

Orali del 5 ottobre 2020

Alessandro Di Gioacchino

5 febbraio 2021

(Estratto del primo orale, potrei aver interpretato male le domande)

Immaginiamo di avere un campione casuale estratto da una popolazione esponenziale: come ne stimiamo la mediana?

Se X_1, \dots, X_n indica il campione, cosa indica X ? La popolazione. Nel problema della stima parametrica, Θ indica il parametro ignoto (che potrebbe non essere esattamente ciò che voglio stimare): non essendo la mediana un parametro della distribuzione, non posso usarlo come Θ . Supponiamo quindi che $\Theta = \lambda$: per stimare il parametro, possiamo usare il reciproco della media campionaria. Infatti sappiamo stimare 'facilmente' il valore atteso di una popolazione, ed il valore atteso di un'esponenziale è $\frac{1}{\lambda}$

Per stimare la mediana di un'esponenziale, che abbiamo dimostrato essere uguale a $\frac{\ln 2}{\lambda}$, è quindi ragionevole utilizzare $\ln 2 \bar{X}$ come stimatore.

Torniamo alla popolazione esponenziale di cui sopra (il candidato non aveva risposto): se ne indichiamo la mediana con m , $m = \frac{\ln 2}{\lambda}$

Vogliamo stimare m , quindi $\tau(\Theta) = m$ e $\tau(\lambda) = \frac{\ln 2}{\lambda}$

La media campionaria è sempre uno stimatore consistente e non deviato per il valore atteso di una popolazione, ma ora vogliamo stimare λ (il parametro di un'esponenziale): non possiamo usare la media campionaria così com'è.

Dobbiamo fare un ragionamento che coinvolga il campione: riscriviamo il fatto che il valore atteso di un'esponenziale è $\frac{1}{\lambda}$ mettendo il campo la media campionaria (suppongo dovesse scrivere qualcosa del tipo: $\frac{1}{\lambda} = \mathcal{E}(\bar{X})$). In conclusione, se voglio stimare λ utilizzo il reciproco della media campionaria; se voglio stimare $m = \frac{\ln 2}{\lambda}$, utilizzo la media campionaria moltiplicata per $\ln 2$

Lo stimatore ottenuto è non deviato per la mediana? Il valore atteso della media campionaria è il valore atteso della popolazione, cioè $\frac{1}{\lambda}$; moltiplicando questo per $\ln 2$, otteniamo proprio quanto vogliamo stimare.

Curve ROC. L'estremo destro della curva è sempre il punto $(1, 1)$?

Abbiamo un'urna con 10 palline, di cui 4 nere e 6 bianche. Vogliamo contare il numero di palline bianche estratte. Quale distribuzione può modellare questa situazione? Potrebbero andar bene la ipergeometrica (estrazione senza re-immissione) o la binomiale (estrazione con re-immissione). Esistono altri modelli che posso applicare?

La distribuzione ipergeometrica è pensata proprio per situazioni di questo tipo. Supponiamo invece di reimmettere le palline una volta estratte: quali sono i parametri del modello binomiale risultante? Come calcoliamo la probabilità di non estrarre alcuna pallina bianca? Qual è la probabilità di estrarre almeno una pallina bianca?

Il coefficiente binomiale $\binom{n}{k}$ indica il numero di modi in cui posso costruire sottoinsiemi di k elementi a partire da n : se $k = 0$, il coefficiente binomiale vale 1 perché l'insieme vuoto è unico.

Dimostrazione: la probabilità dell'evento complementare è uguale ad $1 -$ la probabilità dell'evento. Per dimostrarlo, usiamo gli assiomi di Kolmogorov in due punti: quando diciamo che la probabilità dell'evento certo è uno, e quando calcoliamo la probabilità dell'unione di eventi disgiunti come la somma delle singole probabilità.

Abbiamo una popolazione X , di cui conosciamo solo il valore atteso μ e la deviazione standard σ ; vogliamo stimarne la varianza usando il seguente:

$$T^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

Quando uno stimatore si dice non deviato? Perché è una proprietà desiderabile?

È necessario elevare quella quantità al quadrato per calcolare il valore atteso di T^2 ? Cosa sappiamo di X_i ? Fa parte di un gruppo di variabili aleatorie indipendenti ed identicamente distribuite. Posso dire che il valore atteso di X_i è μ ? Perché sì? Perché seguono la stessa distribuzione della popolazione.

Cos'è uno stimatore? Uno stimatore è una funzione del solo campione; invece T^2 è anche funzione di μ : se μ è ignoto, T^2 non è uno stimatore.

Classificatori naive Bayes. Indichiamo con una variabile aleatoria Y l'esito della classificazione, e condizioniamo rispetto ad altre due variabili aleatorie, X_1 ed X_2 , che rappresentano due attributi di interesse. Su quale ipotesi si basano questi classificatori "ingenui"? Stiamo supponendo che gli eventi X_i (condizionati ad Y) siano indipendenti. C'è un altro punto di interesse, cioè non considerare il denominatore nel teorema di Bayes: perché? Cosa c'è di importante nel denominatore che mi permette di ignorarlo? Non dipende da k

Concetto di indipendenza: a cosa può riferirsi?

Gli assiomi di Kolmogorov sono sempre validi, per cui sappiamo che la probabilità dell'unione di eventi disgiunti è uguale alla somma delle singole probabilità; inoltre questo aspetto non c'entra con l'indipendenza. Cosa deve succedere perché due variabili aleatorie siano indipendenti?

Un evento si verifica; una variabile aleatoria no, però mi permette di ragionare in termini dell'evento "la variabile aleatoria ha assunto un certo valore?"

Che ruolo hanno x ed y , le specificazioni delle due variabili aleatorie? Possono assumere soltanto valori particolari? Non solo la definizione deve valere per ogni x ed y , ma in realtà si fa riferimento alla possibilità che i valori assunti dalle variabili aleatorie appartengano ad un certo insieme:

$$\forall A, B \quad P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \iff X, Y \text{ indipendenti}$$

Quali interessanti proprietà possiamo ricavare sapendo che X ed Y sono indipendenti? Per esempio, se sono indipendenti la varianza della somma è uguale alla somma delle varianze: perché?

Dimostriamo che $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

Ragioniamo sul concetto di indipendenza nell'ambito di statistica descrittiva. Come genero un diagramma di dispersione? Se prendo un generico punto del diagramma, cosa posso dire circa le sue coordinate? Sono le due componenti di una stessa coppia. Come calcoliamo il grado di correlazione dei due campioni? Se l'indice di correlazione campionaria è uno, i valori dei campioni si allineano su una qualsiasi retta. Cosa indica la presenza di una relazione diretta tra due quantità? Che al crescere di una quantità, cresce anche l'altra: per cui va bene ogni retta, purché abbia coefficiente angolare positivo.

Consistenza in statistica inferenziale. Lo scarto quadratico medio misura l'errore fatto stimando la quantità sconosciuta con lo stimatore: e il bias invece? In realtà, lo MSE si può scrivere come somma della varianza e del bias (quindi l'esaminando stava facendo un errore).

Concetto di indipendenza.

Se $P(E \cap F) = P(E)P(F)$, allora gli eventi E ed F sono indipendenti. In che modo questa definizione è legata alla probabilità condizionata?

Come estendiamo la definizione di indipendenza tra due eventi a tre eventi? Non basta imporre che siano indipendenti a coppie. Guarda il foglio #1

Cosa vuol dire che la media campionaria gode della proprietà di linearità?

Calcoliamo la varianza dello stimatore "media campionaria." È importante sottolineare che le conclusioni a cui si arrivano (in particolare che la stima migliora con il crescere del campione) sono legate al fatto che la media campionaria è uno stimatore non distorto per il valore atteso della popolazione.

Modello uniforme discreto. Grafico della funzione di ripartizione. Quanto vale questa funzione tra 0 e 1? Perché è ovvio che valga zero? Quali sono le specificazioni di questa variabile aleatoria? Ricorda sempre una variabile aleatoria discreta uniforme modella l'esito del lancio di un dado.

Varianza della media campionaria.

È sempre vero che la varianza di una somma è uguale alla somma della varianza? Perché la varianza delle singole X_i è uguale alla varianza della popolazione?

Quale teorema possiamo applicare per approssimare la distribuzione della media campionaria?

Per n che tende a $+\infty$, la distribuzione espressa dal teorema del limite centrale non è più approssimata ma esatta. La distribuzione della media campionaria ottenuta tramite il teorema del limite centrale è coerente con il calcolo della sua varianza?

Eterogeneità.

Perché l'indice di Gini è compreso tra zero (incluso) ed uno (escluso)? L'indice vale zero nel caso di massima omogeneità: dimostrazione.

Un'applicazione dell'indice di Gini (ed in generale degli indici di eterogeneità). Gli alberi di decisione determinano la domanda da fare in base al valore dell'indice di Gini (o dell'entropia).

La variabile aleatoria Z è data dalla differenza tra due normali standard, X ed Y

Cosa possiamo dire del valore atteso di Z ? Perché vale anch'esso zero? Non è vero che, siccome X ed Y sono

normali standard, allora anche Z lo è: il concetto di riproducibilità è legato alla famiglia della distribuzione, quindi al massimo potrei dire che anche Z è normale. Cosa c'è di strano nel dire che $Z \sim N(0,0)$? Cosa vuol dire che una variabile aleatoria ha varianza nulla? È sensato che la differenza di due normali standard dia origine ad una costante? Sarebbe come lanciare due dadi, sottrarre il risultato di uno al risultato dell'altro ed ottenere sempre lo stesso numero.

Calcoliamo la varianza di Z , ipotizzando pure che X ed Y siano indipendenti. Cosa succede alla varianza se $Z = X - 2Y$?

Quindi, $Z \sim N(0, \sqrt{2})$

Quale ulteriore modifica devo apportare perché Z sia una normale standard? La trasformazione si applica alla variabile aleatoria, non alle sue specificazioni. Basta quindi dividere Z per $\sqrt{2}$

Abbiamo una popolazione distribuita secondo il modello normale, con media 0 e deviazione standard ignota. Come posso stimare tale parametro? Di quali proprietà gode lo stimatore “deviazione standard campionaria”? È non deviato? Per dirlo, dovremmo calcolarne il valore atteso. Non abbiamo gli strumenti matematici adatti a farlo, anche perché il valore atteso non si può portare dentro una radice quadrata (la deviazione standard è la radice quadrata della varianza).

Curve ROC.