Dispense di Inferenza statistica

Emanuele Izzo, 0307385

Corso di laurea magistrale Informatica
Università Tor Vergata, Facoltà di Scienze MM.FF.NN.
30/03/2022

Documento realizzato in LaTeX

Indice

Capitolo 1

Introduzione

L'**inferenza statistica** (o **statistica inferenziale**) è il procedimento per cui si inducono le caratteristiche di una *popolazione* dall'osservazione di una parte di essa (detta *campione*), selezionata solitamente mediante un esperimento casuale (aleatorio). Si considereranno principalmente campioni casuali semplici di dimensione $n > 1$, che possono venire interpretati come n realizzazioni indipendenti di un esperimento di base, nelle medesime condizioni. Dal momento che si considera un esperimento casuale, si coinvolge il calcolo delle probabilità. Nell'inferenza statistica c'è, in un certo senso, un rovesciamento di punto di vista rispetto al calcolo delle probabilità. Nell'ambito di quest'ultimo, noto il processo di generazione dei dati sperimentali (*modello probabilistico*) siamo in grado di valutare la probabilità dei diversi possibili risultati di un esperimento. Nella statistica il processo di generazione dei dati sperimentali non è noto in modo completo (il processo in questione è, in definitiva, l'oggetto di indagine) e le tecniche statistiche si prefiggono di indurre le caratteristiche di tale processo sulla base dell'osservazione dei dati sperimentali da esso generati.

1.1 Il problema del compleanno

Il **problema del compleanno** è formalizzato come segue: sono presenti k persone all'interno di una stanza, quale è la probabilità che almeno due persone compiano gli anni nello stesso giorno? Per rispondere a questa domanda risulta più semplice calcolare la probabilità che nessuna delle k persone compia gli anni nello stesso giorno.

$$E := \text{"}i \text{ persone compiono gli anni lo stesso giorno"}$$
$$P(E=0) = \frac{365}{365} \cdot \frac{364}{365} \cdot \ldots \cdot \frac{365-(k-1)}{365} = \Pi_{i=0}^{k-1}(1 - \frac{i}{365})$$
$$P(E \geq 1) = 1 - P(E=0) = 1 - \Pi_{i=0}^{k-1}(1 - \frac{i}{365})$$

Questo prodotto è incredibilmente difficile da calcolare. È possibile però approssimarlo: usando l'approssimazione $e^{-x} \cong 1 - x$, possiamo riscrivere il prodotto sopra in un'altra forma.

$$P(E=0) = \Pi_{i=0}^{k-1}(1-\frac{i}{365}) \cong \Pi_{i=0}^{k-1}e^{-\frac{i}{365}} =$$
$$= e^{-\sum_{i=0}^{k-1}\frac{i}{365}} =$$
$$= e^{-\frac{1}{365}\sum_{i=0}^{k-1} i} =$$
$$= e^{-\frac{1}{365}\cdot\frac{k(k-1)}{2}} =$$
$$= e^{-\frac{k(k-1)}{730}}$$

Da cui abbiamo che $P(E \geq 1) = 1 - e^{-\frac{k(k-1)}{730}}$. Possiamo fare alcune osservazioni dal valore appena calcolato:

- $P(E \geq 1) = \frac{1}{2} \Rightarrow k \cong 23$, ossia se sono presenti circa 23 persone, la probabilità che almeno due persone compino gli anni lo stesso giorno è del 50%.
- $P(E \geq 1) = \frac{19}{20} \Rightarrow k \cong 48$, ossia se sono presenti circa 48 persone, la probabilità che almeno due persone compino gli anni lo stesso giorno è del 95%.

1.2 L'inclusione-esclusione

L'**inclusione-esclusione** è la generalizzazione della probabilità dell'unione di due eventi; la formula generale è definibile come segue: la somma delle probabilità dei singoli eventi meno la probabilità di tutte le possibili intersezioni a due degli insiemi più la probabilità di tutte le possibili intersezioni a tre degli insiemi, e così via, fino a sommare o sottrarre la probabilità dell'intersezione tra tutti gli eventi. Esiste una dimostrazione interessante per la formula di inclusione-esclusione: partiamo dal definire una varaibile aleatoria *indicatrice* o *unitaria* per un certo evento A.

$$1_A = \begin{cases} 1 \text{ se } A \text{ avviene} \\ 0 \text{ altrimenti} \end{cases}$$

Un'osservazione importante è che la media di una variabile aleatoria unitaria di un evento è uguale alla probabilità dell'evento, ossia $\mathbb{E}[1_A] = P(A)$. Possiamo inoltre notare alcune cose riguardo queste variabili aleatorie unitarie:

$$1_{A^C} = 1 - 1_A$$
$$1_{A\cap B} = 1_A 1_B$$
$$1_{A\cup B} = 1_A + 1_B$$

Possiamo sfruttare queste variabili aleatorie unitarie in combinazione con le leggi di De Morgan:

$$P((A_1 \cup A_2 \cup ... \cup A_n)^C) =$$
$$= 1 - P(A_1 \cup A_2 \cup ... \cup A_n) =$$
$$= P(A_1^C \cap A_2^C \cap ... \cap A_n^C) =$$
$$= \mathbb{E}((1-1_{A_1})(1-1_{A_1})...(1-1_{A_1}))$$

Essendo quest'ultimo però un prodotto tra quantità numeriche può essere espresso come:

$$\begin{aligned}\mathbb{E}[(1-1_{A_1})(1-1_{A_1})...(1-1_{A_1})] = \mathbb{E}[1-(1_{A_1}+1_{A_2}+...+1_{A_1})+ \\ +(1_{A_1}1_{A_2}+1_{A_1}1_{A_3}+...+1_{A_1}1_{A_n}+1_{A_2}1_{A_3}+...+1_{A_{n-1}}1_{A_n})-...\pm 1_{A_1}1_{A_2}...1_{A_n}]\end{aligned}$$

Operando poi sulla media si ottiene la formula descritta prima.

1.3 Il problema degli ombrelli

Il **problema degli ombrelli** è formalizzato come segue: sono presenti n persone all'interno di una stanza, ognuno con il proprio ombrello. Tutte le persone posano il proprio ombrello all'interno di un portaombrelli; una volta posati tutti gli ombrelli nel portaombrelli, ogni persona, una alla volta, prende uno degli ombrelli al suo interno casualmente. Quale è la probabilità che nessuno abbia preso il proprio ombrello? Questo problema corrisponde, partendo da una sequenza iniziale di n elementi, alla probabilità di scegliere una delle $n!$ permutazioni possibili di essi e avere che un elemento abbia stessa posizione nella permutazione e nella sequenza iniziale. Definiamo le variabili aleatorie $A_k :=$ "l'elemento k-esimo è k" e $E :=$ "n elementi sono nella posizione corretta", la probabilità a cui siamo interessati è $P(E=0) = 1 - P(A_1 \cup A_2 \cup ... \cup A_n)$. Poichè gli eventi non sono indipendenti, è necessario utilizzare la formula di inclusione-esclusione; una cosa interessante che si può appuntare è che le intersezioni sono interpretabili come "1 o più elementi sono al punto giusto" (ossia $E \geq 1$) e le relative probabilità sono facili da calcolare data la struttura delle permutazioni: la probabilità che k sugli n elementi siano nella posizione giusta corrisponde a fissare una delle possibili combinazioni di k elementi e permutare i restanti $n-k$, e la probabilità di ciò è pari a :

$$P(E=k) = \binom{n}{k}\frac{(n-k)!}{n!} = \frac{n!}{k!\,(n-k!\,)}\frac{(n-k)!}{n!} = \frac{1}{k!}$$

Pertanto la probabilità che a noi interessa, considerando i segni delle varie intersezioni, sarà uguale a:

$$P(E=0) = 1 - \sum_{i=1}^{n}(-1)^{i-1}P(E=i) = 1 - (\sum_{i=1}^{n}(-1)^{i-1}\frac{1}{i!}) = \sum_{i=0}^{n}(-1)^{i}\frac{1}{i!} \cong e^{-1}$$

Possiamo notare pertanto che la probabilità che ci interessa si avvicina, con n crescente, molto rapidamente a $e^{-1} \cong \frac{1}{3}$

Capitolo 2

Elementi di probabilità

2.1 La teoria degli insiemi

Definizione (2.1): L'insieme S di tutti i possibili risultati di un particolare esperimento è chiamto lo **spazio campionario** dell'esperimento.

Definizione (2.2): Un **evento** è una qualsiasi collezione di possibili risultati di un esperimento, ossia un qualsiasi sottoinsieme di S (incluso S stesso).

Sia A un evento, ossia un sottoinsieme di S. Diciamo che l'evento A avviene se il risultato dell'esperimento è contenuto nell'insieme A. Definito un secondo evento B, possiamo definire le seguenti relazioni e operazioni:

- *Contenimento*: $A \subset B \Leftrightarrow [x \in A \Rightarrow x \in B]$.
- *Uguaglianza*: $A = B \Leftrightarrow A \subset B$ e $B \subset A$.
- *Unione*: $A \cup B := \{x | x \in A$ o $x \in B\}$.
- *Intersetzione*: $A \cap B := \{x | x \in A$ e $x \in B\}$.
- *Complementare*: $A^C := \{x | x \notin A\}$.

Teorema (2.1): Per qualsiasi tre eventi $A, B, C \in S$:

- *Commutatività*:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- *Associatività*:

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

- *Leggi distributive*:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- *Leggi di De Morgan*:

$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C$$

Definizione (2.3): Due eventi $A, B \in S$ sono **disgiunti** (o **mutualmente esclusivi**) se $A \cap B = \emptyset$. Gli eventi $A_1, A_2, ..., A_n \in S$ sono **disgiunti a due a due** se per ogni $i \neq j$ si ha che $A_i \cap A_j = \emptyset$.

Definizione (2.4): Se $A_1, A_2, ..., A_n \in S$ sono disgiunti a due a due e $\cup_{i=1}^n A_i = S$, allora la collezione $A_1, A_2, ..., A_n$ forma una partizione di S.

2.2 Le basi della teoria della probabilità

2.3 Il momento e il momento centrato

Definizione (2.1): Sia X una variabile aleatoria, il **momento** k-esimo di X è pari a

$$\mu'_n = \mathbb{E}[X^n]$$

Mentre il **momento centrato** k-esimo di X è pari a

$$\mu_n = \mathbb{E}[(X - \mu)^n]$$

Dove $\mu = \mu'_1 = \mathbb{E}[X]$.

Possiamo notare alcune cose interessanti:

- Il primo momento di X è pari alla *media* di X, ossia $\mu_1' = \mathbb{E}[X]$.
- Il secondo momento centrato di X è pari alla *varianza* di X, ossia:

$$\mu_2 = Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$$

- Prese due costanti $a, b \in \mathbb{R}$, abbiamo che:

$$\begin{aligned} &\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \\ &Var[aX + b] = a^2 Var[X] \end{aligned}$$

- Data una seconda variabile aleatoria Y, abbiamo che:

$$\begin{aligned} &\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \\ &\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] && \text{se } X \text{ e } Y \text{ sono indipendenti} \\ &Var[X + Y] = Var[X] + Var[Y] && \text{se } X \text{ e } Y \text{ sono indipendenti} \end{aligned}$$