k$, $P(X_i = j \mid Y = k) = \eta_{ijk}$

  (i valori delle probabilità condizionali nel modello grafico)

  Osservazioni: una sequenza di set di valori, ottenuti da messaggi ricevuti e classificati

  p.es. $D = \{\{Y_1 = 1, X_{11} = 1, X_{12} = 1, ..., X_{1n} = 0\}, \{Y_2 = 0, X_{21} = 0, X_{22} = 1, ..., X_{2n} = 1\}, ...\}$

- *Likelihood Function*

  Sequenza di messaggi

$$L(\{\pi_k, \eta_{ijk}\} \mid D) = P(D \mid \theta) = P(\{\{Y_m = y_m, X_{im} = x_m\}\} \mid \{\pi_k, \eta_{ijk}\})$$

$$= \prod_m P(\{Y_m = y_m, X_{im} = x_m\} \mid \{\pi_k, \eta_{ijk}\}) \quad \text{(i messaggi sono IID)}$$

$$= \prod_m P(Y_m = y_m \mid \{\pi_k, \eta_{ijk}\})\, P(\{X_{im} = x_{im}\} \mid Y_m = y_m, \{\pi_k, \eta_{ijk}\}) \quad \text{(fattorizzazione)}$$

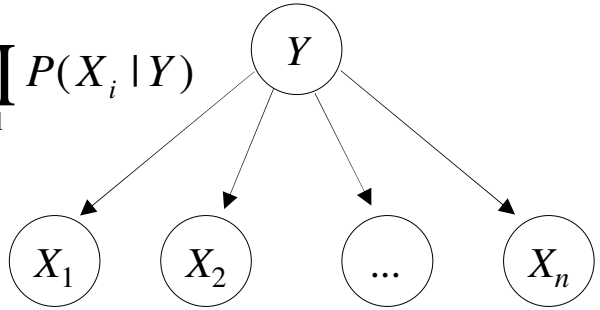
$$= \prod_m P(Y_m = y_m \mid \{\pi_k\})\, P(\{X_{im} = x_{im}\} \mid Y_m = y_m, \{\eta_{ijk}\}) \quad \text{(indipendenza cond.)}$$

$$= \prod_m P(Y_m = y_m \mid \{\pi_k\}) \prod_i P(X_{im} = x_{im} \mid Y_m = y_m, \{\eta_{ijk}\}) \quad (\langle X_i \perp X_j, Y \rangle)$$

- *Log-Likelihood Function*

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \log P(Y_m = y_m \mid \{\pi_k\}) + \sum_m \sum_i \log P(X_{im} = x_{im} \mid Y_m = y_m, \{\eta_{ijk}\})$$

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Apprendimento del filtro *anti-spam*

- *Log-Likelihood Function*

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \log P(Y_m = y_m \mid \{\pi_k\}) + \sum_m \sum_i \log P(X_{im} = x_{im} \mid Y_m = y_m, \{\eta_{ijk}\})$$

Forma alternativa per P*:*

$$P(Y = k \mid \{\pi_k\}) = \pi_k = \prod_k \pi_k^{[Y=k]}$$

$$P(X_i = j \mid Y_m = k, \{\eta_{ijk}\}) = \eta_{ijk} = \prod_j \prod_k \eta_{ijk}^{[X_i=j][Y=k]}$$

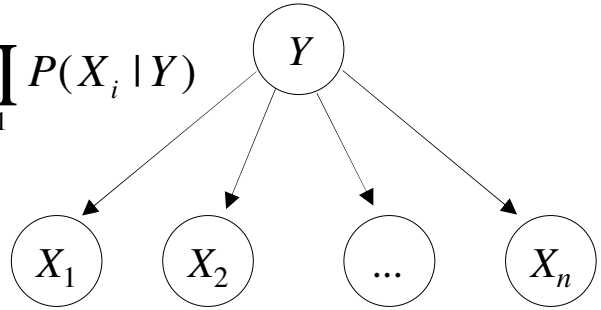
(*Algebraic Follies!*)

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{im} = j][Y_m = k] \log \eta_{ijk}$$

- *Maximum Likelihood Estimation*

  I due termini principali della funzione possono essere massimizzati separatamente

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Apprendimento del filtro *anti-spam*

- *Maximum Likelihood Estimation*

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{im} = j][Y_m = k] \log \eta_{ijk}$$

Ottimizzazione del primo termine:

moltiplicatore di Lagrange

$$\ell^*(\{\pi_k\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \lambda(1 - \sum_k \pi_k)$$

$$\frac{\partial \ell^*}{\partial \pi_k} = \frac{\sum_m [Y_m = k]}{\pi_k} - \lambda$$

numero di messaggi in D classificati come k

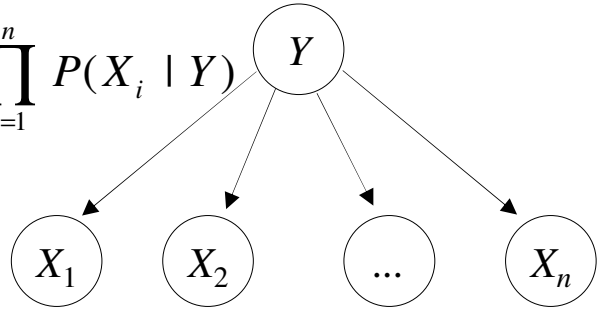
$$\frac{\partial \ell^*}{\partial \pi_k} = 0 \;\Rightarrow\; \pi_k = \frac{N_{Y=k}}{\lambda}$$

numero complessivo di messaggi in D

$$\sum_k \pi_k = 1 \;\Rightarrow\; \sum_k \frac{N_{Y=k}}{\lambda} = 1 \;\Rightarrow\; \lambda = \sum_k N_{Y=k} = N_D$$

$$\pi_k^* = \frac{N_{Y=k}}{N_D}$$ (*Maximum Likelihood Estimator* di π_k)

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Apprendimento del filtro *anti-spam*

- *Maximum Likelihood Estimation*

$$\ell(\{\pi_k, \eta_{ijk}\} \mid D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{im} = j][Y_m = k] \log \eta_{ijk}$$

Ottimizzazione del secondo termine:

$$\ell^*(\{\eta_{ijk}\} \mid D) = \sum_m \sum_i \sum_j \sum_k [X_{im} = j][Y_m = k] \log \eta_{ijk} + \sum_i \sum_k \lambda_{ik} (1 - \sum_j \eta_{ijk})$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = \frac{\sum_m [X_{im} = j][Y_m = k]}{\eta_{ijk}} - \lambda_{ik}$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = 0 \Rightarrow \eta_{ijk} = \frac{N_{X_i = j, Y = k}}{\lambda_{ik}}$$

$$\sum_j \eta_{ijk} = 1 \Rightarrow \sum_j \frac{N_{X_i = j, Y = k}}{\lambda_{ik}} = 1 \Rightarrow \lambda = \sum_j N_{X_i = j, Y = k} = N_{Y = k}$$

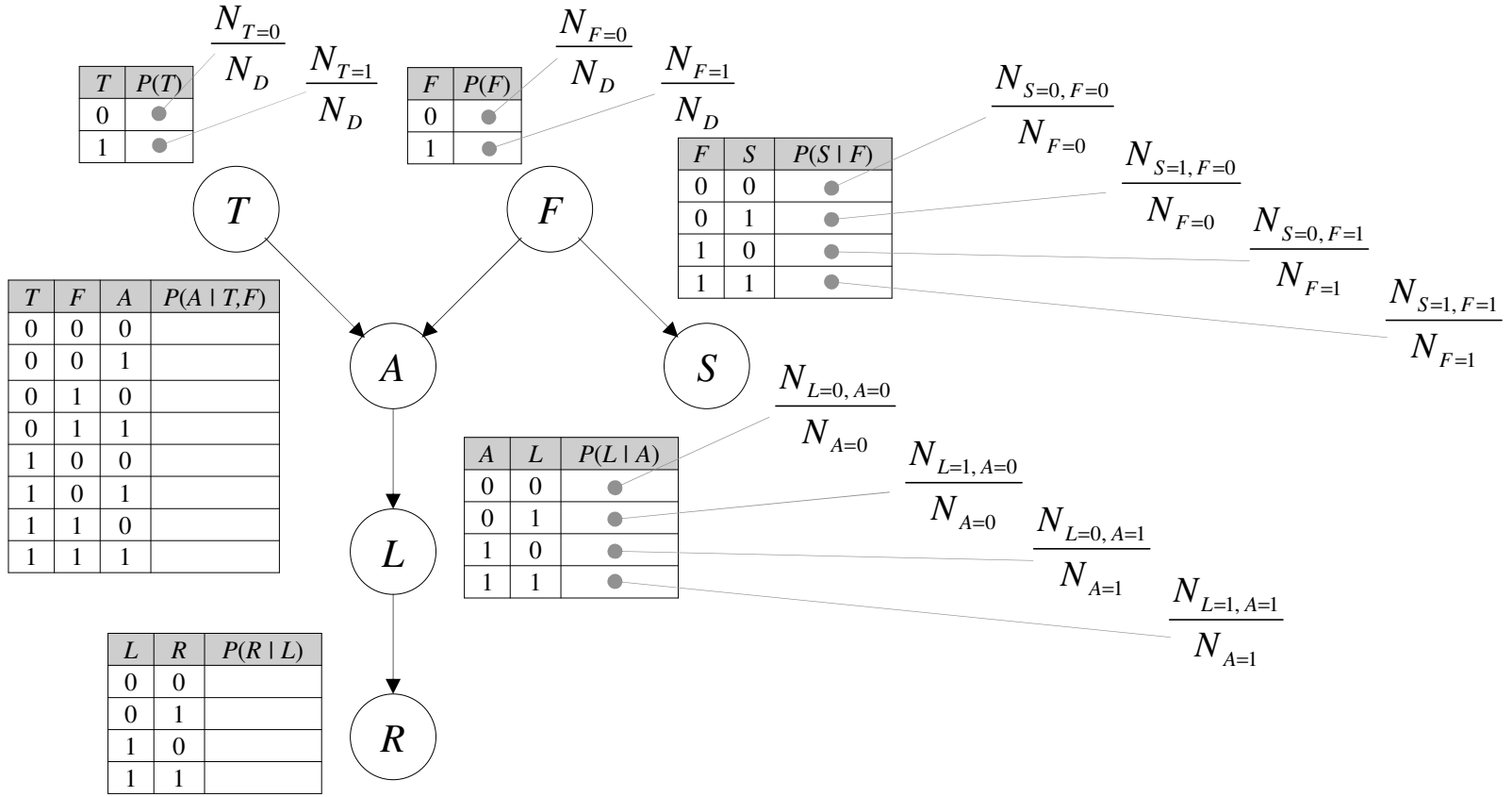
$$\eta^*_{ijk} = \frac{N_{X_i = j, Y = k}}{N_{Y = k}}$$ (*Maximum Likelihood Estimator* di η_{ijk})

Apprendimento delle probabilità di un modello grafico

La *Maximum Likelihood Estimation* del filtro anti-spam si estende al caso più generale

Modello parametrico: il modello grafico del *fire alarm*, con le probabilità come parametri

Osservazioni: una sequenza di set di valori, ottenuta da eventi completamente osservati



Apprendimento bayesiano

- *Maximum a Posteriori Estimation* (*MAP*)

  Invece di massimizzare la sola *likelihood function*, si massimizza la probabilità a posteriori

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\sum_{\theta} P(D|\theta)P(\theta)}$$

  Il che equivale a massimizzare il termine:

$$P(D|\theta)\,P(\theta)$$

  Passando al logaritmo:

$$\log P(D|\theta) + \log P(\theta)$$

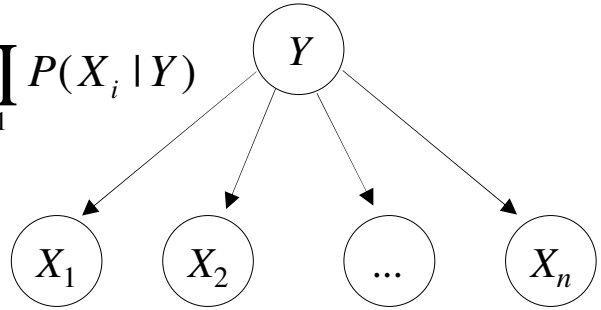
  termine 'a priori'

  Log-Likelihood Function

Vantaggi:

- Maggiore regolarità: alcune combinazioni potrebbero essere mancanti in D
- Una formula per l'apprendimento incrementale:
  i termini a priori possono rappresentare la conoscenza *prima* delle osservazioni D

$$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i \mid Y)$$



Apprendimento del filtro *anti-spam*

- *Maximum a Posteriori Estimation*

  Ripetendo il procedimento di ottimizzazione per:

  $$\log P(D \mid \theta) + \log P(\theta)$$

  Si ottiene:

  $$\pi_k^* = \frac{\alpha_k + N_{Y=k}}{\sum_k \alpha_k + N_D}$$ (*MAP Estimator* di π_k)

  $$\eta_{ijk}^* = \frac{\alpha_{ijk} + N_{X_i=j, Y=k}}{\sum_j \alpha_{ijk} + N_{Y=k}}$$ (*MAP Estimator* di η_{ijk})

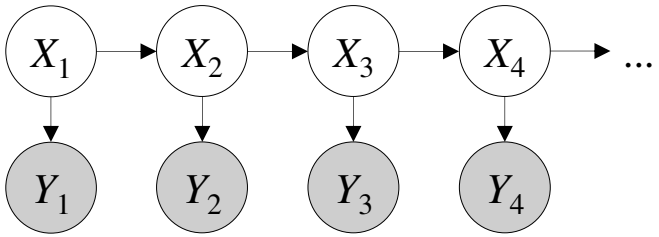
  Dove

  $$\frac{\alpha_k}{\sum_k \alpha_k} \qquad \frac{\alpha_{ijk}}{\sum_j \alpha_{ijk}}$$

  sono le probabilità a priori, sotto forma di frequenze relative
  (o i parametri del modello grafico *prima* dell'arrivo di un nuovo lotto di messaggi D)

Osservazioni incomplete

Esempio: modello 'Hidden Markov'



In genere, i nodi X_i sono *hidden*, nel senso di *non-osservabili*

$$P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = P(X_1)\, P(Y_1 \mid X_1) \prod_{i=2}^{n} P(X_i \mid X_{i-1})\, P(Y_i \mid X_i)$$

- Problema

  Definire i parametri $\{\theta_k\}$ del modello partendo dalle sole osservazioni $\{Y_j\}$

  In altri termini, si tratta di trovare il *MLE* di

  $L(\{\theta_k\} \mid \{X_j\}, \{Y_j\}) = P(\{X_i\}, \{Y_j\} \mid \{\theta_k\})$

  quando $\{X_j\}$ non è noto

  In generale, è un problema di ottimizzazione molto difficile

Algoritmo *Expectation-Maximization* (*EM Algorithm*)

- Algoritmo *Expectation-Maximization* (*EM Algorithm*)

  Problema: definire i parametri $\{\theta_k\}$ del modello partendo dalle sole osservazioni $\{Y_j\}$

  1) Assegnazione di un valore iniziale (p.es. *casuale*) ai parametri $\{\theta_k\}$
  2) Calcolo della distribuzione $P(\{X_i\} \mid \{Y_j\}, \{\theta_k\})$ (*E-step*)

$$P(\{X_i\}|\{Y_j\},\{\theta_k\}) = \frac{P(\{Y_j\}|\{X_i\},\{\theta_k\})\,P(\{X_i\}|\{\theta_k\})}{\sum_{\{X_i\}} P(\{Y_j\}|\{X_i\},\{\theta_k\})\,P(\{X_i\}|\{\theta_k\})}$$

  3) *MLE* del <u>valor medio</u> di $L(\{\theta_k\} \mid \{X_i\}, \{Y_j\})$, data la distribuzione $P(\{X_i\} \mid \{Y_j\}, \{\theta_k\})$ (*M-step*)

$$\{\theta_k^*\} = \arg\max_{\{\theta_k\}} \sum_{\{X_i\}} P(\{X_i\}|\{Y_j\},\{\theta_k\})\,P(\{X_i\},\{Y_j\}|\{\theta_k\})$$

$$= \arg\max_{\{\theta_k\}} \sum_{\{X_i\}} P(\{X_i\}|\{Y_j\},\{\theta_k\}) \log P(\{X_i\},\{Y_j\}|\{\theta_k\})$$

  4) Ritorno al passo 2), usando $\{\theta_k^*\}$, fino a alla convergenza

  Un metodo di ottimizzazione iterativa ed alternata:
  ad ogni passo si utilizzano i *MLE* del passo precedente

  Garantisce la convergenza ad un ottimo *locale* di $L(\{\theta_k\} \mid \{X_j\}, \{Y_j\})$

  In generale, il risultato dipende dai valori casuali inizialmente assegnati a $\{\theta_k\}$

Apprendimento delle strutture

L'apprendimento della struttura di un modello grafico dai dati

- Problema:
  come si misura l'adeguatezza di un modello grafico alle osservazioni

  Si può vedere come l'ottimizzazione di

  $L(Model, \theta_{Model} | D) = P(D | Model, \theta_{Model})$

  Ovviamente non è un problema facile

  Si può affrontare come un problema di ricerca in uno spazio,
  usando la metrica come *euristica*

  Si possono usare metodi di calcolo evolutivo

  Caveat:

  Occorre considerare anche la semplicità del modello

  Ad esempio $L(Model, \theta_{Model} | D)$ sarà massima per

  $$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2)P(X_4 | X_1, X_2, X_3)$$

  (Dipendenza completa)

Minimum Description Length (MDL)

Lunghezza della descrizione del modello + lunghezza descrizione dei dati, dato il modello

$$DL = DL_{Model} + DL_{Data|Model}$$

- Lunghezza della descrizione del modello

Bit usati per un val. di probabilità — Valori assunti da X_i

$$DL_{Model} = \sum_{i=1}^{n} \left[|\Pi_{X_i}| \log_2 n + d(v_i - 1) \prod_{X_j \in \Pi_{X_i}} v_j \right]$$

Insieme dei $parents(X_i)$ — Valori assunti dal $parent\ X_j$

- Lunghezza della descrizione dei dati, dato il modello

Frequenza relativa nei dati D

$$DL_{Data|Model} = -\sum_{i=1}^{n} \sum_{X_i, \Pi_{X_i}} F_D(X_i, \Pi_{X_i}) \log_2 \frac{F_D(X_i, \Pi_{X_i})}{F_D(X_i) F_D(\Pi_{X_i})}$$

Insieme dei $parents(X_i)$

Sono tutte sommatorie rispetto a valori locali calcolati per ciascun nodo

Evolutionary Programming per i modelli

(Wong e Leung, 1999)

- Individui della popolazione

  Modelli grafici

- Funzione di fitness

  Minimum Description Length

- Operatori genetici

  Solo mutazioni