Laboratorio di Statistica Computazionale - 2

Matteo Grigoletto[1]

[1] Il materiale presentato è basato sugli appunti di Domenico De Stefano.

Argomenti di oggi

- Metodi di simulazione Monte Carlo
  - applicazioni al calcolo integrale
  - altre applicazioni
- Campionamento per importanza
- Variabili antitetiche
- Rao-Blackwell

Alcuni esercizi per oggi /1

- Esercizio n. 1.
  a. Valutare via simulazione, fornendo qualche informazione sulla precisione della stima, l'integrale $\int_0^1 h(x)\,dx$, dove $h(x) = [\cos(50\,x) + \sin(20\,x)]^2$. Si usi $B = 10^4$.
  b. Rappresentare graficamente i valori ottenuti per l'integrale, e le rispettive "bande di confidenza", quando si usano $B = 2, 3, \ldots, 10^4$ valori simulati.

- Esercizio n. 2.
  a. Calcolare una stima via simulazione (usando unicamente la generazione da una $U(0,1)$) dell'integrale:

$$\int_0^3 \int_0^3 \int_0^3 \exp(-|x_1 - x_2|^{2+\sqrt{x_3}})\,dx_1\,dx_2\,dx_3 \ .$$

  b. Valutare la precisione della stima ottenuta.

Alcuni esercizi per oggi /2

- Esercizio n. 3.
  a. Con riferimento all'integrale dell'esercizio n. 2, si provi a calcolarlo generando unicamente da una $U(0,3)$.
  b. Valutare, anche in questo caso, la precisione della stima ottenuta.

- Esercizio n. 4. Calcolare una stima via simulazione dell'integrale:

$$I = \int_{-\infty}^{+\infty} \exp(-\sqrt{|x|}) \sin\left(\frac{x^2}{50}\right) dx \ .$$

  In particolare, si usi un approccio basato sul campionamento per importanza generando le x da una normale di media 0 e varianza 400. Si stimi poi lo stesso integrale senza usare il campionamento per importanza, ma riportando il dominio di integrazione all'intervallo $(0,1)$ tramite la trasformazione $y = 1/(1+e^{-x})$.
  Si valutino e confrontino le precisioni della stime ottenute con le due procedure.

Alcuni esercizi per oggi /3

- Esercizio n. 5. Sia Y uniformemente distribuita sull'intervallo $[0, X]$, dove X è una variabile casuale esponenziale con parametro $\lambda = 1$. Si stimino, via simulazione, $E[Y]$ e $P(Y > 1)$, usando le procedure dirette, le procedure basate su Rao-Blackwell e infine usando contemporaneamente Rao-Blackwell e le variabili antitetiche. Si confrontino le precisioni delle diverse procedure di stima.

- Esercizio n. 6. Il cosiddetto *birthday problem* è un risultato, apparentemente paradossale, che mostra che in un gruppo di sole 23 persone la probabilità che due di esse siano nate nello stesso giorno è circa del 50%. Si utilizzi un approccio basato sulla simulazione per mostrare che effettivamente vale questo risultato. Per raggiungere l'obbiettivo, si scriva una funzione per stimare la probabilità che in un gruppo di n persone, con $n > 1$ qualsiasi, almeno due di esse siano nate nello stesso giorno.

Note esercizio 1

- L'integrale di $h(x)$ su $[0,1]$ viene approssimato da
$$\sum h(U_i)/B, \text{ con } U_i \sim U(0,1).$$
- Per integrali unidimensionali è possibile usare la funzione `integrate(<Funzione>, 0, 1)`: si confrontino i risultati ottenuti con le simulazioni con il valore restituito da `integrate`.

Note Esercizio 2

- Dobbiamo trasformare le variabili X_i, $i = 1, 2, 3$, in modo da ottenere nuove variabili, che siano $U(0, 1)$.

Soluzione:

- Ponendo $y_i = x_i/3$, $i = 1, 2, 3$, l'integrale può essere riscritto come:

$$27 \int_0^1 \int_0^1 \int_0^1 \exp(-|3\,y_1 - 3\,y_2|^{2+\sqrt{3\,y_3}})\, dy_1\, dy_2\, dy_3 \ .$$

- Ovvero, l'integrale cercato è $I = E(f(Y_1, Y_2, Y_3))$, dove $Y_1, Y_2, Y_3 \sim IID\ U(0, 1)$, mentre

$$f(Y_1, Y_2, Y_3) = 27\, \exp(-|3\,(y_1 - y_2)|^{2+\sqrt{3\,y_3}}) \ .$$

Note Esercizio 3

- Un'alternativa è procedere via campionamento per importanza. Possiamo scegliere come distribuzione da cui campionare una qualsiasi distribuzione tridimensionale la cui densità non si annulli mai in $[0,3]^3$.

Soluzione:

- Dobbiamo riscrivere I come $E\left(\dfrac{f(X_1,X_2,X_3)}{\psi(X_1,X_2,X_3)}\right)$, dove X_1,X_2,X_3 sono 3 v.c. con densità congiunta $\psi(x_1,x_2,x_3)$, avente supporto che include $[0,3]^3$, mentre $f(x_1,x_2,x_3)=\exp(-|x_1-x_2|^{2+\sqrt{x_3}})$.
- La scelta più banale (e che in realtà ha ben poco del campionamento per "importanza") consiste nell'assumere $X_1,X_2,X_3 \sim IID\ U(0,3)$.
- In questo modo, si ottiene essenzialmente solo una riscrittura della soluzione precedente (senza miglioramenti dell'efficienza).

Note Esercizio 5

- Per generare un valore di Y devo procedere in due passi, generando prima X e poi Y.
- Generando B valori di Y, posso studiare le proprietà della sua distribuzione, con metodi diretti.
- Il vero valore di $E[Y]$ è facile da ottenere analiticamente:

$$E[Y] = E_X[E_Y[Y|X]] = E_X\left[\frac{X}{2}\right] = \frac{1}{2} \ ,$$

  dove abbiamo usato $E[U(0, X)|X] = X/2$. Quest'ultimo risultato suggerisce che la stima per simulazione diretta può essere resa più efficiente applicando Rao-Blackwell.
- Per quanto riguarda $P(Y > 1)$, invece, troviamo:

$$P(Y > 1) = \int_{x=1}^{\infty} P(Y > 1|X = x)\ f(x)\ dx = \int_{x=1}^{\infty} \frac{x-1}{x}\ e^{-x}\ dx$$

  Abbiamo usato $P(Y > 1|X = x) = 0$ se $x < 1$ e $P(Y > 1|X = x) = (x - 1)/x$, altrimenti (e questo suggerisce come applicare Rao-Blackwell). Questo integrale non può essere calcolato esplicitamente.

Note Esercizio 6

- Il giorno di nascita può essere descritto da una variabile casuale discreta uniforme su $1, 2, 3, \ldots, 365$.
- Ad ogni iterazione devono essere generati n giorni di nascita da questa distribuzione.
- A tale scopo, è possibile usare la funzione `sample`.