Pēteris Račinskis
pr20015

Mājas darbs 2.4: MLE un Baijesa estimatori

Uzdevums 1

a. Maksimālā ticamība binomiālam sadalījumam

Piemērā datu kopai
$\vec{x} = (x_1, ..., x_n), x_i \in \{0, 1\}, |\{x_i = 1\}| = r$
par modeli izvēlēts binomiāls sadalījums ar parametru
$\theta : P[x_i = 1]$

un veikta maksimāli ticama modeļa parametra piemeklēšana. Izmantotas trīs piemēra datu kopas - ar n=10, r=7, n=100, r=70 un n=1000, r=700. Ticamības funkcijas $L(\theta|\vec{x})$ argumentu kopa ir ortogonāla parametriskas varbūtības (vai nepārtrauktā gadījumā - blīvuma) funkcijas $f(\vec{x}|\theta)$ argumentiem - ja pierasts domāt par varbūtībām, mainot gadījuma lielumu (un parametrus kā sadalījuma vidējo vērtību vai dispersiju nemaz nepierakstīt), tad ticamības funkcijai gadījuma lielums ir fiksēts un tiek salīdzināti tam atbilstošie varbūtības blīvumi (vai varbūtības, ja runa ir par diskrētiem sadalījumiem kā šeit) pie dažādām parametru θ vērtībām. Jāpiemin, ka piemērā īstenībā aprēķini veikti ar
$L(\theta|r) = L(\theta|g(\vec{x})), g : \{0, 1\}^n \to \mathbb{Z}$

un ticamības funkcija neņem vērā $\binom{n}{r}$ reizināmo binomiālā sadalījuma funkcijā, jo tas ir vienāds visām θ vērtībām - tāpēc $L(\theta|\vec{x}) \neq f(\vec{x}|\theta)$. Piemērā definēta ticamības funkcija pēc θ parametrizētam binomiālam sadalījumam, atrasta logaritmiskā ticamība $l(\theta|\vec{x}) = \log(L(\theta|\vec{x}))$, jo tā saglabā sākotnējās funkcijas ekstrēmus, bet ar to ir vieglāk veikt aprēķinus, un atrasts ticamības funkcijas ekstrēms, iestatot $[l(\theta|\vec{x})]'_\theta = 0$ un atrisinot vienādojumu. Izrādās, ka atrisinājums ir visnotaļ intuitīvi saprotamais $\theta = P[x_i = 1] = \frac{r}{n}$, kas nav īpaši liels pārsteigums.

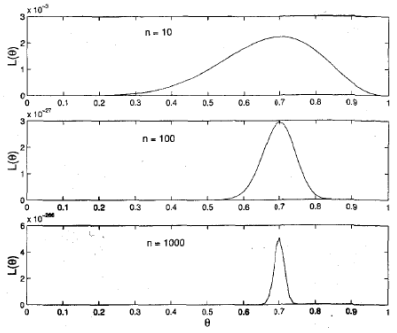

Figure 4.2 The likelihood function for three hypothetical data sets under a Binomial model: $r = 7, n = 10$ (top), $r = 70, n = 100$ (center), and $r = 700, n = 1000$ (bottom).

Piemēra grafikos attēlotas ticamības funkcijas pie dažādiem datu kopas apjomiem. Mazā datu kopā $L(\theta'|\vec{x})$vērtības ir lielākas pie lielākiem $|\theta - \theta'|$, jo $L(\theta|\vec{x})$ vērtības ir atkarīgas no datu kopas izmēriem - lielām datu kopām lielas izlases nobīdes no tās veidojošā sadalījuma ir mazāk ticamas.

b. Maksimālā ticamība normālam sadalījumam

Nākamajā piemērā faktiski atkārtota tā pati procedūra, tikai šoreiz ar citādu datu kopu

$$\vec{x} = (x_1, ..., x_n), x_i \in \mathbb{R}$$

un modelis ir normālais sadalījums. Normālajam sadalījumam vispārīgā gadījumā ir divi parametri - σ, μ - taču piemērā datu kopa ir standartizēta un pieņemts $\sigma = \overline{\sigma} = 1$, atstājot brīvu tikai parametru $\theta = \mu$. Tāpat kā iepriekš, izteikta modelim atbilstošā ticamības funkcija $L(\theta|\vec{x})$, pārvedota logaritmiskajā versijā $l(\theta|\vec{x})$ un atrasts funkcijas ekstrēms (jāpiemin, ka šajā gadījumā ekstrēms tik tiešām atrasts funkcijai ar visu datu kopas vektoru, nevis reducētu skalāru argumentu). Atkal izrādās, ka maksimālai ticamībai $\theta = \mu = \overline{x}$.

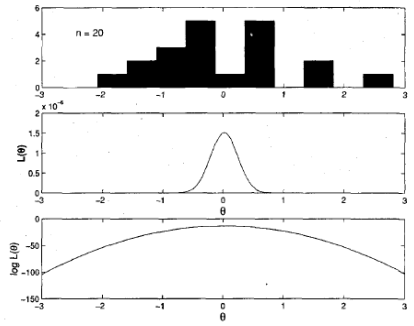

Figure 4.3 The likelihood as a function of θ for a sample of 20 data points from a Normal density with a true mean of 0 and a known standard deviation of 1: (a) a histogram of 20 data points generated from the true model (top), (b) the likelihood function for θ (center), and (c) the log-likelihood function for θ (bottom).

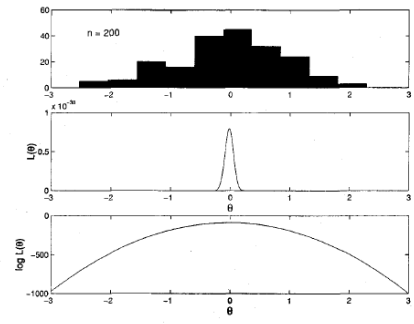

Figure 4.4 The likelihood function for the same model as in figure 4.3 but with 200 data points: (a) a histogram of the 200 data points generated from the true model (top), (b) the likelihood function for θ (center), and (c) the log-likelihood function for θ (bottom).

Tāpat kā binomiālā sadalījuma gadījumā, lielākām datu kopām lielu θ nobīžu ticamība ir zemāka, nekā mazām.

Uzdevums 2

Piebilde par kursa materiālu “Parametriskie modeļi: estimatori, MLE, Beijess”, 6. lpp., piemērā $d = \Theta k + q$ ar normāli sadalītiem parametriem Θ, σ: parametra σ vidējā vērtība uzdota vispirms tekstā kā n (kas jau tiek izmantots kā datu kopas izmērs), pēc tam bez paskaidrojuma formulā nomainīta uz k (kas, gan ar indeksiem, jau tiek izmantots kā nobraukto kilometru skaits modelī). Ieteikums nomainīt gan formulā, gan tekstā uz kādu citu burtu, jo beigās iegūtais vienādojums ar diviem dažādiem k var būt mulsinošs.

a. Beijesa estimators, binomiāls modelis, beta a priori

Šajā piemērā par pamatu ņemts 4.4 paraugs ar tāpat pieņemtu pēc θ parametrizētu binomiālu sadalījumu kā modeli, no kura formulas, bez liekām ceremonijām izmetot datu kopai konstantos faktoriāļus, atkal iegūta tā pati ticamības funkcija skalāram argumentam $L(\theta|r) = \theta^r(1-\theta)^{n-r}$, tikai parametrizēta ortogonāli - $p(r|\theta) \propto L(r|\theta) = L(\theta|r)$. Beta sadalījumam bez komentāriem izmesta katriem α, β nemainīgā daļa ar gamma funkcijām un iegūta ļoti līdzīga proporcionalitātes sakarība $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, ko, ja vēlētos, varētu citādi izteikt, piemēram, kā $L_0(\theta|\alpha,\beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Tad pēc dažām algebriskām manipulācijām ticamības novērtējumu pēc Beijesa metodes izsaka kā

$$L_{Bayes}(\theta|r) = L(r|\theta)L_0(\theta|\alpha,\beta) = \theta^{\alpha+r-1}(1-\theta)^{\beta+n-r-1}$$

kas faktiski ir tāda pati izteiksme, kādu iegūtu, izmantojot nevis beta sadalījumu, bet pielietojot MLE metodi binomiālam sadalījumam, apvienojot reālos datus ar fiktīvu datu kopu, kuras izmērs $n = \alpha + \beta - 2$un $|\{x_i = 1\}| = \alpha - 1$. Tātad, ja radusies intuitīva izpratne par MLE metodi un tās pielietojumu binomiālajam sadalījumam, šo izpratni var vispārināt uz Beijesa estimatoru ar beta a-priori sadalījumu.

b. Beijesa estimators, normālais sadalījums abos

4.11. piemērā atkal pieņemts, ka datu kopa ir izlase no $N(\theta, \alpha), \alpha = const.$, un tiek meklēta brīvā parametra θ vērtība, kas reizē vislabāk atbilst datu kopai un iepriekš pieņemtajam (*a priori*) sadalījumam $\theta \sim N(\theta_0, \alpha_0); \theta_0, \alpha_0 = const.$. Pēc tās pašas $p(\theta|\vec{x}) = p(\vec{x}|\theta)p(\theta)$ metodes tiek reizināti abi sadalījumi un, dažus algebriskus trikus vēlāk, iegūts *a posteriori* sadalījums $\theta \sim N(\theta_1, \alpha_1)$, kur

$$\theta = \frac{\alpha_0^{-1}\theta_0 + \overline{x}n\alpha^{-1}}{\alpha_0^{-1} + n\alpha^{-1}}$$

un attiecīgi, tāpat kā piemērā lekciju materiālos,

$$\lim_{\alpha_0 \to 0}(\theta) = \theta_0$$

$$\lim_{\alpha_0 \to \infty}(\theta) = \overline{x}$$

$$\lim_{n \to 0}(\theta) = \theta_0$$

$$\lim_{n \to \infty}(\theta) = \overline{x}$$

Neformāli to var izteikt sekojoši: *a priori* sadalījums dominē rezultātā pie mazām *a priori* sadalījuma dispersijām un pārāk mazām datu kopām. Datu kopa dominē, ja *a priori* dispersija ir liela vai datu kopa ir pietiekami liela.