Teoria dei fenomeni aleatori e della stima – 24/07/2019

- Ogni esercizio viene valutato da 0 a 6 punti.
- Verranno valutate solo le parti scritte in penna – non usare correttori.
- Non riportare solo il risultato, ma cerca di argomentare sinteticamente la risposta.
- Riportare il proprio nome, cognome, e numero di matricola su ogni foglio consegnato.
- Esercizio da escludere dal punteggio finale:

1. Si consideri la parola *piccola*. Pescando un anagramma della parola a caso, qual è la probabilità che la parola pescata non abbia le due lettere c in posizione adiacente? *Un anagramma è una parola ottenuta permutando le lettere della parola di partenza.*

2. Due v.a. continue X e Y hanno legge congiunta $f_{X,Y}(x,y) = ke^{-x-y}$ per $x > 0$ e $y > 0$, e $f_{X,Y}(x,y) = 0$ altrove. Determinare:

   (a) il valore della costante k.

   (b) le marginali di X e Y, e dire se le v.a. sono indipendenti. Giustificare la risposta.

   (c) il valore atteso $\mathsf{E}[XY]$.

3. Sia $X \sim \mathcal{U}[-1,1]$ e $U = \sqrt{|X|}$. Trovare la legge di probabilità di U.

4. Un trasmettitore è dotato di $n = 50$ antenne, ognuna delle quali trasmette un segnale X_i con $\mathsf{E}[X_i] = 0$, $\mathsf{Var}[X_i] = 1$, e $\mathsf{Var}[X_i^2] = 0.5$. Tutti i segnali sono indipendenti e identicamente distribuiti. Calcolare un'approssimazione alla probabilità che la potenza istantanea trasmessa per antenna $S = \frac{1}{50}\sum_{i=1}^{50} X_i^2$ superi il valore 1:

   (a) usando la diseguaglianza di Markov

   (b) usando il teorema centrale del limite

   Dire quale delle due approssimazioni è più significativa e perché.

5. Si vuole ottenere una stima numerica della costante di Euler-Mascheroni $\gamma = \int_0^\infty -\log(x)e^{-x}dx$. Usando n campioni indipendenti $U_i \sim \mathcal{U}[0,1]$ per $i = 1,\ldots,n$ proporre un algoritmo di Importance Sampling per la stima di γ.
   Suggerimento: parte della funzione integranda è la legge di probabilità di una v.a. nota.

6. Una sorgente di dati emette i simboli $X \in \{a,b,c,d\}$ rispettivamente con probabilità $\{1/2, 1/4, 1/8, 1/8\}$ in maniera indipendente da simbolo a simbolo. Si consideri un messaggio di $n = 1000$ simboli emesso dalla sorgente.

   (a) Qual è la lunghezza del messaggio non compresso (in bit)?

   (b) Mediamente qual è la lunghezza minima della stringa (in bit) che possiamo usare per codificare questo messaggio senza perdita di informazione?

   (c) Proporre un codice ottimale per la compressione del messaggio emesso dalla sorgente.

Soluzioni

Problema 1

Siccome lo spazio di probabilità è uniforme (tutte gli anagrammi sono equiprobabili) calcoliamo i casi favorevoli e i casi totali. Gli anagrammi totali sono 7!, dove abbiamo considerato distinte anche le parole che scambiano di posto le 2 lettere c. Ora contiamo tutti gli anagrammi che contengono le lettere cc in posizioni adiacenti: per far ciò è sufficiente considerare la coppia cc come un unico simbolo, e quindi il numero totale di anagrammi è $2 \cdot 6!$, dove il 2 conta anche lo scambio tra le lettere c. In definitiva, la probabilità cercata è:

$$p = 1 - \frac{2 \cdot 6!}{7!} = \frac{5}{7}.$$

Problema 2

1. Il valore di k si ottiene imponendo l'integrale della ddp a 1:

$$\int_0^\infty \int_0^\infty e^{-x-y} dxdy = \left(\int_0^\infty e^{-x} dx \right)^2 = 1,$$

dunque si ha $k = 1$.

2. Per simmetria, le marginali di X e Y avranno la stessa forma:

$$f_Y(x) = f_X(x) = \int_0^\infty e^{-x-y} dy = e^{-x}.$$

Siccome il prodotto delle marginali è uguale alla legge congiunta:

$$f_{X,Y}(x, y) = e^{-x-y} = f_X(x) f_Y(y), \qquad \forall (x, y)$$

possiamo concludere che le v.a. X e Y sono indipendenti.

3. Il valore atteso da calcolare è

$$\mathsf{E}[XY] = \int_0^\infty \int_0^\infty xy e^{-x-y} dxdy = \left(\int_0^\infty x e^{-x} dx \right)^2 = (\mathsf{E}[X])^2 = 1,$$

dove nell'ultimo step abbiamo usato il fatto che $X \sim \text{Exp}(1)$.

Problema 3

Definiamo una v.a. di comodo $Y = |X|$. Siccome tutti i valori negativi di X vengono mappati nei valori positivi con lo stesso valore assoluto, si ha semplicemente $Y \sim \mathcal{U}[0, 1]$. Ora si calcola la legge di $U = \sqrt{Y}$ tramite il metodo della cumulata:

$$F_U(u) = \Pr(U \le u) = \Pr(\sqrt{Y} \le u) = \Pr(Y \le u^2) = u^2, \qquad 0 \le u \le 1,$$

dove nell'ultimo passaggio abbiamo usato $\Pr(Y \le y) = y$ per $0 \le y \le 1$. Differenziando rispetto a u si ottiene $f_U(u) = 2u$ per $0 \le u \le 1$ e zero altrove.

Problema 4

1. Siccome S è una v.a. positiva si può applicare la diseguaglianza di Markov:

$$\Pr(S > 1) \le \frac{\mathsf{E}[S]}{1} = \mathsf{E}[X_1^2] = \mathsf{Var}[X_1] = 1$$

dove abbiamo usato $\mathsf{E}[X_1] = 0$.

2. Siccome tutte le X_i sono i.i.d. si può applicare il teorema centrale del limite:

$$\begin{aligned}
\Pr(S > 1) &= \Pr\left(\sum_{i=1}^{50} X_i^2 > 50 \right) \\
&= \Pr\left(\sum_{i=1}^{50} X_i^2 - 50 > 0 \right) \\
&= \Pr\left(\frac{\sum_{i=1}^{50} X_i^2 - 50}{\sqrt{\mathsf{Var}[\sum_i^{50} X_i^2]}} > 0 \right) \approx \Pr(Z > 0) = 1/2.
\end{aligned}$$

L'approssimazione suggerita dalla diseguaglianza di Markov non è utile, dato che si limita a dire che la probabilità non può superare 1. L'approssimazione suggerita dal CLT è più realistica, perché mediare 50 contributi indipendenti potrebbe già portare vicino ad una distribuzione Gaussiana.

Problema 5

L'integrale si può reinterpretare come segue:

$$\gamma = \int_0^\infty -\log(x)e^{-x}dx = \mathsf{E}\left[-\log(X)\right]$$

dove $X \sim \text{Exp}(1)$. Un possibile algoritmo per generare i campioni esponenziali e per stimare l'integrale è il seguente:

1. Genero n campioni indipendenti $U_i \sim \mathcal{U}[0,1]$ per $i = 1, \ldots, n$.
2. Genero i campioni esponenziali come $X_i = -\log(U_i)$ per $i = 1, \ldots, n$.
3. Stimo numericamente l'integrale come $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^n -\log(X_i) = \frac{1}{n}\sum_{i=1}^n \log(\log(U_i))$.

Problema 6

1. Se si rinuncia a comprimere il messaggio, per ogni simbolo bisogna usare 2 bit, quindi la lunghezza totale del messaggio è $2n = 2000$ bit.
2. Mediamente la lunghezza minima della stringa è

$$n \cdot H(X) = n\left(\frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{2}{8}\log_2 8\right) = n\left(\frac{1}{2} + \frac{1}{2} + \frac{3}{4}\right) = n\frac{7}{4} = 1750 \text{ bit}$$

3. Le informazioni dei singoli simboli della sorgente sono

$$i(a) = 1 \text{ bit}, \qquad i(b) = 2 \text{ bit}, \qquad i(c) = i(d) = 3 \text{ bit}.$$

Un codice C che assegna lo stesso numero di bit alle autoinformazioni è ottimale, dunque un possibile codice è il seguente:

$$C(a) = 0, \qquad C(b) = 10, \qquad C(c) = 110, \qquad C(d) = 111.$$