Datu statiskā apstrāde konspekts

Jānis Erdmanis

2013. gada 22. janvāris

Varbūtība

- Novērtējums notikumam notikt pie atbiltošiem apstākļiem
- Raksturo skaitlis robežās $[0, 1]$
- Simetriskiem notikumiem jeb A un $\bar{A}$, ja abi ir izslēdzoši varbūtība novērtējama kā 0.5
- Tiek novērtēta ar frekvences definīciju $P(A) = \lim \frac{n}{N}$
  - Nevar nodrošināt bezg. skaitu eksperimentu
- Varbūtības aksiomas
  - $P(A) >= 0$
  - Notikt jebkuram notikumam - $P(E) = 1$
  - Ja notikumi ir savstarpēji izslēdzoši - $P(A + B) = P(A) + P(B)$
- Nosacījuma varbūtība - $P(A\,B) = P(A)\,P(B|A)$
- Varbūtība notikt notikumam A ar īpašību B jeb varbūtību summa zinot varbūtības notikt A_i notikumam un katram A_i notikumam varbūtību uzrādīt īpašību B:

$$P(B) = \sum P(A_i)\,P(B|A_i)$$

- Neatkarīgi notikumi - $P(A|B) = P(A)$

Sadalījuma kumultatīvā un blīvuma funkcija

- $F(x) = P(\xi \leq x)$
- $f(x) = P(x \leq \xi \leq x + dx)$
- $P(a \leq \xi \leq b) = \int_a^b f(x)dx$

Raksturojošie parametri

- Matemātiskā cerība - $E(\xi) = \sum x_i P(\xi = x)$
- Momenti vispārīgi - $\mu_l = E\{(\xi - \hat{x})^l\}$
- Ja l=2, tad tā ir dispersija. Var interpretēt kā inerces momentu
- Reducētais mainīgais - $u = \frac{\xi - \hat{x}}{\sigma(\xi)}$
- Visvarbūtīgākā vērtība
- Mediāna - $F(x_{0.5}) = P(\xi < x_{0.5}) = 0.5$
- Kvantila - $x_q \to q$, ja $F(x_q) = q$

Vairāku mainīgo sadalījumu, momenti.

- $H(\xi, \upsilon) = (\xi - a)^l(\upsilon - b)^m$
- Ja l vai $m = 2$, tad dispersija
- Ja $l, m = 1$, tad kovariācija
- Kovariācijas matrica - $C = E\{(\vec{\xi} - \hat{\vec{x}})(\vec{\xi} - \hat{\vec{x}})^T\}$, kur $E\{\ \}$ iedarbojās kā lineārs operators!
- Korelācijas koeficients - $\rho(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$
- Korelācijas koeficienta nozīmība parādās, ja izsaka - $\sigma^2(u \pm v) = 2(1 \pm \rho(u, v))$

Mainīgo transformācija

- Ja abi sadalījumi raksturo vienu notikumu - $g(y)dy = f(x)dx$
- Analogi - $g(y, \dots) = |\frac{\partial(y,\dots)}{\partial(u,\dots)}|f(u, \dots)$
- Lineāra transformācija - $\vec{\upsilon} = T\vec{\xi} + \vec{a}$
- $C_y = E\{(\vec{y} - \hat{\vec{y}})(\vec{y} - \hat{\vec{y}})^T\} = E\{(T\vec{x} + T\vec{a} - T\hat{\vec{x}} - T\vec{a})(T\vec{x} + \vec{a} - T\hat{\vec{x}} - T\vec{a})^T\} = E\{T(\vec{x} - \hat{\vec{x}})(T(\vec{x} - \hat{\vec{x}}))^T\} = TC_xT^T$
- Ja nepieciešams iegūt transformētā mainīga $\vec{y}$ kovariācijas matricu, tad izvirza $\vec{y} = \vec{f}(\vec{x})$ Teilora rindā:

$$\vec{y} = \vec{f}(\vec{x_0}) + \left.\frac{\partial(\vec{f})}{\partial(\vec{x})}\right|_{\vec{x}=\vec{x_0}} (\vec{x} - \vec{x_0}) = \vec{f}(\vec{x_0}) + \nabla\vec{f}|_{\vec{x}=\vec{x}_0}(\vec{x} - \vec{x_0})$$

Nepārtraukti sadalījumi

- Sadalījuma raksturīgā funkcija - $\phi(t) = E\{e^{itx}\}$, kurai īpašība $\frac{\partial^n \phi}{\partial t^n} = i^n \lambda_n$
- Standarta normālais sadalījums(paredzēts reducētajam mainīgajam!):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

  - Var izmantot formā - $gausa\_sad = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$
- Standarta eksponenciālais sadalījums - $f(x) = e^{-x}$
- Košī sadalījums apraksta kā daļiņas izkārtojās uz ekrāna, ja tās izšautas no punktveida avota ar vienmērīgi sadalītu varbūtību

Diskrētie sadalījumi

- Binomiālais apraksta varbūtību notikt notikumam A n reizes no k mēģinājumiem, ja $P(A) + P(\bar{A}) = 1$
  - Izvēlās n unikālas neatklātas monētas no k
  - Nosaka varbūtību izlasei, kurai visas n monētas ir ar notikumiem A, bet $(k - n)$ monētas ar notikumiem $\hat{A}$
  - Ja p apraksta varbūtību notikt pozitīvam notikumam jeb A, novērtējot no mērijuma $s^2\{S(p)\} = \frac{\kappa}{n^2}\left(1 - \frac{\kappa}{n}\right)$ ir iespējams novērtēt novērtēto pozitīvo notikumu skaitu κ, kā:

$$\Delta\kappa = \sqrt{s^2\{S(p)\}n^2} \approx^{n \gg \kappa} \sqrt{\kappa}$$

- Puasona tiek iegūts $\lambda = np$ tiek saglabāts konstants, bet $n \to \infty$:

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Hiperģeometriskais - varbūtība, ka no n izvilktām bumbiņām k ir baltas, bet $(n - k)$ melnas, pastāvot galīgam skaitam izvelkamo bumbiņu.
  - Ja $n \ll N$, tad iegūst Binomiālo

Izlases

- Eksperimentā tiek iegūta gadījuma rakstura mainīgo kopa, ko sauc par *izlasi*. To pieņemts apzīmēt ar $\vec{\xi} = (\xi_1, \ldots, \xi_n)$
- Ģenerālkopa - visas iespējamās izlases
- Izlasei definē arī varbūtības blīvumu, kas raksturo varbūtību no ģenerālkopas *izvilkt* šādus mērijumus - $g(\vec{x})$
- Ja mērijumi ir savstarpēji neatkarīgi - $g(\vec{x}) = \prod g_i(x_i)$
- Izlases sadalījuma funkcija - $W_n(x) = \frac{n[x>x_i]}{n}$

Novērtējumi

- Parametri, kuri iegūti no izlases bet kuri raksturo ģenerālkopu tiek saukti par novērtējumiem
- Novērtējums $\hat{a}$ ir labs:
  - ja ir konverģējošs:
$$\lim_{N\to\infty} \hat{a} = a$$
  - nenovirzīts - $E\{\hat{a}\} = a$
- Matemātiskā cerība:
  - $\bar{\xi} = \frac{1}{n} \sum \xi_i$
  - $\sigma^2(\bar{\xi}) = \frac{1}{n}\sigma^2(\xi)$, kas saista vidējās vērtības novērtējuma dispersiju ar sadalījuma mainīgā ξ dispersiju!
- Sadalījuma $g_1(\xi_1) = \ldots = g_n(\xi_n) = f(\xi)$ dispersijas novērtējums:
  - $s^2 = \frac{1}{n-1} \sum (\xi_i - \bar{\xi})^2$
  - $\Delta s^2 = s^2 \sqrt{\frac{2}{n-1}}$
  - kovariāciju novērtē analogi dispersijai

Maksimālās ticamības metode

- Izlases varbūtības funkcija:
  - Zināms sadalījums, kuram pakļaujās ξ_i
  - Zināmi parametri $\vec{\lambda}$, kuri definē sadalījumu
  - Tad var definēt varbūtību *mazā apkārtnē* iegūt tieši šādu izlasi:
$$L(\vec{\xi}|\vec{\lambda}) = \prod f(\xi_i|\vec{\lambda})$$
- Mērķis noskaidrot $\vec{\lambda}$, lai varbūtība šādai *izlasei* no *ģenerālkopas* būtu maksimāla!

Izlases informācija

- Vispārīgi novērtējums S ir nobīdīts: $B(\lambda) = E(S) - \lambda$
- Lai novērtējums būtu *labs*, novērtētā parametra $\sigma^2(S)$ jābūt pēc iespējas mazākai, bet šie abi lielumi ir savstarpēji saistītīti ar *informācijas nevienādību*:
$$I(\lambda) = E\left\{ \left(\sum_{i=1}^{N} \frac{f'(x_i|\lambda)}{f(x_i|\lambda)} \right)^2 \right\} = N\, E\left\{ \left(\frac{f'(x|\lambda)}{f(x|\lambda)} \right)^2 \right\} = E(\ell_\lambda'^2) = -E(\ell'')$$
$$\sigma^2(S) \geq \frac{1 + B_\lambda'(\lambda)}{I(\lambda)}$$
$$\ell_\lambda' = A(\lambda)(S - E(S))$$
  - Var parādīt, ka mazākā dispersija novērtētājam S ir tad, ja izlases varbūtība ir formā - $L = d\, e^{B(\lambda)S + C(\lambda)}$

Mazākā kvadrātu metode

- Tiek uzskatīts, ka mērijums $\vec{y}$ ir formā - $\vec{y} = T\vec{\lambda} + \vec{a_0} + \vec{\varepsilon}$
  - Kur katrs ε_i ir pēc *Centrālās robežteorēmas* ir ar *Gausa sadalījumu*
  - T ir transformācijas matrica jeb Teilora rindas otrais loceklis, kas aprēķināts pie $\vec{t}$:

$$y_i = \nabla_\lambda f(t_i|\vec{\lambda})\vec{\lambda} + a_i + \varepsilon_i$$

  - $\vec{\lambda}$ ir novērtējums parametriem no izlases - $\{(t_1, y_1, \sigma_1), \ldots, (t_n, y_n, \sigma_n)\}$

- Tiek meklēti tadi parametri λ, lai varbūtība šādai izlasei būtu vislielākā. Zinot $\vec{\varepsilon}$ sadalījumu problēma uzrakstāma kā:

$$L = \prod f(\varepsilon_i, \sigma_i|\vec{\lambda}) = \prod \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$$

- No iepriekšējās izteiksmes iegūst, ka jāmeklē minimums - $\chi^2(\vec{\lambda}) = \sum \frac{\varepsilon_i^2}{\sigma_i^2}$
  - Summu, ja $\vec{\lambda} \nrightarrow T$, var minimizēt pēc šadas shēmas:
    - Pārrakstam problēmu, ja $C_\varepsilon^{-1} = H^T H$, $H\vec{y} = \vec{c}'$, $T' = HT$, formā:

$$\chi^2(\vec{\lambda}) = \sum \frac{\varepsilon_i^2}{\sigma_i^2} = \vec{\varepsilon_j}^T C_\varepsilon^{-1} \vec{\varepsilon} = [H(\vec{y} - T\vec{x})]^T H(\vec{y} - T\vec{x}) = \|H(\vec{y} - T\vec{x})\|^2 = \|\vec{c}' - T'\vec{x}\|^2$$

    - Ja mērijumi ir 3, tad izteiksme ir attālums telpā starp diviem punktiem. Ar brīvajām $\vec{x}$ koordinātēm tiek definēta telpa, kurā konkrētie mērijumi attēlojās kā punkts. Šajā telpā ir definējams $\vec{c}'$, kas veido tajā apakštelpu, kas 3. mērijumiem varētu būt plakne vai taisne. Mazākais attēlums starp šādiem objektiem būs tad, ja *mērijumu punkts* projicēsies uz $\vec{c}'$ apakštelpu. Varam $\vec{c}' = \vec{c}' - T'\vec{x} + \vec{c}'$, tad redzams, ja attālums līdz apakštelpai ir minimāls, tad $\vec{c}' - T'\vec{x}$ ir ortognāls visiem vektoriem, ko veido $T'\vec{x}$.
    - No iepriekšējā sprieduma risinām - $T'\vec{x} \cdot (\vec{c}' - T'\vec{x}) = 0$
    - Piereizinot ar T'^T un atrisinot LNHVS iegūst:

$$\vec{x} = (T^T C_\varepsilon^{-1} T)^{-1} C_\varepsilon^{-1} \vec{y}$$

    - Dispersijas varam novērtēt, ja pieņemam $(T^T C_\varepsilon T)^{-1} C_\varepsilon^{-1}$ par transformācijas matricu no mainīgā $\vec{\varepsilon}$ uz $\vec{x}$:

$$C_x = (T^T C_\varepsilon^{-1} T)^{-1}$$

Slīdošais vidējais

- Ja $t_i - t_{i-1} = \Delta t = const$, tad par laikrindu sauc - $\upsilon_i = y_i(t_i) + \varepsilon$
- Vienkāršākais veids kā aprakstīt tendenci ir kā nesvērto vidējo vērtību, kas apraksta tendenci tad, ja $y = \alpha + \beta\, t$
- Vispārīgā gadījumā veic regresiju ar polinomu ap izvēlētu apkārtni. Lai procesu atvieglotu tiek izmantots, ka argumenta ass vērtības var uzskatīt par veseliem skaitļiem.

Statistikas testi

- Testa procedūra:
  - Formulē noliedzamo hipotēzi H_0
  - Izvēlās nozīmības kritēriju
  - Atkarībā no H_0 tiek izvēlēts statistikas tests
  - Aprēķina testa statistiku
  - Nosaka varbūtību šādai statistikai attiecībā pret nozīmības kritēriju
- Testa statistikas:

- $X^2 = \sum u_i^2$, kur $u_i = \frac{y_i - f(x_i)}{\sigma_i}$
- Fišera - $\frac{s_1^2}{s_2^2} < F_{1-\alpha/2}(df_1, df_2)$
- Stjudenta(vidējai vērtība) - $u < t_{1-\alpha/2}$
- Stjudenta(divām izlasēm):
  * Konstruē lielumu - $\Delta = \bar{x} - \bar{y}$
  * Nosaka šī lieluma dispersiju - $\sigma^2(\Delta) = \sigma^2(\bar{x}) + \sigma^2(\bar{y})$
  * Pieņem, ka $\bar{x}$ un $\bar{y}$ nāk no viena sadalījuma un iegūst - $s^2 = \frac{(N_1-1)s_x^2 + (N_2-1)s_y^2}{(N_1-1)+(N_2-1)}$
  * Novērtē dispersiju zinot to ģenerālkopai - $s_\Delta^2 = s_{\bar{x}}^2 + s_{\bar{y}}^2 = \frac{N_1+N_2}{N_1 N_2} s^2$
  * Konstruē statistiku - $\frac{|\bar{x}-\bar{y}|}{s_\Delta} > t_{1-\alpha/2}$

χ^2 sadalījums

- Ja pieņem, ka kļūdu novirzes pakļaujās Gausa sadalījumam, tad katram $u_i = \frac{y_i - f(x_i)}{\sigma_i}$ varbūtība uz savas ass - $f(u_i) = e^{-u^2/2}$
- Varbūtības blīvums atrasties kādā telpas apgabalā $du_1 \ldots du_N$:

$$f(u_1, \ldots, u_N) = \prod_{i=1}^{N} e^{-u_i^2/2} = exp(\sum \frac{-u_i^2}{2}) = exp(-\frac{1}{2}\chi^2)$$

- Varbūtība, ka χ būs robežās $[\chi, \chi + d\chi]$ ir proporcionāla sfēras čaulas tilpumam $\chi^{N-1} d\chi$, selko - $P(\chi) = \chi^{N-1} e^{-\frac{1}{2}\chi^2} d\chi$
- Lai iegūtu χ^2 sadalījumu veic transformāciju $P(\chi^2)\, d\chi^2 = P(\chi)\, d\chi$ un iegūto sadalījumu normē:

$$P(\chi^2) \propto \chi^{N-2} e^{-\chi^2/2}$$

- Sadalījuma īpašības:
  - $E(\chi^2) = df$
  - $\sigma^2(\chi^2) = 2df$
  - Ja $df \to \infty$, tad tas tiecās uz normālo sadalījumu