

A sztochasztika alapjai

MBNXK262

11. előadás: Becslések

Kevei Péter

2024/25 tavasz

Maximum likelihood módszer

Példa

Egy dobozban két pénzérme van. Az egyik szabályos, a másik cinkelt, $0,7$ valószínűséggel ad fejet. Az egyik érmével 4-szer dobunk. Az eredmény 3 fej és 1 írás. Vajon melyik érmével dobtunk?

Maximum likelihood módszer

Tekintsünk egy $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mezőt, ahol $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$, ahol $\Theta \subset \mathbb{R}^k$ (általában egy-, néha kétdimenziós). Egy adott $\mathbf{x} = (x_1, \dots, x_n)$ realizáció esetén azt a θ paramétert fogadjuk el, mely mellett a legnagyobb a valószínűsége az adott realizációnak.

Példa

$\xi_1, \dots, \xi_n$ független Bernoulli(θ)-eloszlású.

$$\mathbf{P}_\theta((\xi_1, \dots, \xi_n) = (x_1, \dots, x_n)) =$$

Maximum Likelihood gondolat

$$\mathbf{P}_\theta ((\xi_1, \dots, \xi_n) = (x_1, \dots, x_n)) L_\theta(x_1, \dots, x_n) = \theta^{s_n} (1 - \theta)^{n - s_n},$$

ahol $s_n = \sum_{i=1}^n x_i$.

A θ becslése az a $\hat{\theta}$ érték, melyre

$$L_{\hat{\theta}}(\mathbf{x}) = \sup_{\theta \in [0,1]} L_\theta(\mathbf{x}).$$

Adott s, n értékek mellett keressük a

$$L_\theta = \theta^s(1 - \theta)^{n-s}$$

függvény maximumát a $\theta \in [0, 1]$ intervallumon.

ML általában

Definíció

Ha a háttéreloszlás diszkrét, akkor a *likelihood-függvény*

$$L_{\theta}(\mathbf{x}) = L_{\theta}(x_1, \dots, x_n) = \mathbf{P}_{\theta}(\xi = \mathbf{x}) = \prod_{j=1}^n \mathbf{P}_{\theta}(\xi = x_j),$$

ha folytonos, akkor a *likelihood-függvény*

$$L_{\theta}(\mathbf{x}) = L_{\theta}(x_1, \dots, x_n) = \prod_{j=1}^n f_{\theta}(x_j).$$

A θ paraméter *maximum likelihood becslése* az $\mathbf{x}$ minta alapján

$$\hat{\theta}(\mathbf{x}) = \operatorname{argmax}_{\theta \in \Theta} L_{\theta}(\mathbf{x}).$$

Log-likelihood

$$\ell_{\theta}(\mathbf{x}) = \log L_{\theta}(\mathbf{x}).$$

Hipergeometrikus eloszlás

Egy területen meg akarjuk becsülni az ott élő madarak ismeretlen N számát. Meggyűrűzünk M madarat, majd befogunk n madarat, melyek közül m van meggyűrűzve. Tegyük föl, hogy $m \geq 1$, különben fogjunk még madarat. Megadjuk N ML becslését.

A likelihood függvény

$$L_N(m) = \mathbf{P}_N(\xi = m) =$$

Ez milyen N -re maximális?

Egyszerű számolással

$$\frac{L_{N+1}(m)}{L_N(m)} = \frac{(N+1-m)(N+1-n)}{(N+1)(N+1-M-n+m)} > 1.$$

pontosan akkor teljesül, ha

$$N < \frac{nM}{m} - 1.$$

Ezek szerint az $(L_N(m))_{N \geq M}$ sorozat monoton nő $[nM/m]$ -ig, ahol $[\cdot]$ az egészrészfüggvény. Ezek szerint N ML becslése

$$\hat{N} = \left\lceil \frac{nM}{m} \right\rceil.$$

Exponenciális eloszlás

Egyenletes eloszlás

Momentumok módszere

Tfh $\theta = (\theta_1, \dots, \theta_k)$, azaz $k \geq 1$ paraméterünk van.

Vezessük be a

$$m_j = \mathbf{E}_{\theta}(\xi^j) = g_j(\theta_1, \dots, \theta_k), \quad j = 1, 2, \dots, k,$$

jelölést. Tfh, az első k momentum egyértelműen meghatározza a paramétereket. Ekkor vannak olyan $h_1, \dots, h_k$ függvények (m inverze), hogy

$$h_i(m_1, \dots, m_k) = \theta_i, \quad i = 1, 2, \dots, k.$$

Definíció

A $\theta = (\theta_1, \dots, \theta_k)$ momentum becslése a

$$\hat{\theta}_i = h_i(\hat{m}_1, \dots, \hat{m}_k), \quad i = 1, 2, \dots, k,$$

statisztika, ahol $\hat{m}_i$ az empirikus i -edik momentum, azaz

$$\hat{m}_i = \frac{1}{n} \sum_{j=1}^n x_j^i, \quad i = 1, 2, \dots, k.$$

A nagy számok gyenge törvénye szerint $\lim_{n \rightarrow \infty} \hat{m}_i = m_i$. Innen pedig egyszerűen adódik, hogy $\hat{\theta}_i$ konzisztens becslése θ_i -nek minden i -re.

Poisson-eloszlás

Exponenciális eloszlás

Normális

Lineáris regresszió – motiváció

x ismert bemenet, y ismert kimenet. Kapcsolat $y = f(x) + \text{hiba}$. Feltesszük, hogy a hiba véletlen, várható értéke 0. A hiba eredhet a mérés pontatlanságából (mérési hiba), hozzáadott zajból.

Nem ismerjük az f függvényt, ezt akarjuk meghatározni. Ez a mesterséges intelligencia egyik alapfeladata, a *tanulás*.

Lineáris regresszió

A legegyszerűbb esetet vizsgáljuk, amikor $f(x) = ax + b$ lineáris függvény, ahol $a, b \in \mathbb{R}$ nem ismertek. Ekkor az a, b értékek meghatározása, becslése a feladat.

Adott az (x_i, y_i) , $i = 1, 2, \dots, n$ minta. Nem várunk pontos lineáris illeszkedést, nem lesz olyan a, b , hogy $y_i = ax_i + b$ minden $i = 1, 2, \dots, n$ esetén teljesüljön. Négyzetes hibára minimalizálunk, azaz keressük azt az $(\hat{a}, \hat{b})$ párt, melyre

$$h(a, b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

négyzetes hiba minimális.

$h(a, b)$ pontosan akkor minimális, ha

$$b = \bar{z}_n = \bar{y}_n - a\bar{x}_n.$$

és

$$a = \frac{c_n}{v_n} = \frac{\sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Ez a *legkisebb négyzetek módszere*.

Emlékeztető: elméleti lineáris regresszió

(ξ, η) véletlen vektorváltozó. η függő változó, ξ független változó. Keressük az

$$h(a, b) = \mathbf{E} \left[(\eta - (a\xi + b))^2 \right]$$

függvény minimumhelyét, azaz a legjobb a, b választást.
A legjobb lineáris közelítést a

$$g(x) = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)} (x - \mathbf{E}(\xi)) + \mathbf{E}(\eta)$$

függvény adja. Ő a regressziós egyenes.

Konfidenciaintervallumok

Definíció

A $(T_1(\xi), T_2(\xi))$ statisztikapárral definiált intervallum *legalább* $1 - \varepsilon$ megbízhatósági szintű konfidenciaintervallum a $\psi(\theta)$ paraméterre, ha

$$\mathbf{P}_\theta(T_1(\xi) < \psi(\theta) < T_2(\xi)) \geq 1 - \varepsilon, \quad \forall \theta \in \Theta,$$

ahol $\varepsilon > 0$ előre adott kicsi szám.

Konfidenciaintervallum normális eloszlás várható értékére ismert szórás esetén

Legyenek $\xi_1, \dots, \xi_n \sim N(\mu, \sigma_0^2)$ független véletlen változók, ahol μ az ismeretlen paraméter, σ_0^2 ismert. Megadunk μ -re pontosan $1 - \varepsilon$ megbízhatósági szintű konfidenciaintervallumot.

Konfidenciaintervallum

Tétel

Ha ξ, η független normális eloszlású véletlen változók akkor $\xi + \eta$ is normális eloszlású.

Melyik becslés NEM torzítatlan

- ▶ exponenciális eloszlás paraméterének ML becslése
- ▶ Poisson-eloszlás paraméterének momentumbecslése
- ▶ normális eloszlás várható értékének momentumbecslése
- ▶ mintaátlag

Teszt

Poisson-eloszlás paraméterét becsüljük. Az 1, 0, 1, 2 mintához tartozó momentumbecslés

- ▶ 2
- ▶ 4
- ▶ 1
- ▶ más

Teszt

Egy 2 paraméterű exponenciális eloszlásból veszünk 3 elemű mintát. A mintaátlag

- ▶ véletlen mennyiség, várható értéke 3
- ▶ determinisztikus, $1/2$
- ▶ véletlen mennyiség, várható értéke $1/2$
- ▶ egyik sem

Teszt

A madárfogós feladatban mit becsülünk?

- ▶ a befogott madarak számát
- ▶ az egészséges madarak arányát
- ▶ az összes madár számát
- ▶ a beteg madarak arányát