Modellazione dei dati, valutazione di parametri e ipotesi

Simulazione – Lezione n. 10

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

Contenuti

- Obiettivi e problemi della modellazione dei dati
- Intervalli di confidenza
- Test di ipotesi
- Dipendenze fra i dati
- Coefficiente di correlazione

Modellazione dei dati

- Obiettivi
  - Riconoscere il tipo di distribuzione
  - Stimarne i parametri
  - In sostanza, formulare un'ipotesi
- Problema
  - Una volta arrivati a un'ipotesi ...
  - ... occorre darne una misura dell'affidabilità
- Strumenti
  - Intervalli di confidenza
  - Test di ipotesi

Intervallo di confidenza

- Misura dell'affidabilità di una stima
  - Esprime un margine di errore d ...
  - ... e la probabilità α di commettere un errore più grande
  - Limiti di confidenza, estremi dell'intervallo: $\pm d$
  - Livello di confidenza, probabilità di non errore: $1-\alpha$
  - Definito per una distribuzione o per un campione
- Uso tipico
  - Media del campione
  - Come stima della media della distribuzione

$$P(\mu_X - d < \bar{X}_n < \mu_X + d) = 1-\alpha = P(\bar{X}_n - d < \mu_X < \bar{X}_n + d)$$

Errore standard

- Media del campione $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Varianza del campione $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$
- Errore standard del campione $s\bar{e}_n = \sqrt{\frac{S_n^2}{n}} = \frac{S_n}{\sqrt{n}}$
- Espressione dell'intervallo di confidenza

$$P(\mu_X - c\,s\bar{e}_n < \bar{X}_n < \mu_X + c\,s\bar{e}_n) = 1-\alpha$$

Distribuzione T di Student (Gosset)

- Distribuzione di c come variabile casuale
  - Dipende dai gradi di libertà ($k-1$, con k i valori osservati)
  - Considera una o entrambe le "code"
  - Nota e invertibile
- Calcolo dell'intervallo di confidenza
  - Dato d $\quad c = \frac{d}{s\bar{e}_n} \quad \alpha = F_{T(k-1,\,2)}(c)$
  - Dato α $\quad c = F^{-1}_{T(k-1,\,2)}(\alpha) \quad d = c\,s\bar{e}_n$

Esempio

- Lancio di una moneta (testa 1, croce 2)
  - Esperimento di 10 lanci, esiti: 2, 1, 1, 2, 1, 2, 2, 1, 1, 1
  - Calcoliamo

$$\bar{X}_n = 1.4 \qquad S_n^2 = 0.267 \qquad s\bar{e}_n = \sqrt{\frac{S_n^2}{n}} = 0.163$$

- Fissiamo d oppure α

$$d = 0.5 \qquad c = \frac{d}{s\bar{e}_n} = 3.062 \qquad \alpha = F_{T(1,2)}(c) = 0.201$$

$$\alpha = 0.1 \qquad c = F^{-1}_{T(1,2)}(\alpha) = 6.314 \qquad d = c\, s\bar{e}_n = 1.031$$

Quanta confidenza?

- Affidabilità di una stima
  - L'intervallo di confidenza è espresso da due valori, d e α
  - Legati: se stringo d la probabilità di errore α cresce
  - Dipende soprattutto dalla dimensione n del campione
- Confidenza del 95%
  - Sembrerebbe una quasi certezza
  - Se d è anche piccolo, una quasi certezza di buona precisione
- Attenzione
  - Rimane 1/20 di errore, e l'errore non è quantificato
  - Molto più probabile sbagliare che fare 12 ai dadi

Test di ipotesi

- Ipotizzare la distribuzione di un campione
  - Quanto possiamo fare affidamento sull'ipotesi?

$$H_0 : P(x=i) = p_i \qquad i = 1,..k$$

- Ipotesi nulla H_0
  - Se, sotto l'ipotesi formulata, i dati di un campione sono molto dubbi, l'ipotesi va scartata
  - Altrimenti non si dovrebbero trarre conclusioni
  - Si può però stimare la probabilità di commettere un errore se, sulla base di quei dati, si decidesse di scartare l'ipotesi
  - In sostanza, valutando la probabilità dei dati dubbi

Dati e ipotesi

- Data una VC discreta X, con valori 1, .. k
  - Sia, per l'ipotesi H_0, $P(x = i) = p_i$ per $i = 1, .. k$
  - Sia dato un campione di valori di X di dimensione n
  - Il numero di occorrenze attese di i, sotto l'ipotesi H_0, è np_i
  - Sia N_i il numero di occorrenze di i nel campione (sono VC)

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

- Misura della distanza fra i dati e l'ipotesi
  - Sospettare di valori bulgari (0 per esempio)

X^2 e distribuzione χ^2

- Valore di X^2 calcolato su un campione
  - Più è grande e più l'ipotesi è dubbia, ma è solo un campione
  - Qual'è invece la probabilità che il valore teorico sia minore?
  - Ovvero, quale è la probabilità di sbagliare rifiutando H_0?

$$P_{H_0}(X^2 \le t) \approx P(\chi^2_{k-1} \le t)$$

- Distribuzione di χ^2
  - Nota, dipende dai gradi di libertà ($k - 1$)
  - L'approssimazione vale per n grande ...
  - ... e per np_i non troppo piccoli (> 5 o > 10)

Esempio

- Esperimento su un dado
  - $k = 6$ valori, $n = 1000$ lanci (campione)
  - $N_1 = 172$, $N_2 = 164$, $N_3 = 158$, $N_4 = 170$, $N_5 = 173$, $N_6 = 163$
  - H_0 : distribuzione uniforme (dado onesto)

$$\forall\, i\ p_i = \frac{1}{6} \qquad np_i = 166.67 \qquad X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = 1.052$$

$$P(X^2 \le 1.052) \approx F_{\chi^2_5}(1.052) = 0.958$$

- Confidenza $1 - \alpha$
  - Valori di α standard: 0.05, 0.01, 0.001

Estensione alle distribuzioni continue

- Intervalli invece che valori
  - Campione di n valori
  - k intervalli sul dominio della VC $-\infty, x_1, x_2, \ldots x_i, \ldots x_{k-1}, +\infty$
  - N_i = numero delle occorrenze degli x t.c. $x_i < x \leq x_{i+1}$
  - $p_i = F_X(x_{i+1}) - F_X(x_i)$ dove F_X è identificata sotto H_0
- Scegliere bene gli intervalli
  - Vincoli su np_i
  - Intervalli non uniformi, ma equiprobabili
  - E non troppo piccoli
  - Controllare np_i e aggiustare

Gradi di libertà

- Parametro di funzioni di distribuzione
  - La distribuzione T di Student
  - La distribuzione χ^2
- Significato
  - Numero di informazioni libere
  - Rispetto al calcolo o alla stima di altre informazioni
- Esempi
  - Campione di 2 dati: media 2 GdL, varianza 1 GdL
    infatti i valori sono equidistanti dalla media
  - Campione di n esiti su k valori: k–1 GdL
    infatti $N_k = n - N_1 - N_2 - .. - N_{k-1}$

Dipendenze fra dati

- Variabili casuali diverse, ma fra loro legate
  - Perché, in qualche modo, associate agli stessi fenomeni
  - Esempio: il tipo di cliente e il servizio richiesto
- Analisi
  - In genere le dipendenze sono sospettabili o addirittura note
  - Nei dati si cercano conferme (oltre che i numeri)
  - Individuare le dipendenze e trattare separatamente i dati
- Strumenti
  - Parametri delle distribuzioni
  - Coefficiente di correlazione, scatterplot

Coefficiente di correlazione

- Misura della dipendenza fra due VC X e Y
  - X e Y VC indipendenti $\Rightarrow \varrho_{XY} = 0$
  - $\varrho_{XY} = 1$, correlazione lineare positiva (crescente)
  - $\varrho_{XY} = -1$, correlazione lineare negativa (decrescente)
  - Calcolabile su un campione (stima)

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$\bar{\rho}_{XY} = \frac{\sum_i^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{(n-1)\sigma_X \sigma_Y}$$

Riferimenti

- G. Gallo, *Note di Simulazione*, capp. 3.2, 3.3, 4.3