Probabilità e Processi Stocastici (455AA)

Lezione 6

Dario Trevisan – *https://web.dm.unipi.it/trevisan*

8/10/2025

Richiami

- Variabili aleatorie $X = \{X = x\}_{x \in E}$

Richiami

- Variabili aleatorie $X = \{X = x\}_{x \in E}$
- Densità discreta/continua

Richiami

- Variabili aleatorie $X = \{X = x\}_{x \in E}$
- Densità discreta/continua
- Variabili composte $g(X)$ e transformazione delle densità

Richiami

- Variabili aleatorie $X = \{X = x\}_{x \in E}$
- Densità discreta/continua
- Variabili composte $g(X)$ e transformazione delle densità
- Variabili congiunte e marginali

Richiami

- Variabili aleatorie $X = \{X = x\}_{x \in E}$
- Densità discreta/continua
- Variabili composte $g(X)$ e transformazione delle densità
- Variabili congiunte e marginali
- Formula di Bayes

$$p(Y = y|X = x) \propto p(Y = y)L(Y = y; X = x)$$

Indipendenza

Indipendenza: caso discreto

L'indipendenza tra sistemi di alternative si traduce per variabili discrete:

- Siano $X_1 \in E_1, \ldots X_k \in E_k$ variabili aleatorie con densità discreta (rispetto ad una informazione nota I). Allora esse si dicono indipendenti (condizionatamente ad I) se vale

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k | I) = \prod_{i=1}^{k} P(X_i = x_i | I),$$

per ogni $x_1 \in E_1$, $x_2 \in E_2$, $\ldots$, $x_k \in E_k$, o equivalentemente, per ogni sottoinsieme $J \subseteq \{1, \ldots, k\}$,

$$\begin{aligned} &P(X_j = x_j \text{ per ogni } j \in J | I, X_\ell = x_\ell \text{ per ogni } \ell \notin J) \\ &= P(X_j = x_j \text{ per ogni } j \in J | I). \end{aligned}$$

Indipendenza: caso discreto

L'indipendenza tra sistemi di alternative si traduce per variabili discrete:

- Siano $X_1 \in E_1, \ldots X_k \in E_k$ variabili aleatorie con densità discreta (rispetto ad una informazione nota I). Allora esse si dicono indipendenti (condizionatamente ad I) se vale

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k | I) = \prod_{i=1}^{k} P(X_i = x_i | I),$$

per ogni $x_1 \in E_1$, $x_2 \in E_2$, $\ldots$, $x_k \in E_k$, o equivalentemente, per ogni sottoinsieme $J \subseteq \{1, \ldots, k\}$,

$$\begin{aligned} &P(X_j = x_j \text{ per ogni } j \in J | I, X_\ell = x_\ell \text{ per ogni } \ell \notin J) \\ &= P(X_j = x_j \text{ per ogni } j \in J | I). \end{aligned}$$

- Il membro a sinistra è la densità discreta della variabile congiunta $(X_1, \ldots, X_k)$, mentre a destra abbiamo il prodotto delle densità discrete delle marginali.

Indipendenza: caso continuo

- Siano $X_1 \in \mathbb{R}^{d_1}, \ldots X_k \in \mathbb{R}^{d_k}$ variabili aleatorie con densità continua (rispetto ad una informazione nota I). Allora esse si dicono indipendenti (condizionatamente ad I) se la variabile congiunta $X = (X_1, \ldots, X_k)$ ammette densità continua e vale

$$p(X = x|I) = p(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k|I) = \prod_{i=1}^{k} p(X_i = x_i|I),$$

per ogni $x_1 \in \mathbb{R}^{d_1}$, $x_2 \in \mathbb{R}^{d_2}$, $\ldots$, $x_k \in \mathbb{R}^{d_k}$, o equivalentemente, per ogni sottoinsieme $J \subseteq \{1, \ldots, k\}$,

$$\begin{aligned} &p(X_j = x_j \text{ per ogni } j \in J|I, X_\ell = x_\ell \text{ per ogni } \ell \notin J) \\ &= p(X_j = x_j \text{ per ogni } j \in J|I). \end{aligned}$$

Indipendenza: caso generale

- Possiamo immaginare definizioni valide anche per i casi “misti”, in cui alcune variabili sono discrete e altre continue. Ma è possibile dare una definizione generale (che include quelle sopra).

Indipendenza: caso generale

- Possiamo immaginare definizioni valide anche per i casi "misti", in cui alcune variabili sono discrete e altre continue. Ma è possibile dare una definizione generale (che include quelle sopra).
- Siano $X_1 \in E_1, \ldots X_k \in E_k$ variabili aleatorie (generali). Allora esse si dicono indipendenti (condizionatamente ad una informazione nota I) se vale

$$P(X_1 \in U_1, X_2 \in U_2, \ldots, X_k \in U_k | I) = \prod_{i=1}^{k} P(X_i \in U_i | I),$$

per ogni $U_1 \subseteq E_1$, $U_2 \subseteq E_2$, $\ldots$ $U_k \subseteq E_k$, o equivalentemente, per ogni sottoinsieme $J \subseteq \{1, \ldots, k\}$,

$$\begin{aligned} &P(X_j \in U_j \text{ per ogni } j \in J | I, X_\ell \in U_\ell \text{ per ogni } \ell \notin J) \\ &= P(X_j \in U_j \text{ per ogni } j \in J | I). \end{aligned}$$

Indipendenza e composizione

Vale il seguente risultato:

- Siano $X_1 \in E_1, \ldots X_k \in E_k$ variabili aleatorie (generali). Allora esse sono indipendenti (condizionatamente ad una informazione nota I) se e solo se, dato un qualsiasi sottoinsieme $J \subseteq \{1, \ldots, k\}$, qualsiasi affermazione A associata alle variabili $\{X_j\}_{j \in J}$ è indipendente (sapendo I) da qualsiasi affermazione B associata alle rimanenti variabili $\{X_\ell\}_{\ell \in \{1,\ldots,k\} \setminus J}$.

Indipendenza e composizione

Vale il seguente risultato:

- Siano $X_1 \in E_1, \ldots X_k \in E_k$ variabili aleatorie (generali). Allora esse sono indipendenti (condizionatamente ad una informazione nota I) se e solo se, dato un qualsiasi sottoinsieme $J \subseteq \{1, \ldots, k\}$, qualsiasi affermazione A associata alle variabili $\{X_j\}_{j \in J}$ è indipendente (sapendo I) da qualsiasi affermazione B associata alle rimanenti variabili $\{X_\ell\}_{\ell \in \{1,\ldots,k\} \setminus J}$.
- Una **conseguenza fondamentale**: ogni variabile ottenuta tramite funzione delle sole $(X_j)_{j \in J}$, è indipendente da ogni variabile ottenuta tramite funzione delle sole $(X_\ell)_{\ell \notin J}$.

Reti bayesiane

Reti Bayesiane

Presentiamo un metodo grafico per rappresentare le densità di variabili aleatorie (una specie di estensione dei diagrammi ad albero per variabili invece di eventi)

- Passo zero: si fissa un ordinamento tra le variabili (che corrisponde all'ordine in cui i sistemi di alternative vengono aggiunti nella costruzione del grafo ad albero): $X_1, X_2, \ldots X_k$

Il diagramma è un grafo orientato su k nodi corrispondenti alle k variabili, viene costruito in k passi:

- nel primo passo si introduce solamente il nodo corrispondente alla variabile X_1;

Il diagramma è un grafo orientato su k nodi corrispondenti alle k variabili, viene costruito in k passi:

- nel primo passo si introduce solamente il nodo corrispondente alla variabile X_1;
- nel passo i-esimo, si introduce il nodo corrispondente alla variabile X_i, e si considera la densità di X_i condizionata a tutte le variabili inserite $X_1, \ldots X_{i-1}$,

$$P(X_i = x_i | I, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1).$$

Il diagramma è un grafo orientato su k nodi corrispondenti alle k variabili, viene costruito in k passi:

- nel primo passo si introduce solamente il nodo corrispondente alla variabile X_1;
- nel passo i-esimo, si introduce il nodo corrispondente alla variabile X_i, e si considera la densità di X_i condizionata a tutte le variabili inserite $X_1, \ldots X_{i-1}$,

$$P(X_i = x_i | I, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1).$$

- si individua un sottoinsieme $J \subseteq \{1, \ldots, i-1\}$ tale che la densità dipenda solo dalle $(X_j)_{j \in J}$, ossia, per ogni $x_1, \ldots, x_i$, valga

$$\begin{aligned} &P(X_i = x_i | I, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \\ &\quad = P(X_i = x_i | I, X_j = x_j \text{ per ogni } j \in J). \end{aligned}$$

Il diagramma è un grafo orientato su k nodi corrispondenti alle k variabili, viene costruito in k passi:

- nel primo passo si introduce solamente il nodo corrispondente alla variabile X_1;
- nel passo i-esimo, si introduce il nodo corrispondente alla variabile X_i, e si considera la densità di X_i condizionata a tutte le variabili inserite $X_1, \ldots X_{i-1}$,

$$P(X_i = x_i | I, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1).$$

- si individua un sottoinsieme $J \subseteq \{1, \ldots, i-1\}$ tale che la densità dipenda solo dalle $(X_j)_{j \in J}$, ossia, per ogni $x_1, \ldots, x_i$, valga

$$\begin{aligned} &P(X_i = x_i | I, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \\ &\quad = P(X_i = x_i | I, X_j = x_j \text{ per ogni } j \in J). \end{aligned}$$

- si inseriscono gli archi orientati (frecce) da ciascun nodo corrispondente alle variabili X_j, $j \in J$, verso il nodo corrispondente ad X_i.

Date k variabili aleatorie indipendenti $X_1, \ldots X_k$, l'algoritmo produce il diagramma:

- Si consideri una variabile aleatoria Λ tale che, condizionatamente ad essa, le variabili $T_1, \ldots, T_k$ sono indipendenti (un esempio concreto è $\Lambda = \lambda$ individua il parametro delle variabili T_i che hanno legge esponenziale).

- Si consideri una variabile aleatoria Λ tale che, condizionatamente ad essa, le variabili $T_1, \ldots, T_k$ sono indipendenti (un esempio concreto è $\Lambda = \lambda$ individua il parametro delle variabili T_i che hanno legge esponenziale).
- La densità congiunta ha la forma

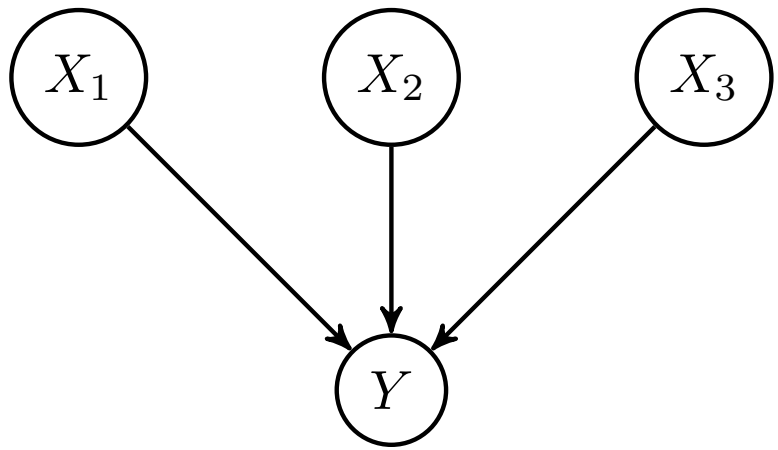
$$P(\Lambda, T_1, T_2, T_3, T_4) = P(\Lambda)P(T_1|\Lambda)P(T_2|\Lambda)P(T_3|\Lambda)P(T_4|\Lambda).$$

- Si consideri una variabile aleatoria Λ tale che, condizionatamente ad essa, le variabili $T_1, \ldots, T_k$ sono indipendenti (un esempio concreto è $\Lambda = \lambda$ individua il parametro delle variabili T_i che hanno legge esponenziale).
- La densità congiunta ha la forma

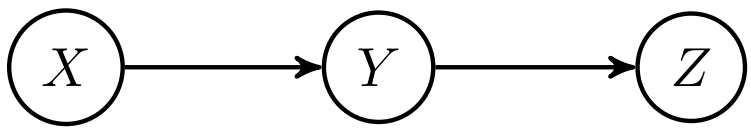
$$P(\Lambda, T_1, T_2, T_3, T_4) = P(\Lambda)P(T_1|\Lambda)P(T_2|\Lambda)P(T_3|\Lambda)P(T_4|\Lambda).$$

- La rete bayesiana costruita inserendo prima la variabile Λ e poi le rimanenti è la seguente:

Si considerino k variabili $X_1, \ldots, X_k$ indipendenti tra loro (rispetto all'informazione iniziale) e sia $Y = g(X_1, \ldots, X_k)$ (ad esempio $Y = X_1 + \ldots + X_k$ nel caso di variabili a valori in $\mathbb{R}$). La rete bayesiana è rappresentata in figura.



Si consideri la rete bayesiana rappresentata in figura. Condizionando rispetto ad Y, si ottiene che X e Z sono indipendenti. Questo è un semplice esempio di *catena di Markov*.



Cenni ai metodi numerici

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)
  b. Si stabiliscono probabilità *a priori* per Θ (ad es. uniformi) e *verosimiglianza* $L(\Theta = \theta; X = x) = P(X = x | \Theta = \theta)$

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)
  b. Si stabiliscono probabilità *a priori* per Θ (ad es. uniformi) e *verosimiglianza* $L(\Theta = \theta; X = x) = P(X = x | \Theta = \theta)$
  c. Si aggiorna il modello incorporando i dati *osservati* (Bayes)

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)
  b. Si stabiliscono probabilità *a priori* per Θ (ad es. uniformi) e *verosimiglianza* $L(\Theta = \theta; X = x) = P(X = x | \Theta = \theta)$
  c. Si aggiorna il modello incorporando i dati *osservati* (Bayes)
  d. Si stimano i parametri Θ (ad es. mediante MAP o MLE)

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)
  b. Si stabiliscono probabilità *a priori* per Θ (ad es. uniformi) e *verosimiglianza* $L(\Theta = \theta; X = x) = P(X = x | \Theta = \theta)$
  c. Si aggiorna il modello incorporando i dati *osservati* (Bayes)
  d. Si stimano i parametri Θ (ad es. mediante MAP o MLE)
  e. Si fanno *previsioni* su eventuali variabili non osservate usando

Dalla teoria alla pratica

- In teoria abbiamo tutti gli strumenti per affrontare problemi concreti:
  a. Si definisce un *modello* con parametri Θ e quantità osservabili X (una opportuna rete bayesiana)
  b. Si stabiliscono probabilità *a priori* per Θ (ad es. uniformi) e *verosimiglianza* $L(\Theta = \theta; X = x) = P(X = x | \Theta = \theta)$
  c. Si aggiorna il modello incorporando i dati *osservati* (Bayes)
  d. Si stimano i parametri Θ (ad es. mediante MAP o MLE)
  e. Si fanno *previsioni* su eventuali variabili non osservate usando
- Il problema è che nella pratica la numerosità sia dei dati osservati sia dei parametri è elevata e questo si richiede di calcolare densità di probabilità su spazi di *dimensione elevata*, che diventa impraticabile analiticamente.

Densità coniugate

Per certe funzioni di verosimiglianza, si possono introdurre delle densità a priori (*coniugate*) particolarmente convenienti perché la densità a posteriori è nella stessa classe.

- In n esperimenti indipendenti, ciascuno con probabilità di successo $Y = y$, il numero X di successi ha una densità binomiale

$$L(Y = y; X = k) = P(X = k|Y = y) = \binom{n}{k} y^k (1 - y)^{n-k}.$$

Densità coniugate

Per certe funzioni di verosimiglianza, si possono introdurre delle densità a priori (*coniugate*) particolarmente convenienti perché la densità a posteriori è nella stessa classe.

- In n esperimenti indipendenti, ciascuno con probabilità di successo $Y = y$, il numero X di successi ha una densità binomiale

$$L(Y = y; X = k) = P(X = k|Y = y) = \binom{n}{k} y^k (1 - y)^{n-k}.$$

- Se la densità di Y a priori è *Beta* di parametri α, $\beta > 0$, ossia

$$p(Y = y) \propto y^{\alpha-1}(1 - y)^{\beta-1},$$

allora la densità a posteriori avendo osservato $X = k$ successi è ancora Beta:

$$p(Y = y|X = k) \propto y^{\alpha+k-1}(1 - y)^{\beta+n-k-1}.$$

Problema

Una routine probabilistica fornisce l'output desiderato con probabilità $p \in [0, 1]$, inizialmente non nota. Ogni applicazione dell'algoritmo produce esiti tra di loro indipendenti. Si modellizza inizialmente p come una variabile uniforme su $[0, 1]$.

- Dopo aver effettuato 100 esperimenti, avendo osservato che l'output è corretto su 60 di questi, scrivere la densità a posteriori per p e riconoscere una densità beta. Determinare la stima di massimo a posteriori.

Problema

Una routine probabilistica fornisce l'output desiderato con probabilità $p \in [0, 1]$, inizialmente non nota. Ogni applicazione dell'algoritmo produce esiti tra di loro indipendenti. Si modellizza inizialmente p come una variabile uniforme su $[0, 1]$.

- Dopo aver effettuato 100 esperimenti, avendo osservato che l'output è corretto su 60 di questi, scrivere la densità a posteriori per p e riconoscere una densità beta. Determinare la stima di massimo a posteriori.
- Come cambia la risposta se invece si effettuano 1000 esperimenti e si osserva un output corretto su 600 di questi?

Soluzione

Tecniche numeriche

Le densità coniugate non risolvono completamente il problema, in particolare se la numerosità delle osservazioni è molto grande.

Si ricorre a metodi numerici. Due approcci fondamentali:

- Approssimare tutta la densità a posteriori di Y

Tecniche numeriche

Le densità coniugate non risolvono completamente il problema, in particolare se la numerosità delle osservazioni è molto grande.

Si ricorre a metodi numerici. Due approcci fondamentali:

- Approssimare tutta la densità a posteriori di Y
- Determinare solamente la stima di massima verosimiglianza y_{MAP}

Integrazione numerica

Nel primo caso il problema è quindi di approssimare numericamente un integrale

$$P(Y \in U|I) = \int_U p(Y = y|I)dy.$$

- Negli esempi seguiamo un approccio elementare: calcolare $p(Y = y|I)$ su una griglia

Integrazione numerica

Nel primo caso il problema è quindi di approssimare numericamente un integrale

$$P(Y \in U|I) = \int_U p(Y = y|I)dy.$$

- Negli esempi seguiamo un approccio elementare: calcolare $p(Y = y|I)$ su una griglia
- Problema: se Y è a valori in $[0, 1]^d \subseteq \mathbb{R}^d$ e il passo è δy, si devono calcolare circa

$$\frac{1}{\delta^d}$$

valori (e d è molto grande).

Integrazione numerica

Nel primo caso il problema è quindi di approssimare numericamente un integrale

$$P(Y \in U|I) = \int_U p(Y = y|I)dy.$$

- Negli esempi seguiamo un approccio elementare: calcolare $p(Y = y|I)$ su una griglia
- Problema: se Y è a valori in $[0, 1]^d \subseteq \mathbb{R}^d$ e il passo è δy, si devono calcolare circa

$$\frac{1}{\delta^d}$$

  valori (e d è molto grande).
- Approcci alternativi cercano di ottimizzare la scelta dei punti, anche mediante scelte (pseudo)-*casuali* (metodi Monte Carlo).

Metodi di ottimizzazione

Nel secondo caso il problema è di determinare il punto di massimo di una funzione

$$y_{\mathrm{MLE}} = \arg\max \{L(Y = y; X = x)\}.$$

- Il problema generale è studiatissimo e vi sono molteplici algoritmi (Newton, ascesa gradiente...)

Metodi di ottimizzazione

Nel secondo caso il problema è di determinare il punto di massimo di una funzione

$$y_{\text{MLE}} = \arg\max \{L(Y = y; X = x)\}.$$

- Il problema generale è studiatissimo e vi sono molteplici algoritmi (Newton, ascesa gradiente...)
- In R le funzioni `nlm()` e `optim()` permettono di applicare i principali.

Metodi di ottimizzazione

Nel secondo caso il problema è di determinare il punto di massimo di una funzione

$$y_{\text{MLE}} = \arg\max \{L(Y = y; X = x)\}.$$

- Il problema generale è studiatissimo e vi sono molteplici algoritmi (Newton, ascesa gradiente...)
- In R le funzioni `nlm()` e `optim()` permettono di applicare i principali.
- Se la numerosità delle osservazioni X è troppo elevata, si ricorre a metodi probabilistici.

Metodi di ottimizzazione

Nel secondo caso il problema è di determinare il punto di massimo di una funzione

$$y_{\mathrm{MLE}} = \arg\max \{L(Y = y; X = x)\}.$$

- Il problema generale è studiatissimo e vi sono molteplici algoritmi (Newton, ascesa gradiente...)
- In R le funzioni `nlm()` e `optim()` permettono di applicare i principali.
- Se la numerosità delle osservazioni X è troppo elevata, si ricorre a metodi probabilistici.
- A volte $Y = (Y_{\mathrm{par}}, Y_{\mathrm{hid}})$ e si richiede di massimizzare solo la verosimiglianza di Y_{par}. Necessità di ricorrere a metodi "misti" (ad esempio Expectation-Maximization).

Problema

Il numero di clienti che entrano in un negozio in un dato giorno è rappresentato da una variabile con densità Poisson di parametro $\Lambda > 0$ (inizialmente non noto). Si vuole stimare λ osservando gli ingressi in tre giorni consecutivi. Si suppone che a giorni diversi corrispondano numeri di ingressi indipendenti tra loro.

- Supponendo di avere osservato in sequenza $X_1 = 4$, $X_2 = 5$ e 3 ingressi nei tre giorni, fornire una stima di massima verosimiglianza per λ.

Problema

Il numero di clienti che entrano in un negozio in un dato giorno è rappresentato da una variabile con densità Poisson di parametro $\Lambda > 0$ (inizialmente non noto). Si vuole stimare λ osservando gli ingressi in tre giorni consecutivi. Si suppone che a giorni diversi corrispondano numeri di ingressi indipendenti tra loro.

- Supponendo di avere osservato in sequenza $X_1 = 4$, $X_2 = 5$ e 3 ingressi nei tre giorni, fornire una stima di massima verosimiglianza per λ.
- Supponendo che Λ sia a priori distribuito come una variabile esponenziale di parametro 1, avendo osservato la stessa sequenza 4, 5, 3 di ingressi, fornire la densità a posteriori e la stima di massimo a posteriori.

Problema

Il numero di clienti che entrano in un negozio in un dato giorno è rappresentato da una variabile con densità Poisson di parametro $\Lambda > 0$ (inizialmente non noto). Si vuole stimare λ osservando gli ingressi in tre giorni consecutivi. Si suppone che a giorni diversi corrispondano numeri di ingressi indipendenti tra loro.

- Supponendo di avere osservato in sequenza $X_1 = 4$, $X_2 = 5$ e 3 ingressi nei tre giorni, fornire una stima di massima verosimiglianza per λ.
- Supponendo che Λ sia a priori distribuito come una variabile esponenziale di parametro 1, avendo osservato la stessa sequenza 4, 5, 3 di ingressi, fornire la densità a posteriori e la stima di massimo a posteriori.
- Come cambiano le risposte se invece si osservano in sequenza $X_1 \leq 4$, $X_2 \leq 5$ e $X_3 \leq 3$ ingressi?

Soluzione