Esperimenti al calcolatore

Andrea Marin

Università Ca' Foscari di Venezia
Dipartimento di Informatica
Corso di Probabilità e Statistica

2009

Indice

Diverse definizioni di probabilità

Storicamente si sono affermate diverse definizioni di probabilità. Le principali sono:

- Definizione classica
- Definizione frequentista
- Definizione soggettiva
- Definizione assiomatica

Definizione classica

- Probabilmente dovuta a Laplace (1800 ca.)
- La probabilità di un evento è definita come il rapporto tra il numero di casi favorevoli all'evento e il numero di casi possibili, purchè questi ultimi siano tutti equiprobabili
- Quali problemi con questa definizione?

Problemi della definizione classica

- È una definizione circolare. Necessita del concetto di equiprobabile per dare la definizione di probabilità
- È applicabile solo in caso di eventi equiprobabili
- Il numero di risultati possibili deve essere finito $\Rightarrow$ non utilizzabile al continuo

Definizione frequentista

- Introdotta da von Mises (1883-1953)
- La probabilità di un evento è definita come il limite cui tende la frequenza relativa dell'evento al crescere del numero degli esperimenti
- Se n_A è il numero di osservazioni dell'evento a su n esperimenti, allora la probabilità $P(A)$ è:

$$P(A) = \lim_{n \to +\infty} \frac{n_A}{n}$$

- Negli esperimenti di laboratorio faremo riferimento a questa definizione (in opposizione a quella assiomatica vista nel corso di teoria)

Limiti della definizione frequentista

- Richiede la ripetibilità degli eventi: intuitivamente invece è possibile assegnare una probabilità anche agli eventi non ripetibili
- La definizione del limite p_a in questo caso non è compatibile con quella dell'analisi.

$$\forall \epsilon > 0, \quad \exists N : \forall n > N, \quad |n_a/N - p_a| < \epsilon$$

- È possibile determinare N?

Definizione soggettiva

- Introdotta fa de Finetti (1906-1985) e Savage (1917-1971)
- La probabilità di un evento è definita come il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica e 0 altrimenti
- Le probabilità devono essere assegnate in modo da evitare la vincita o la perdita certa (criterio di coerenza)
- Problemi: la soggettività (e.g. Superenalotto)

Definizione assiomatica

- Introdotta da Kolmogorov nel 1933
- Non si tratta di una definizione operativa (la probabilità è vista come misura)
- Il calcolo è basato su un insieme di assiomi
- (Dettagli visti a lezione)

Il paradosso delle 3 carte

- Si hanno a disposizione 3 carte. Una ha entrambe le facce rosse, una ha una faccia rossa e una bianca, l'ultima due bianche. Si dispongono le carte voltate a caso e si pesca una carta osservando che il dorso è rosso. Qual è la probabilità che la faccia sia anch'essa rossa?
- Intuzione?
- Approccio classico?
- Approccio frequentista?
- Approccio assiomatico?

Approccio classico

- Identifichiamo ciascuna faccia delle carte con un numero:
  - RR$\longrightarrow$ 12
  - RB $\longrightarrow$ 34
  - BB $\longrightarrow$ 56
- Vediamo tutte le possibili scelte:

Scelta	Altra faccia	Possibile?	Favorevole?
1	2	SI	SI
2	1	SI	SI
3	4	SI	NO
4	3	NO	NO
5	6	NO	NO
6	5	NO	NO

- Numero dei favorevoli su numero dei possibili $= 2/3$

Approccio assiomatico

- Si vuole calcolare
  $P\{\text{faccia nascosta rossa}|\text{faccia visibile rossa}\} = P\{Fn = R|Fv = R\}$
- Notiamo che $P\{Fn = R|Fv = R\} =$
  $P\{\text{Carta tutta rossa}|Fv = R\} = P\{C = RR|Fv = R\}$
- Per il teorema di Bayes abbiamo:

$$P\{C = RR|Fv = R\} = \frac{P\{Fv = R|C = RR\}P\{C = RR\}}{P\{Fv = R\}}$$

- Attribuiamo la misura di probabilità agli eventi
  - Nota che questo viene fatto di solito usando un approccio classico/frequentista
  - $P\{Fv = R\} = 1/2$ (3 facce su 6 sono rosse)
  - $P\{C = RR\} = 1/3$
  - $P\{Fv = R|C = RR\} = 1$

La funzione rand

- `rand`: restituisce un numero pseudo-causale nell'intervallo $(0, 1)$, distribuzione uniforme
- `rand(d)`: restituisce una matrice $d \times d$ di numeri pseudo-casuali nell'intervallo $(0, 1)$, d.u.
- `rand(d1,d2)`: restituisce una matrice $d1 \times d2$ di numeri pesudo-casuli

Ma prima...

Citazione

La generazione dei numeri casuali è troppo importante per essere lasciata al caso. [R.R. Coveyou]

Utilità dei generatori di numeri pseudo-casuali

I generatori di numeri casuali (RNGs) sono utlizzati in molti contesti:

- Esperimenti statistici
- Simulazione di sistemi stocastici
- Analisi numerica basata su metodi Monte-Carlo
- Algoritmi probabilistici
- Computer games
- Crittografia
- Protocolli di comunicazione sicuri

Caratteristiche di un generatore pseudo-casuali

Un generatore di numeri pseudo-casuali simula il comportamento di una sequenza di v.c. indipendenti

- il programma che implementa il generatore è deterministico
- a parità di seme iniziale la sequenza di output è la stessa

Caratteristiche:

- Ciascun numero dell'intervallo dovrebbe avere la stessa probabilità di estrazione
- Le estrazioni non dovrebbero essere correlate

Definizione di RNG

Definition (RNG)

Un RNG può essere definito come una struttra $(S; \mu; f; U; g)$ dove:

- S insieme finito di stati
- μ è una distribuzione di probabilità su S usata per la selezione dello stato iniziale S_0 (seme)
- $f : S \to S$ è una funzione di transizione
- U: è un insieme finito di simboli
- $g : S \to U$ è la funzione di output

Se s_n è lo stato all'n-ma estrazione, lo stato successivo è: $s_{n+1} = f(s_n)$. L'uscita corrispondente è $g(s_{n+1}) \in U$.

Periodo del generatore

Definition (Periodo del generatore)

Il periodo di un generatore RNG è il più piccolo $\ell > 0$ tale che $s_i = s_{\ell+i}$ per ogni $s_i \in S$.

- Proprietà: si può dimostrare che $\ell \leq |S|$.
- Valori grandi di ℓ sono auspicabili

Esempio

Realizzazione di un RNG nell'intervallo $[0, 1]$

- $S = \{0, \ldots, 2^{20}\}$
- $U = [0, 1]$ (attenzione!)
- $$f(s_i) = s_{i+1} = \begin{cases} s_i + 1 & \text{se } s_i < 2^{20} \\ 0 & \text{se } s_i = 2^{20} \end{cases}$$
- $g(s_i) = u_i = s_i / 2^{20}$
- $s_0 = 0$

Attenzione: pessimo generatore! (anche se si pu'o aumentare la periodicità a piacere)

Proprietà di un buon generatore

- *Lunghezza del periodo*: periodi lunghi assicurano che non vi siano cicli prevedibili
- *Efficienza*: uso di poche risorse tempo/memoria
- *Ripetibilità*: partendo dallo stesso seme riproducono la stessa sequenza
- *Portabilità*: devono essere indipdendenti dalla piattaforma

Generatori lineari congruenziali

- Sono generatori molto efficienti e diffusi
- Sono caratterizzati da una funzione f dalla forma:

$$s_{i+1} = f(s_i) = [a \cdot s_i + c](\mathrm{mod} m)$$

- a viene chiamato moltiplicatore
- c incremento

Esempio didattico

Consideriamo il generatore congruenziale con $a =$, $c = 6$, $m = 5$, $s_0 = 1$.
La sequenza generata è la seguente:
$s_0 = 1$, $s_1 = 4$, $s_2 = 3$, $s_3 = 0$, $s_4 = 1$, ...

- Periodo $\ell = 4$
- Alcuni numeri non vengono estratti...

Implementazioni di RNG congruenziale

Implementazione	m	a	c	g
Borland C/C++	2^{32}	22695477	1	bits 30...16
glibc (usato da GCC)	2^{32}	1103515245	12345	bits 30...0
ANSI C: Watcom, Digital Mars, CodeWarrior, IBM VisualAge C/C++	2^{32}	1103515245	12345	bits 30...16
Borland Delphi, Virtual Pascal	2^{32}	134775813	1	bits 63...32 of (seed * L)
Microsoft Visual/Quick C/C++	2^{32}	214013	2531011	bits 30...16
Random class nelle API di Java	2^{48}	25214903917	11	bits 48...17

Problemi dei generatori congruenziali

- Periodo lungo al massimo m
- Il periodo ℓ è determinato dai parametri a, c, m, s_0
- C' una marcata correlazione tra chiamate successive al generatore

Cos'è la correlazione?

- Prendiamo uno spazio k dimensionale
- Siano $(n_1, \ldots, n_k)$ le coordinate di un punto dello spazio
- Se le coordinate del punto sono estrazioni successive del generatore allora i punti si dispongono su iperpiani
- Il numero di iperpiani massimo è dato da $m^{\frac{1}{k}}$.
- Vedi applicazione Java...

RNG Marsenne Twister

- RNG molto veloce (paragonabile alla `rand` dell'ANSI C
- Richiede relativamente poche risorse
- Sviluppato nel 1997 da Matsumoto e Nishimura
- 'E il generatore di default di Octave
- Bassa correlazione
- periodo di 2^{19937}

ma attenzione... Dal manuale di Octave:
"Do not use for cryptography without securely hashing several returned values together, otherwise the generator state can be learned after reading 624 consecutive values."

Modifica del seed

- `v = rand(''state'')`: legge il seed corrente e lo mette nella variabile `v`.
- `rand(''state'', val)`: imposta il seed
- Per defualt il seed è impostato col valore estratto da `/dev/urandom`. Se non è disponibile si usa il clock di sistema.
- `rand(''seed'',val)`: imposta il seed del generatore a congruenza. Se l'ultima impostazione è stata fatta con *seed* verrà utilizzato il generatore congruenziale, se è stata fatta con *state* il Marsenne Twister

Attenzione: nelle relazioni, per una questione di replicabilità degli esperimenti allegare un file con il seed utilizzato.

Esempi

- `n = floor(rand * 100)`: estrae un numero casuale tra 0 e 99
- `d = floor(rand * 6) +1`: estrae un numero casuale tra 1 e 6
- `n = floor(rand(100,1)*50)+ones(100,1)`: estrae un vettore colonna di numeri casuali tra 0 e 49

Quante prove?

- Selezioniamo N numeri casuali in $(0, 1)$ con distribuzione uniforme
- Quanto ci aspettiamo sia la media dei numeri selezionati?
  - 0.5
- Test con Octave...

N	media	errore
10	0.44142	0.05858
50	0.47188	0.028125
100	0.49910	$9.0E-4$
500	0.50276	0.0027559
1000	0.50870	0.0086974

- Come fare quando non si conosce il valore atteso?

Primo esercizio

Si dispone di un dado a sei facce. Un esperimento consiste in questo: lanciato il dado, lo si rilancia soltanto se il valore uscito è minore o uguale a 3. In questo caso, il valore ottenuto è accettabile se la somma dei due lanci è minore o uguale di 6. Riassumendo, l'esito dell'esperimento può essere:

- accettabile con un lancio ($R \geq 4$),
- accettabile con due lanci ($2 \leq R \leq 6$),
- non accettabile.

Si desidera conoscere:

1. Il valore medio dei punteggi totalizzati dagli esperimenti considerati accettabili
2. La probabilità che un esperimento sia accettabile

a- Scrivere un'applicazione che stimi il valore medio e la probabilità per simulazione b- Confrontare i risultati sperimentali con quelli teorici

Da consegnare entro il 22/03/2010

- Documento PDF che illustri i risultati della simulazione e analisi teoriche
  - Introduzione
  - Formalizzazione del problema
  - Analisi teorica
  - Descrizione del simulatore
  - Descrizione dell'esperimento
  - Esisti dell'esperimento
  - Conclusione e discussione dei risultati
- Sorgenti e altri file necessari per ripetere gli esperimenti
- Il tutto va inviato come cartella compressa sulla pagina del corso su Moodle
  - il nome del file deve essere cognome-matricola.zip
  - In relazione e sorgenti indicate Nome, Cognome, Matricola