Identificazione di distribuzioni di probabilità

Simulazione – Lezione n. 8

Corso di Laurea in Informatica Applicata
Università di Pisa, sede di La Spezia

Contenuti

- Identificazione di una distribuzione
- Distribuzioni note
- Profili ...
- ... e segni particolari
- Campionamento e stimatori

Dati da modellare

- Identificazione delle variabili casuali
  - Attributi delle entità, delle loro istanze o delle loro azioni
  - Ricostruire la distribuzione che regola la VC
  - Studiando i dati
- Conoscere le distribuzioni notevoli
  - Conoscere le loro proprietà, per cercarle nel contesto giusto
  - Conoscere profili e segni particolari, per riconoscerle
- Strumenti di riserva
  - Distribuzioni modellabili
  - Per ricostruire da un campione una distribuzione qualsiasi
  - Per costruire ad hoc una distribuzione ragionevole

Identificazione di una distribuzione

- Profilo
  - Funzione di densità di probabilità
  - Funzione cumulativa di distribuzione della probabilità
  - La prima, in genere, è più caratteristica
- Segni particolari
  - Intervallo di definizione, media, mediana, moda, varianza
  - Parametri di forma/scala, propri di alcune distribuzioni
  - Aggiustare il profilo per validare il confronto
- Strumenti per il confronto
  - Foto segnaletiche dei profili
  - Identi-kit: metodo di costruzione del profilo dal campione

Distribuzione uniforme

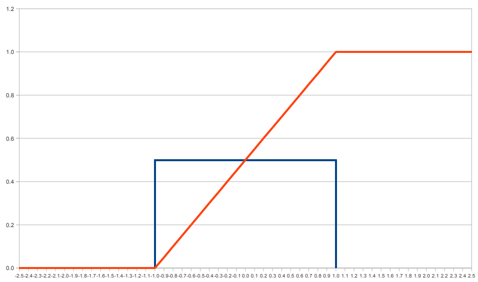
- Continua nei reali, parametri *a* e *b*

$$f_X(x) = \begin{cases} 0 & x<a \\ \dfrac{1}{b-a} & x \in [a,b] \\ 0 & x>b \end{cases} \qquad F_X(x) = \begin{cases} 0 & x<a \\ \dfrac{x-a}{b-a} & x \in [a,b] \\ 1 & x>b \end{cases}$$

- Media e varianza

$$E[X] = \frac{a+b}{2} \qquad Var[X] = \frac{(b-a)^2}{12}$$

Uniforme, profilo



Distribuzione esponenziale

- Continua nei reali positivi, parametro λ

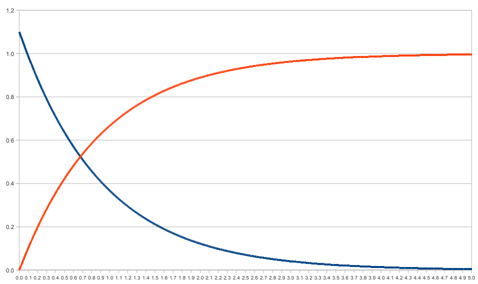
$$f_X(x) = \lambda e^{-\lambda x} \qquad F_X(x) = 1 - e^{-\lambda x}$$

- Media e varianza

$$E[X] = \frac{1}{\lambda} \qquad Var[X] = \frac{1}{\lambda^2}$$

- Caratteristiche
  - Modella il tempo fra eventi successivi
  - Legata alla distr. discreta di Poisson (n. eventi nel tempo)

Esponenziale, profilo



Distribuzione normale

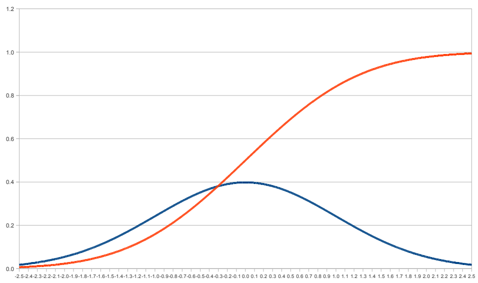
- Continua nei reali, parametri μ e σ

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

- Caratteristiche
  - Modella l'errore rispetto a un obiettivo
  - Media e deviazione standard sono parametri della funzione
  - La funzione di distribuzione è definita solo come integrale
  - Con $\mu = 0$ e $\sigma^2 = \sigma = 1$, è la dist. normale standard N(0, 1)
  - Data z che sia N(0, 1) e x che sia N(μ, σ), allora:

$$z = \frac{x-\mu}{\sigma} \qquad x = \sigma z + \mu$$

Normale, profilo



Distribuzione di Weibull-Rayleigh

- Continua nei reali positivi, parametri k e λ

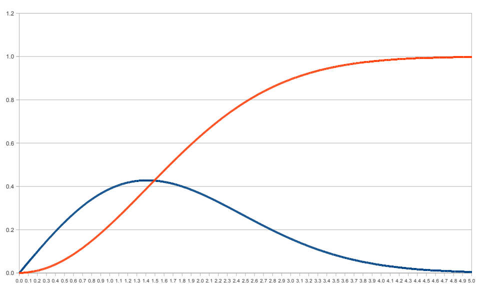
$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \qquad F_X(x) = 1 - e^{-(x/\lambda)^k}$$

- Media e varianza

$$E[X] = \lambda\, \Gamma(1+\frac{1}{k}) \qquad Var[X] = \lambda^2\, \Gamma(1+\frac{2}{k}) - \mu^2$$

- Caratteristiche
  - Modella il tempo fra eventi di guasto (Rayleigh)
  - Parametri di forma $k > 0$ e di scala $\lambda > 0$
  - k = 1 esponenziale 1/λ, k = 2 Rayleigh, k = 3.4 quasi normale

Weibull-Rayleigh, profilo



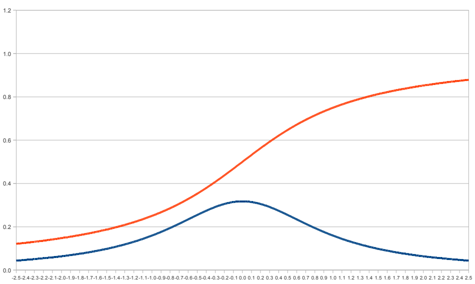
Distribuzione di Cauchy-Lorentz

- Continua nei reali, parametri x_0 e γ

$$f_X(x) = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right] \qquad F_X(x) = \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}$$

- Caratteristiche
  - Modella fenomeni di diffusione omogenea
  - Media e varianza non sono definite
  - Mediana e moda sono un parametro della funzione, x_0
  - Con $x_0 = 0$ e $\gamma = 1$, è la dist. Cauchy standard C(0, 1)

Cauchy, profilo



Distribuzione triangolare

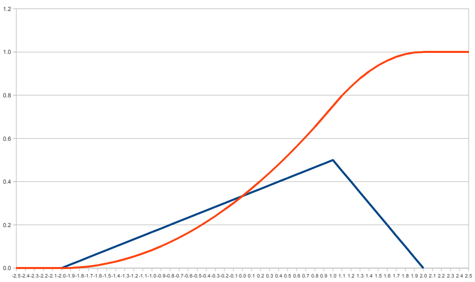
- Continua nei reali, parametri a, b e c

$$f_X(x) = \begin{cases} 0 & x<a \\ \dfrac{2(x-a)}{(b-a)(c-a)} & a \le x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & c \le x < b \\ 0 & x>b \end{cases} \qquad F_X(x) = \begin{cases} 0 & x<a \\ \dfrac{(x-a)^2}{(b-a)(c-a)} & a \le x \le c \\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)} & c \le x < b \\ 1 & x>b \end{cases}$$

- Media e varianza

$$E[X] = \frac{a+b+c}{3} \qquad Var[X] = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

Triangolare, profilo



Considerazioni

- Distribuzioni notevoli
  - Legate a fatti reali, naturali o comuni
  - Uniforme, esponenziale, normale
  - Rayleigh, Cauchy-Lorentz
- Distribuzioni costruite
  - Weibull, da esponenziale a Rayleigh
  - Triangolare
- Obiettivo
  - È sempre possibile riprodurre un campione
  - Anche senza riconoscerne la distribuzione
  - Ma bisogna provarci

L'identi-kit

- Supponendo di avere un testimone
  - Dobbiamo costruire un identi-kit
  - Da confrontare con le foto segnaletiche
- Il testimone: un campione di dati
  - Un numero finito di dati
  - Da trattare come valori di una VC
  - Di cui si vuole identificare la distribuzione
- Occorrono
  - Licenza di trattare insiemi di dati come valori di una VC
  - Meccanismi di costruzione dell'identi-kit
  - Meccanismi di individuazione dei segni particolari

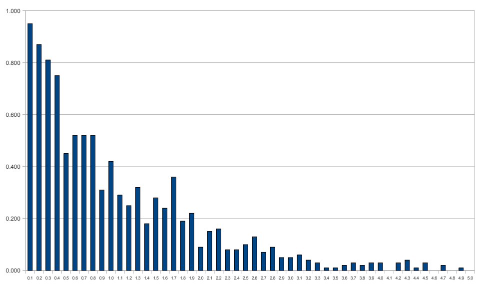
La licenza

- Possiamo trattare come “equivalenti”
  - Un numero n di valori di una VC X con distribuzione F_X
  - I valori di n VC, X_1, X_2, ... X_n, tutte con distribuzione F_X
  - Purché le n VC X_1, X_2, ... X_n siano indipendenti
- Esempio
  - X_i rappresenta il tipo di servizio dell’i-esimo cliente
  - Ogni cliente è indipendente dai precedenti
  - Ma tutti condividono le stesse esigenze di servizio
  - In altre parole, “scelgono” allo stesso modo
  - Allora posso rappresentare i servizi richiesti con una VC la cui distribuzione è la stessa di un campione di scelte

Campionamento

- Tracciare un’istogramma di f_X
  - Partendo da un campione di dati
  - Che sia, con buone motivazioni, rappresentativo
- Ipotesi di distribuzione discreta
  - Valori noti o evidenti dal campione, spesso pochi
  - Contare la frequenza dei singoli valori
- Ipotesi di distribuzione continua
  - Valori non chiaramente identificabili
  - Partizionarli in classi di equivalenza (intervalli)
  - Contare la frequenza dei valori in una classe di equivalenza

Campionamento, esempio



Campionamento, trucchi del mestiere

- Obiettivo
  - Riconoscere la forma
  - Localizzare alcuni segni particolari
- Occhio ai buchi
  - Giocare sulla definizione degli intervalli
  - Troppo larghi: profilo troppo scalettato
  - Troppo stretti: buchi o frastagliature troppo evidenti
- Normalizzazione del conteggio
  - Un istogramma di frequenze è una probabilità nel campione
  - Se la distribuzione è discreta
  - Se la distribuzione è continua il valore va normalizzato

Media del campione

- Media di un numero finito di valori
  - $X_1, X_2, \ldots X_n$, VC indipendenti con stessa distribuzione F
  - Non è nota F, né i suoi segni particolari (media, varianza, ...)

  $$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- Stimatore della media di X con distribuzione F

  $$\mu_X = E[X] = \sum_i x_i f_X(x_i)$$
- Note
  - Calcolata come se il campione avesse distribuzione uniforme
  - La probabilità dei valori vs la frequenza delle loro repliche

Varianza del campione

- Varianza di un numero finito di valori
  - $X_1, X_2, \ldots X_n$, VC indipendenti con stessa distribuzione F
  - Non è nota F, ma è calcolata la media del campione

  $$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
- Stima della varianza di X con distribuzione F

  $$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
- Note
  - Le differenze tendono a 0 al crescere del campione
  - Si usano entrambi come stimatori

Riferimenti

- G. Gallo, *Note di Simulazione*, capp. 3.2 e 4.3
- M. Pidd, Computer Simulation in Management Science, Cap. 11, J. Wiley & Sons, 1998