Variabili casuali, distribuzioni uniformi discrete

Simulazione & Logistica, I modulo
Lezione n. 8

Corso di Laurea in Informatica Applicata
Università di Pisa, sede di La Spezia
A.a. 2008/09, I semestre

Contenuti

- Variabili casuali
- Distribuzioni discrete
- Distribuzione uniforme discreta
- Generare valori casuali con distribuzione uniforme
- In pratica: generatori di base in GS DSLibs

Modellazione stocastica

- Problema di simulazione
  - Alcuni dati della simulazione non sono deterministici
  - Si comportano in modo probabilistico
- Generatori pseudocasuali
  - Generatori di dati
  - Componenti di libreria del motore di simulazione
  - Parametrizzati per riprodurre dati probabilistici
- Dati in ingresso al progetto di simulazione
  - Parte del sistema da modellare
  - Analisi dei dati per trovare i parametri dei generatori
  - Che produrranno i dati per il simulatore

Spazio di probabilità

- Definito dalla tripla (Ω, $\mathscr{F}$, P)
- Ω spazio campione
  - Un insieme di elementi
  - Es. i possibili esiti di un esperimento
- $\mathscr{F}$ spazio degli eventi, famiglia di sottoinsiemi di Ω
  - $\Omega \in \mathscr{F}$
  - $A \in \mathscr{F} \Rightarrow \Omega \setminus A \in \mathscr{F}$
  - $A, B \in \mathscr{F} \Rightarrow A \cup B \in \mathscr{F}$
- $P : \mathscr{F} \to [0, 1]$, funzione di probabilità
  - $P(A) \geq 0, \ \forall A \in \mathscr{F}$
  - $P(\Omega) = 1$
  - $(A, B \in \mathscr{F}) \wedge (A \cap B = \varnothing) \Rightarrow P(A \cup B) = P(A) + P(B)$

Altre proprietà

- Derivabili dalle proprietà della definizione
- $\forall A, B \in \mathscr{F}$
  - $A \cap B \in \mathscr{F}$
  - $B \setminus A \in \mathscr{F}$
  - $P(\varnothing) = 0$
  - $P(\Omega \setminus A) = 1 - P(A)$
  - $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Variabile casuale

- Dato uno spazio di probabilità (Ω, $\mathscr{F}$, P)
- Una variabile casuale X è una funzione

  $X : \Omega \to \mathbb{R}$ tale che
  $\forall\, r \in \mathbb{R}, \{\omega \in \Omega: X(\omega) \leq r\} \in \mathscr{F}$
- Funzione di distribuzione (caratterizza sempre una VC)

  $F_X(x) = P(\{\omega \in \Omega: X(\omega) \leq x\}) = P(X \leq x)$
- Terminologia
  - Il termine “variabile casuale” è tradizionale
  - Non rispetta i significati usuali di “variabile” e di “casuale”
  - È pertanto ambiguo, ma l’uso è ormai consolidato

Significati intuitivi

- Variabile, in matematica
  - Un simbolo per rappresentare uno qualsiasi di un insieme di valori
  - Usata nel calcolo letterale, es. per le funzioni e i loro domini
- Variabile, in informatica
  - Un identificatore associato a un contenitore (memoria)
  - Legato a un insieme stabilito di valori (tipo)
  - Può assumere un valore (assegnamento)
  - Rappresenta il valore nelle espressioni (accesso)
- Variabile casuale
  - Una variabile che non si può assegnare
  - Che rappresenta un valore nelle espressioni
  - Il valore non è mai noto, ma rispetta una legge di probabilità

Uso delle variabili casuali

- Una variabile casuale è una mappa
- Rappresentazione di un fenomeno non prevedibile
  - Es.: l'operazione richiesta da un cliente di un servizio
  - Es.: il tempo di arrivo del cliente
- Analisi della variabile casuale (del fenomeno)
  - Raccolta di dati (serie storiche, campioni, indagini ...)
  - Studio del comportamento probabilistico, ovvero studio della funzione di distribuzione (empirica)
- Riproduzione della variabile casuale
  - Tramite generatori con la "stessa" funzione di distribuzione
  - Per modellare il fenomeno in una simulazione

Esempio, variabile casuale

- A uno sportello si possono fare diverse operazioni:
  - incasso di un assegno, bonifico, versamento
  - Spazio campione: $\Omega = \{ \, inc, bon, ver \, \}$
  - Spazio degli eventi: $\mathscr{F} = 2^{\Omega}$
- Definiamo la mappa: $X : \Omega \rightarrow \mathbb{R}$
  - $X\,(inc) = 0,\; X\,(bon) = 1,\; X\,(ver) = 2$
- È una variabile casuale, infatti:
  - $r < 0$ $\quad \{ \, \omega \in \Omega : X(\omega) \le r \, \} = \varnothing$
  - $0 \le r < 1$ $\quad \{ \, \omega \in \Omega : X(\omega) \le r \, \} = \{ \, inc \, \}$
  - $1 \le r < 2$ $\quad \{ \, \omega \in \Omega : X(\omega) \le r \, \} = \{ \, inc, bon \, \}$
  - $2 \le r$ $\quad \{ \, \omega \in \Omega : X(\omega) \le r \, \} = \{ \, inc, bon, ver \, \} = \Omega$

Esempio, analisi dei dati

- Supponiamo di avere una registrazione di operazioni
  - 10 dati, assunti come campione rappresentativo
  - *inc*, *ver*, *ver*, *inc*, *ver*, *bon*, *inc*, *bon*, *bon*, *inc*
  - 4 *inc*, 3 *bon*, 3 *ver*
- Funzione di distribuzione, definizione

  $F_X(x) = P(\{\ \omega \in \Omega : X(\omega) \leq x\ \}) = P(X \leq x)$
- Funzione di distribuzione (empirica)
  - $x < 0$, $F_X(x) = P(\varnothing) = 0.0$
  - $0 \leq x < 1$, $F_X(x) = P(\{inc\}) = 0.4$
  - $1 \leq x < 2$, $F_X(x) = P(\{inc, bon\}) = 0.7$
  - $2 \leq x$, $F_X(x) = P(\{inc, bon, ver\}) = 1.0$

Distribuzioni discrete

- Variabile casuale discreta

  $X: \Omega \rightarrow x_1, x_2, \ldots, x_k, \ldots$
- Valori in un insieme finito o al più numerabile, è possibile parlare di probabilità di un valore
- Funzione di densità discreta

  $f_X(x) = P(X = x)$ se $x = x_i$ per $i = 1, 2, \ldots k, \ldots$

  $f_X(x) = 0$ altrimenti
- Funzione di distribuzione discreta

  $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$

Media, varianza e dev. standard

- Media

  $\mu_X = E[X] = \sum_i x_i f_X(x_i)$
- Varianza, media degli scarti quadratici rispetto alla media

  $\sigma^2_X = Var[X] = \sum_i (x_i - \mu_X)^2 f_X(x_i)$

  σ_X deviazione standard
- Sono misure della dispersione di X
  - Possono essere stimate empiricamente su un campione
  - Per descrivere la distribuzione di una variabile casuale
  - Per verificare la distribuzione di un generatore
  - Utili nell'analisi dei dati d'ingresso (min, max, mediana, ...)

Distribuzione uniforme discreta

- Una variabile casuale discreta
  $X: \Omega \rightarrow 1, 2, ..., n$
- È una distribuzione uniforme se la f. di densità:
  $f_X(x) = P(X = x) = 1/n$ se $x = 1, 2, ..., n$
  $f_X(x) = 0$ altrimenti
- Media $\mu_X = E[X] = (n + 1) / 2$
- Varianza $\sigma^2_X = Var[X] = (n^2 - 1) / 12$
- È la distribuzione degli eventi equiprobabili
  - Es. ($n = 6$) il lancio di un dado
  - Importante per la costruzione di altre distribuzioni

Generatori di numeri pseudocasuali

- Perché pseudocasuali
  - Macchine deterministiche non generano numeri casuali
  - Alcuni fenomeni fisici sono casuali
- Sequenze cicliche difficili da prevedere
  - Con numeri "sufficientemente" grandi
  - Di periodo "sufficientemente" grande
  - Punto di partenza dato, a seconda degli usi non predicibile
  - Tabelle di 1 000 000 di numeri casuali della RAND C. (1955)
- Requisiti per i generatori usati in simulazione
  - Devono produrre sequenze ripetibili
  - Devono produrre "buone" sequenze
  - Devono essere efficienti

Metodi congruenziali

- Formula generale di un LCG
  - $X_{i+1} = (aX_i + c) \bmod m$
  - X_0 primo numero della sequenza o seme
- Generatori moltiplicativi
  - $c = 0$
  - Al più $m - 1$ numeri generati
  - Lo 0 non compare e non può essere usato come seme
  - Se X_0 e m sono primi fra loro, tutti i numeri sono generati
- Generatori noti
  - Il congruenziale con $m = 2^{32}$ di Knuth & Lewis
  - Il *Minimal Standard* di Park & Miller (con trucco di Schrage)

Caccia al miglior generatore

- Qualità di un generatore
  - Dimensione del periodo e dell'insieme dei valori
  - Proprietà statistiche della distribuzione dei valori
- Prova classica
  - Generazione di punti equidistribuiti in spazi n dimensionali
  - Esempio: scatterplot a 1, 2, ... n dimensioni
- Lagged Fibonacci
  - Serie di Fibonacci: $S_n = S_{n-1} + S_{n-2}$
  - Generatore: $S_n = (S_{n-h} \circ S_{n-k}) \mod m$
  - L'operatore in genere è la somma, il prodotto o lo xor bit a bit
  - Il *Mersenne Twister* MT19937 è una delle variazioni LFG più note

Riferimenti

- M. Pidd, *Computer Simulation in Management Science*, Capp. 11, J. Wiley & Sons, 1998
- G. Gallo, *Note di Simulazione*, capp. 3.1 e 4.4
- Numerical Recipes in C, Cambridge University Press http://www.nrbook.com/
- P. Hellekalek, *Good RNG are (not so) easy to find*, Mathematics and Computers in Simulation, n. 46 (1998)
- G. Marsaglia, *The Marsaglia Post*, sci.stat.math, 1999.01.12