Kako se lotiš: Statistika

Patrik Žnidaršič

Prevedeno 14. april 2024

1 Centralni limitni izrek

Izrek (centralni limitni izrek)**.** *Naj bodo* $X_1, X_2, \ldots$ *neodvisne in enako porazdeljene slučajne spremenljivke s končnim drugim momentom. Označimo* $\mu_1 = E(X_1)$, $\sigma_1^2 = \operatorname{var}(X_1)$ *in* $S_n = X_1 + \cdots + X_n$. *Za* $W_n = \frac{S_n - \mu_1 n}{\sigma_1 \sqrt{n}}$ *potem velja*

$$\lim_{n\to\infty} P(W_n \leq w) = \phi(w)$$

enakomerno za $w \in \mathbb{R}$, *torej*

$$\lim_{n\to\infty} \sup_{w\in\mathbb{R}} |P(W_n \leq w) - \phi(w)| = 0.$$

Izrek lahko uporabimo za ocenjevanje porazdelitve vsote veliko IID slučajnih spremenljivk. Izračunamo $\mu_1 = E(X_1)$ in $\sigma_1^2 = \operatorname{var}(X_1)$ in ocenimo verjetnost kot

$$P(S_n \leq x) = \phi\left(\frac{x - n\mu_1}{\sigma_1\sqrt{n}}\right).$$

Za x vzamemo mejo, ki nas zanima, pri čemer za vsote z vrednostmi na mreži $a\mathbb{Z} + b$ za $a, b \in \mathbb{N}$ po dogovoru vzamemo srednjo vrednost; torej za ocenjevanje verjetnosti, da je padlo manj kot M pik, npr. vzamemo $x = M - \frac{1}{2}$, za več kot M pa $x = M + \frac{1}{2}$.

Povemo lahko tudi nekaj o napaki te ocene. Če so $X_1, X_2, \ldots$ neodvisne, za $\mu_n = E(S_n)$ in $\sigma_n^2 = \operatorname{var}(S_n)$ velja

$$\sup_{x\in\mathbb{R}} \left| P(S_n \leq x) - \phi\left(\frac{x - \mu_n}{\sigma_n}\right) \right| \leq \frac{0.5583}{\sigma_n^3} \sum_{k=1}^{n} E\big((X_k - E(X_k))^3\big).$$

Pri tem nismo predpostavili, da so spremenljivke enako porazdeljene, torej lahko to oceno uporabimo v več primerih. Če npr. pokažemo, da desna stran konvergira k 0 za $n \to \infty$, lahko pokažemo rezultat CLI tudi za različno porazdeljene slučajne spremenljivke.

2 Konvergenca slučajnih spremenljivk

Zaporedje $X_1, X_2, \ldots$ konvergira proti X

- ŠIBKO, če je za vsako zvezno in omejeno h

$$\lim_{n\to\infty} E(h(X_n)) = E(h(X)),$$

- V VERJETNOSTI, če je za vsak $\varepsilon > 0$

$$\lim_{n\to\infty} P(d(X_n, X) > \varepsilon) = 0,$$

- SKORAJ GOTOVO, če je

$$P\Big(\{\lim_{n\to\infty} X_n = X\}\Big) = 1.$$

Tretja točka implicira drugo, druga pa prvo. Če imajo spremenljivke vrednosti v $\mathbb{R}$, je prva točka ekvivalentna pogoju, da

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$$

za vsak x z $P(X = x) = 0$.

3 Cenilke

Recimo, da imamo model nekega dogajanja in želimo preveriti, če drži vodo. Karakteristika modela y je neka lastnost (parameter, ...), ki je za ta model značilna, mi pa je ne poznamo. Iz opažanj lahko ocenimo vrednost te karakteristika $\hat{y}$, oceni pravimo CENILKA. Za cenilko definiramo PRIČAKOVANO ali SREDNJO KVADRATIČNO NAPAKO

$$\mathrm{MSE}(\hat{y}\,|\,y) = E((y - \hat{y})^2)$$

($\hat{y}$ je slučajna spremenljivka, pisana z malo). Poleg tega definiramo PRISTRANSKOST

$$\mathrm{Bias}(\hat{y}\,|\,y) = E(\hat{y}) - y.$$

Cenilka je NEPRISTRANSKA, če je $\mathrm{Bias}(\hat{y}\,|\,y) = 0$. Pri takih cenilkah je MSE enaka varianci, in lahko definiramo

$$\mathrm{SE}(\hat{y}) = \mathrm{RMSE}(\hat{y}\,|\,y) = \sqrt{\mathrm{var}(\hat{y})},$$

v splošnem pa dobimo

$$\mathrm{var}(\hat{y}) = \mathrm{MSE}(\hat{y}\,|\,y) - \left(\mathrm{Bias}(\hat{y}\,|\,y)\right)^2.$$

Če podatke dobivamo postopoma, dobimo zaporedje cenilk $(\hat{y}_i)_i$. Za tako zaporedje pravimo, da je ŠIBKO DOSLEDNO, če

$$\hat{y}_n \xrightarrow[n\to\infty]{d} y.$$

Zadosten pogoj je, da srednja kvadratna napaka konvergira k 0, čemur pravimo DOSLEDNOST.

Pogosta tema je iskanje NAJBOLJŠE NEPRISTRANSKE LINEARNE CENILKE (NNLC). To je preprosto cenilka, ki ima izmed vseh linearnih cenilk najmanjšo srednjo kvadratno napako. Če so $X_1, \dots, X_n$ nekorelirane z enakimi pričakovanimi vrednostmi in variancami, je NNLC kar povprečje.

4 Slučajni vektorji

Za slučajni vektor $\underline{X}$ definiramo PRIČAKOVANO VREDNOST

$$E(\underline{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix},$$

za par $\underline{X}, \underline{Y}$ pa KOVARIANČNO MATRIKO

$$\operatorname{Cov}(\underline{X}, \underline{Y}) = \begin{bmatrix} \operatorname{cov}(X_1, Y_1) & \cdots & \operatorname{cov}(X_1, Y_n) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, Y_1) & \cdots & \operatorname{cov}(X_n, Y_n) \end{bmatrix} = E(\underline{X}\underline{Y}^T) - E(\underline{X})E(\underline{Y})^T.$$

Za deterministično matriko A je $E(A\underline{X}) = AE(\underline{X})$, podobno za slučajno matriko M velja $E(AM) = AE(M)$. Analogni enakosti dobimo za množenje z desne. Za deterministični matriki A, B in slučajna vektorja $\underline{X}, \underline{Y}$ je

$$\operatorname{Cov}(A\underline{X}, B\underline{Y}) = A \operatorname{Cov}(\underline{X}, \underline{Y}) B^T.$$

Dva omembe vredna trika pri obračanju slučajnih matrik sta s sledjo. Ker je sled linearna preslikava, se lepo obnaša s pričakovano vrednostjo, in je $E(\operatorname{sl}(M)) = \operatorname{sl}(E(M))$. Spomnimo se, da je skalar enak svoji sledi, da je sled linearna, $\operatorname{sl}(A + B) = \operatorname{sl} A + \operatorname{sl} B$, in da lahko ciklično zamenjujemo argumente:

$$\operatorname{sl}(ABC) = \operatorname{sl}(BCA) = \operatorname{sl}(CAB).$$

5 Pridobivanje cenilk

Pri enostavnem slučajnem vzorčenju imamo populacijo velikosti N in vzorec velikosti n. Iz populacije izberemo enote $K_1, \dots, K_n$, $K_i \neq K_j$, tako, da so vse n-terice enako verjetne. Označimo $X_i = X_{K_i}$. Imamo nepristranske cenilke

$$\hat{\mu} = \frac{1}{n}(X_1 + \ldots + X_n),$$

za $S = \sum_i (X_i - \hat{\mu})^2$

$$\hat{\sigma^2} = \frac{N-1}{N}\frac{S}{n-1},$$
$$\hat{\mu^2} = \hat{\mu}^2 \frac{N-n}{N}\frac{S}{n(n-1)},$$

in napaka

$$\hat{\mathrm{SE}^2} = \frac{N-n}{N-1}\frac{\hat{\sigma^2}}{n}.$$

Pri stratificiranem vzorčenju je populacija razdeljena na k stratumov z velikostmi $N_1, \ldots, N_k$. Označimo $w_i = N_i/N$. Na vsakem stratumu vzorčimo enostavno slučajno.

Poznamo dve pogosti metodi za pridobivanje cenilk. Prva je METODA MOMENTOV, kjer cenilke izpeljemo iz predpisov za momente (potrebujemo toliko momentov, kolikor je argumentov v porazdelitvi). Potem poiščemo inverzno preslikavo, s katero izrazimo cenilke z momenti;

$$(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) = F^{-1}(\overline{X}, \overline{X^2}, \ldots, \overline{X^m}).$$

Druga metoda je METODA NAJVEČJEGA VERJETJA. Pri tej metodi iz opažanja $X = (X_1, \ldots, X_n)$ in gostote $f_X(x \,|\, \theta)$ pridobimo cenilko za θ kot

$$\hat{\theta} = \arg\max L(\theta \,|\, x)$$

kjer je $L(\theta \,|\, x) = f_X(x \,|\, \theta)$ verjetje. V diskretnem primeru zamenjamo f_X z verjetnostjo $P(X = x \,|\, \theta)$. Pogosto se splača namesto verjetja maksimizirati njegov logaritem, torej rešiti $\partial_\theta \log L = 0$.