PSC: Progettazione di sistemi di controllo **a.a. 2010-2011**

Lezione 12 — 15 Novembre

Docente: Luca Schenato *Stesori: Barbazza R., Cordioli M., Milani B.*

12.1 HIDDEN MARKOV MODELS

Si parte dall'esigenza di voler descrivere il comportamento di sistemi a stati finiti in cui le probabilità di transizione tra stati non dipendono dal particolare valore assunto dal tempo, ma soltanto dalla coppia di stati interessati nella transizione.
In una catena di Markov ogni stato può essere messo in corrispondenza con un evento fisico osservabile, per cui una situazione reale (ovvero una successione di eventi) può essere modellizzata attraverso una sequenza ordinata di stati assunti dalla catena in determinati istanti temporali.
Si può estendere il modello di Markov in modo da contemplare la situazione in cui l'osservazione sia legata allo stato attraverso una funzione probabilistica: in altre parole, il modello risultante è un processo stocastico che emette simboli osservabili appoggiandosi su un altro processo stocastico non osservabile (nascosto) che determina il cambio di stato. Le probabilità di emissione dei simboli cambiano da uno stato all'altro.
Un modello (catena) di Markov nascosto può essere descritto attraverso cinque elementi:

- l'insieme dei possibili stati che può assumere la catena di Markov

$$x \in V = \{v_1, \dots, v_N\}$$

- il vettore delle probabilità che la catena si trovi in un particolare stato all'istante iniziale

$$\pi_0 \in \mathbb{R}^{N\times 1}, \quad \pi_0(i) \geq 0 \ \forall i \quad \text{e} \quad \sum_{i=1}^{N} \pi_0(i) = 1$$

$$\pi_0(i) \ = \ p(x_0 = i) \ = \ p(x_0 = v_i)$$

- la matrice di transizione

$$A \in \mathbb{R}^{N\times N}$$

  il cui elemento di posizione $[A]_{ij} = a_{ij}$ indica la probabilità che la catena si trovi nello stato v_j condizionatamente al fatto che all'istante precedente si trovasse nello stato v_i, ovvero

$$p\big(x_{t+1} = j \mid x_t = i\big) = a_{ij} \geq 0;$$

Per tale matrice vale la relazione $\sum_{j=1}^{N} a_{ij} = 1$, ovvero la catena di Markov dallo stato i transita (all'istante successivo) con probabiltà 1 in uno degli N stati (può rimanere anche in i). Ponendo $\mathbb{1}^T = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}$ la condizione precedente si può riscrivere come $A\,\mathbb{1} = \mathbb{1}$.
Le matrici con elementi non negativi le cui righe sommano a 1 si definiscono matrici stocastiche per righe.

- l'insieme delle uscite osservabili

$$y \in W = \big\{w_1, \ldots, w_M\big\}$$

- la matrice $C \in \mathbb{R}^{N\times M}$, cosiddetta matrice delle emissioni, i cui elementi $\big[C\big]_{ij} = c_{ij}$ rappresentano la probabilità che l'uscita osservata sia w_j condizionatamente al fatto che la catena si trova nello stato v_i, ovvero

$$p\big(y_t = j \mid x_t = i\big) = c_{ij} \geq 0$$

Anche per la matrice C vale la proprietà $\sum_{j=1}^{M} c_{ij} = 1$ che può essere espressa come $C\,\mathbb{1}_M = \mathbb{1}_N$.

Nella descrizione dell'evoluzione nel tempo di una catena di Markov nascosta si incontrano i concetti di cammino e osservazione: il primo è un insieme ordinato di stati in cui la catena si viene a trovare sequenzialmente mentre l'osservazione è un insieme ordinato di simboli (uscite) che vengono osservate in sequenza.
Indicando con x_t lo stato in cui si trova la catena all'istante t, mentre con $x^t = \big(x_0, \cdots, x_t\big)$ la sequenza di stati in cui la catena si viene a trovare in $t+1$ istanti successivi e con $Y^h = \big(y_0, \cdots, y_h\big)$ la successione di uscite osservate in $h+1$ istanti successivi; in queste ipotesi si possono formulare diversi problemi sull'evoluzione della catena di Markov:

- definendo $p\big(x_t \mid Y^h\big) = \widehat{p}_{t|h} \in \mathbb{R}^N$ come la probabilità che la catena si trovi all'istante t nello stato x_t sulla base delle osservazioni Y^h, si può cercare lo stato che massimizza tale probabilità (calcolo della probabilità 'forward'), ovvero

$$\widehat{x}_{t|h} = \arg\max_i\ \widehat{p}_{t|h}\big(i\big).$$

- si può cercare la sequenza di stati che massimizzano ad ogni istante la probabilità che la catena si trovi in essi sulla base delle osservazioni (calcolo della probabilità 'backward'), ovvero

$$\widehat{x}^t_{t|h} = \big(\widehat{x}_{0|h}, \cdots, \widehat{x}_{t|h}\big) \in \underbrace{V \times \cdots \times V}_{\text{t+1 volte}}$$

- si può cercare la traiettoria dello stato della catena con più alta probabilità sulla base delle osservazioni Y^h (problema della decodifica), ovvero

$$\widehat{\widehat{x}}^t_{t|h} = \arg\max_{x_0,\cdots,x_t} \; p\big(x^t \mid Y^h\big) \in \underbrace{V \times \cdots \times V}_{\text{t+1 volte}}$$

  Da notare che in generale $\widehat{\widehat{x}}^t_{t|h} \neq \widehat{x}^t_{t|h}$.

Sulla base dell'analogia con il caso continuo per cui vale

$$p\big(x_t \mid Y^t\big) \propto p(y_t \mid x_t) \int_{x_{t-1}} p\big(x_t \mid x_{t-1}\big) p\big(x_{t-1} \mid Y^{t-1}\big) \, d\, x_{t-1}$$

si può scrivere la seguente relazione che definisce esplicitamente la probabilità che la catena si trovi nello stato $x_t = i$ sulla base delle osservazioni Y^t:

$$\underbrace{p\big(x_t = i \mid Y^t\big)}_{\widehat{p}_{t|t}(i)} \propto \big(y_t = \bar{y}_t \mid x_t = i\big) \left(\sum_{j=1}^N p\big(x_t = i \mid x_{t-1} = j\big) \underbrace{p\big(x_{t-1} = j \mid Y^{t-1}\big)}_{\widehat{p}_{t-1|t-1}(j)} \right)$$

da cui

$$\underbrace{\widehat{p}_{t|t}(i) \propto c_i(\bar{y}_t) \underbrace{\sum_{j=1}^N a_{ij} \, \widehat{p}_{t-1|t-1}(j)}_{\text{predizione}}}_{\text{aggiornamento}}$$

dove

$$c_i(\bar{y}_t) = p\big(y_t = \bar{y}_t \mid x_t = i\big)$$

e

$$C(\bar{y}_t) = \big[\; c_1(\bar{y}_t) \quad \cdots \quad c_N(\bar{y}_t) \; \big]^T$$

è la colonna della matrice C relativa all'uscita $\bar{y}_t$, $\bar{y}_t \in W$. Scrivendo in forma compatta otteniamo:

- la predizione

$$\hat{p}^T_{t|t-1} = \hat{p}^T_{t-1|t-1} A$$

- la predizione a più passi

$$\hat{p}^T_{t|h} = \hat{p}^T_{h|h} A^{t-h}$$

- e il filtraggio

$$\widehat{p}_{t|t} = \frac{\widehat{p}_{t|t-1} * c(\bar{y}_t)}{||\widehat{p}_{t|t-1} * c(\bar{y}_t)||_1} \quad \equiv \quad p^T_{t|t} = \frac{\text{diag}[c(\bar{y}_t)] p_{t|t-1}}{||\text{diag}[c(\bar{y}_t)] p_{t|t-1}||_1}$$

dove $.*$ indica il prodotto elemento per elemento di due vettori colonna (simbologia Matlab).

Consideriamo ora il caso più corrispondente allo smoothing, nel quale vengono considerate tutte le misure (anche future) per la predizione dello stato, cioè $p(x_t|Y^T)$. In particolare l'applicazione delle usiali regole sull'indipendenza condizionata e di Bayes si ottiene:

$$\begin{aligned} p\big(x_t = i \mid Y^T\big) &= \frac{p\big(Y^t, x_t = i\big) p\big(\widetilde{Y}^{t+1} \mid x_t = i\big)}{p\big(Y^T\big)} \\ &= \frac{\alpha_{t|t}\big(i\big) \beta_{t|t+1}\big(i\big)}{p\big(Y^T\big)} \\ &= \frac{\alpha_{t|t}\big(i\big) \beta_{t|t+1}\big(i\big)}{\sum_{i=1}^N \alpha_{t|t}\big(i\big) \beta_{t|t+1}\big(i\big)} \end{aligned}$$

Il termine relativo ad $\alpha_{t|t}$ si ottiene facilmente in maniera simile a $\hat{p}_{t|t}$ visto in precedenza e che si puo' riassumere nelle sequenti equazioni:

- $\alpha^T_{t+1|t} = \alpha^T_{t|t} A$
- $\alpha_{t+1|t+1} = \alpha_{t+1|t} \; .* \, c\big(\bar{y}_{t+1}\big)$
- $\alpha_{0|-1} = \pi_0$

Da un punto di vista teorico tali espressioni sono soddisfacenti, però da un punto di vista computazionale tale calcolo potrebbe risultare troppo oneroso. Non è necessario normalizzare ad ogni passo, meglio se lo si fà ogni tanto: in letteratura esistono tecniche che operano proprio in questo senso. Si veda per esempio L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE, vol 77(2), pp. 257-286

Analizzando invece il secondo termine relativo a $\beta_{t|t+1}$ otteniamo:

$$\begin{aligned}\beta_{t|t+1}(i) &= p(\widetilde{Y}^{t+1} \mid x_t = i) = \sum_{j=1}^{N} p(\widetilde{Y}^{t+1}, x_{t+1} = j \mid x_t = i) \\ &= \sum_{j=1}^{N} p(\widetilde{Y}^{t+1} \mid x_{t+1} = j, x_t = i) p(x_{t+1} = j \mid x_t = i) \\ &= \sum_{j=1}^{N} p(\widetilde{Y}^{t+1} \mid x_{t+1} = j) p(x_{t+1} = j \mid x_t = i) \\ &= \beta_{t+1|t+1}(j)\ a_{ij}\end{aligned}$$

nella quale (secondo passaggio), prima utilizziamo la formula di Bayes, e poi (penultimo passaggio) sfruttiamo il fatto che le misure future condizionando rispetto a x_{t+1} e x_t sono indipendenti da x_t. Alla fine si ottiene una espressione moltiplicativa tra due termini, dove il primo per definizione è proprio $\beta_{t+1|t+1}(j)$, mentre il secondo a_{ij}. Questo significa che l'elemento j-esimo del vettore $\beta_{t|t+1}$ si ottiene prendendo la riga i di a moltiplicato per il vettore $\beta_{t+1|t+1}$.
Scrivendo in forma compatta otteniamo

$$\beta_{t|t+1} = A\beta_{t+1|t+1}$$

In pratica per mantenera la stessa analogia del vettore $\alpha_{t|t}$ è opportuno prendere il trasposto di tutta l'espessione:

$$\beta^T_{t|t+1} = \beta^T_{t+1|t+1} A^T$$

in cui si nota che nelle catene di Markov l'inversione della dinamica comporta a prendere la trasposta della matrice A.
Invece, per quel che riguarda l'aggiornamento

$$\beta_{t+1|t+1}(i) = p(\widetilde{Y}^{t+1} \mid x_{t+1} = i) = \underbrace{p(y_{t+1} \mid x_{t+1} = i)}_{c_i(\bar{y}_{t+1})} \underbrace{p(\widetilde{Y}^{t+2} \mid x_{t+1} = i)}_{\beta_{t+1|t+2}(i)}$$

dove si nota che la misura t-esima e le successive sono indipendenti tra di loro una volta che ho condizionato rispetto a x_{t+1}. Trascrivendo sempre in forma compatta otteniamo l'espressione

$$\beta_{t|t} = c(\bar{y}_t)\ .*\ \beta_{t|t+1}$$

mentre l'inizializzazione (che in questo caso invece è la condizione finale) è data da

$$\beta_{T|T} = \underbrace{p(y_T \mid x_T = i)}_{c_i(\bar{y}_T)}$$

Quindi, riassumendo l'algoritmo di forward-backward calcola lo stato con maggiore probabilità in un certo istante date tutte le misure fino all'istante T, cioè la sequenza $\{\widehat{x}_{0|T}, \ldots, \widehat{x}_{T|T}\}$ per ogni $t = 0, \ldots, T$. Tale sequenza è generalmente diversa dalla seguente, data da:

$$\{\widehat{\widehat{x}}_{0|T}, \ldots, \widehat{\widehat{x}}_{T|T}\} = \arg\max_{x_0,\cdots,x_T} \; p\big(x_0, x_1 \ldots, x_t \mid Y^T\big)$$

Quest'ultima intatti è la probabilità di una particolare sequenza: tra tutte le possibili prendo quella con probabilità maggiore. Si verifica velocemente che sono vere le due seguenti relazioni:

$$p\big(\widehat{\widehat{x}}_{0|T}, \ldots, \widehat{\widehat{x}}_{T|T} \mid Y^T\big) \geq p\big(\widehat{x}_{0|T}, \ldots, \widehat{x}_{T|T} \mid Y^T\big)$$

$$p\big(\widehat{\widehat{x}}_{t|T} \mid Y^T\big) \leq p\big(\widehat{x}_{t|T} \mid Y^T\big)$$

Per calcolare la sequenza $\{\widehat{\widehat{x}}_{0|T}, \ldots, \widehat{\widehat{x}}_{T|T}\}$ si utilizza un algoritmo particolare della programmazione dinamica: l'algoritmo di Viterbi, che verrà analizzato nella lezione seguente.