② # Энтропия и дивергенция (дискр. случай)

Пусть P – распр., т.ч. символы $\{a_1 \dots a_k\}$ имеют вероятности $\{p_1 \dots p_k\}$

$$H(P) = -\sum_{j=1}^{k} p_j \log p_j \quad \text{– энтропия}$$

Утв. 1) $H(P) \geqslant 0$ $\quad = 0 \Leftrightarrow \exists j: p_j = 1$ $\quad (0 \log 0 = 0)$

2) $H(P) \leqslant \log k$ $\quad = \log k \Leftrightarrow \forall j \ p_j = 1/k$

▲ 1) следует из того, что $-\log p_j \geqslant 0$

2) $$H(P) = E \log \frac{1}{p(\xi)} \leqslant \log E \frac{1}{p(\xi)} = \log \sum_{j=1}^{k} p_j \frac{1}{p_j} = \log k$$

где $\xi \sim P$ $\quad$ ↑ по нер-ву Йенсена

□

Общий случай

P, Q – доминир.

$$H(P) = -E \log p(\xi), \text{ где } \xi \sim P$$ энтропия

$$H(P, Q) = -E \log q(\xi), \text{ где } \xi \sim P$$ кросс-энтропия

$$KL(P, Q) = H(P, Q) - H(P) = E \log \frac{p(\xi)}{q(\xi)}, \text{ где } \xi \sim P$$ дивергенция

Пример $\quad P = U(0, 1/2) \qquad p(x) = 2\, I\{x \in (0, 1/2)\}$

$$H(P) = -\log 2 < 0$$

Свойства KL :

1) $KL(P, Q) \geq 0 \qquad = 0 \Leftrightarrow P = Q$

$\blacktriangle \quad -KL(P, Q) = E \log \dfrac{q(\xi)}{p(\xi)} \underset{\text{Йенсен}}{\leq} \log E \dfrac{q(\xi)}{p(\xi)} = \log \int p(x) \dfrac{q(x)}{p(x)} dx =$

$= \log 1 = 0$

т.к. log − строго выпуклый, то равенство достигается при $P = Q$

2) $KL(P, Q) \neq KL(Q, P)$

3) Пусть $X_1 \ldots X_n$ − выборка из $P_\theta \in \{ P_\theta \mid \theta \in \Theta \}$ − дискретное

$KL(\hat{P}_n, P_\theta) = E_{\hat{P}_n} \log \dfrac{\hat{p}_n(X)}{p_\theta(X)} = \dfrac{1}{n} \sum_{i=1}^{n} \log \dfrac{1/n}{p_\theta(X_i)} =$

дискр. нл-ти

$$= -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i) - \log n = -\frac{1}{n} l_x(\theta) - \log n$$

Вывод: $KL(\hat{P}_n, P_\theta) \to \min_{P_\theta} \iff l_x(\theta) \to \max_\theta$

Свойства ОМП

Теорема Экстремальное свойство правдоподобия [L1 – L3]

$$\forall \theta_0, \theta_1 \in \Theta \quad \text{т.ч.} \quad \theta_0 \neq \theta_1 \qquad P_{\theta_0}\left(L_X(\theta_0) > L_X(\theta_1)\right) \to 1 \quad \text{при} \quad n \to \infty$$

$\blacktriangle$ $$L_X(\theta_0) > L_X(\theta_1) \iff \frac{L_X(\theta_0)}{L_X(\theta_1)} > 1 \iff \frac{1}{n}\log\frac{L_X(\theta_0)}{L_X(\theta_1)} > 0$$

$$\frac{1}{n}\log\frac{L_X(\theta_0)}{L_X(\theta_1)} = \frac{1}{n}\log\prod_{i=1}^{n}\frac{p_{\theta_0}(x_i)}{p_{\theta_1}(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\theta_0}(x_i)}{p_{\theta_1}(x_i)} \xrightarrow[364]{P_{\theta_0}\text{-п.н.}}$$

$$\to E_{\theta_0}\log\frac{p_{\theta_0}(x_i)}{p_{\theta_1}(x_i)} = KL\left(P_{\theta_0}, P_{\theta_1}\right) > 0$$

$\neq$ т.к. $\theta_0 \neq \theta_1$ и L1

Получаем $$P_{\theta_0}\left(L_X(\theta_0) > L_X(\theta_1)\right) \to P\left(KL\left(P_{\theta_0}, P_{\theta_1}\right) > 0\right) = 1$$

Следствие: Если $|\Theta| < \infty$, то ОМП существует и явл. сост. оц.

[L1 - L3]

Δ Пусть θ_0 — истинное знач парам.

Тогда $P_{\theta_0}(\forall \theta_1 \neq \theta_0 : L_x(\theta_0) > L_x(\theta_1)) =$

$$= P_{\theta_0}\Big(\bigcap_{k=1}^{|\Theta|} L_x(\theta_0) > L_x(\theta_k)\Big) \to 1$$

← конечное

Теорема С вер-тью $\to 1$ ур-е $\dfrac{\partial L_x(\theta)}{\partial \theta} = 0$ имеет корень, являющийся сост. оц. θ

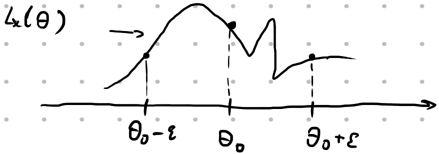
[L1 - L5]

▲ Пусть θ_0 – истинное значение параметра

$L4 \Rightarrow \exists \varepsilon > 0$ т.ч. $(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \Theta$

По т. об экст. св-ве:

$$P_{\theta_0}\left(L_x(\theta_0) > L_x(\theta_0 - \varepsilon),\ L_x(\theta_0) > L_x(\theta_0 + \varepsilon) \right) \to 1$$



с вер $\to 1$

$\Rightarrow$ с вер $\to 1$ на $(\theta_0-\varepsilon,\ \theta_0+\varepsilon)$ имеется корень ур-я правдоподобия.

Пусть $\tilde{\theta}_0$ — ближайший к θ_0 корень ур-я (не факт, что $\tilde{\theta}_0 \in (\theta_0-\varepsilon;\ \theta_0+\varepsilon)$)

Тогда $P(|\tilde{\theta}-\theta_0| > \varepsilon) \to 0$

Т.к. это выполнено $\forall \varepsilon > 0$, то получаем, что

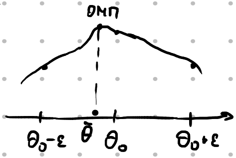
$$\tilde{\theta} \xrightarrow{P_{\theta_0}} \theta_0 \Rightarrow \tilde{\theta} \text{ — сост. оц. } \theta \qquad \square$$

Следствие [L1 – L5] Если $\forall n$ $\forall X_1 \ldots X_n$ $\dfrac{\partial l_X(\theta)}{\partial \theta} = 0$ имеет единственный корень, то $\tilde{\theta}$ — сост. оц. θ

$$P(\tilde{\theta} = \hat{\theta}_{ОМП}) = 1 \Rightarrow \hat{\theta}_{ОМП} \text{ — сост.}$$

▲

$P_{\theta_0}\Bigg($



$\Bigg) \to 1$

Теорема [L1 – L9]

1) $\tilde{\theta}$ – сост. оц. θ, являющаяся решением ур-я правдоподобия

Тогда $\tilde{\theta}$ – а.н.о. θ с а.д. $i^{-1}(\theta)$

В частности, если $\tilde{\theta}$ – ед. реш. ур-я правдоподобия,

Тогда $\tilde{\theta}$ – а.н.о. θ с а.д. $i^{-1}(\theta)$

2) Пусть $\hat{\theta}$ – а.н.о. с а.д. $\sigma^2(\theta)$, т.ч. $\sigma^2(\theta)$ непр. по θ

Тогда $\sigma^2(\theta) \geq i^{-1}(\theta)$ (б/д)

Следствие [L1 – L9] ОМП – сост., а.н.о., асимптотически эффективно (т.е. имеет наим. а.д. среди всех а.н.о. с непр. дисп.)

Примечание если L1–L9 нету, то можно получить более крутые св-ва.

$$X_1 \ldots X_n \sim U[0, \theta]$$

$$n(\theta - X_{(n)}) \xrightarrow{d_\theta} Exp(1)$$ ← это круче ас. норм

$$\left(X_{(n)} \pm c/\underline{\underline{n}}\right)$$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N$$

$$\left(\hat{\theta} \pm c/\underline{\underline{\sqrt{n}}}\right)$$

Док-во п. 1)

Вклад $U_x(\theta) = \dfrac{\partial L_x(\theta)}{\partial\theta}$

разложим по Тейлору $U_x(\tilde{\theta})$ в точке θ: (в форме Лагранжа)

$$U_x(\tilde{\theta}) = U_x(\theta) + U_x'(\theta)(\tilde{\theta}-\theta) + \frac{1}{2}U_x''(\theta^*)(\tilde{\theta}-\theta)^2$$

где θ^* лежит между $\tilde{\theta}$ и θ

следствие: $\theta^* \xrightarrow{P_\theta} \theta$ [θ^* не сост. оц., т.к. не факт, что оц., т.к. может зависеть от θ]

т.к. $\tilde{\theta} \xrightarrow{P_\theta} \theta$

$U_x(\tilde{\theta}) = 0$, т.к. $\tilde{\theta}$ — решение ур-я $U_x(\theta) = 0$

$$\Rightarrow \quad -U_x(\theta) = (\tilde{\theta}-\theta)\left(U_x'(\theta) + \frac{1}{2}U_x''(\theta^*)(\tilde{\theta}-\theta)\right)$$

$$\frac{-\sqrt{n}\; U_x(\theta)}{U_x'(\theta) + \frac{1}{2} U_x''(\theta^*)(\tilde{\theta}-\theta)} = \sqrt{n}\,(\tilde{\theta}-\theta)$$

$$\|$$

$$\frac{\overbrace{-\sqrt{n}\; U_x(\theta)\; 1/n}^{\to N(0,\, i(\theta))}}{\underbrace{\frac{1}{n} U_x'(\theta)}_{\to i(\theta)} + \frac{1}{2n} \underbrace{U_x''(\theta^*)}_{<\infty} \underbrace{(\tilde{\theta}-\theta)}_{\to 0}} \to N(0,\, i^{-1}(\theta))$$

1) $$\sqrt{n}\,\frac{1}{n} U_x(\theta) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U_{x_i}(\theta) - \overbrace{E_\theta U_{x_1}(\theta)}^{0}\right) \xrightarrow[\text{ЦПТ}]{d_\theta} N\big(0,\, \underbrace{D_\theta U_{x_1}(\theta)}_{i(\theta)}\big)$$

2) $$\frac{1}{n} U_x'(\theta) = \frac{1}{n}\sum_{i=1}^{n} U_{x_i}'(\theta) \xrightarrow[\text{УЗБЧ}]{P_\theta} E_\theta U_{x_1}'(\theta) = -i(\theta)$$

3) $$\hat{\theta} - \theta \xrightarrow{P_\theta} 0 \quad \Rightarrow \quad \text{хотим} \quad \left|\frac{1}{n} U_x''(\theta^*)\right| \le C$$

$$\left|\frac{1}{n}U''_x(\theta^*)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|U''_{x_i}(\theta^*)\right| \leq \frac{1}{n}\sum_{i=1}^{n}H(x_i) \xrightarrow[\text{УЗБЧ}]{P_\theta} E_\theta H(x_1) < +\infty$$

для $\theta^* \in (\theta - c;\ \theta + c)$ по L9

т.к. $\theta^* \xrightarrow{P_\theta} \theta$, то $P_\theta(\theta^* \in (\theta \pm c)) \to 1$

$\Rightarrow$ $\frac{1}{n}U''_x(\theta)$ огр. с вер. $\to 1$

т.е. $\exists\, c(\theta)\ \forall\theta:\ P_\theta\left(\left|\frac{1}{n}U''_x(\theta)\right| \leq c(\theta)\right) \to 1$

2) без док-ва

Натуральный градиент

Пример $X_1 \ldots X_n \sim N(a, s)$ // $s = \sigma^2$

$$L_x(\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log s - \frac{1}{2s}\sum_{i=1}^{n}(x_i - a)^2$$

$$\frac{\partial L_x(\theta)}{\partial a} = -\frac{1}{s}\sum_{i=1}^{n}(a - x_i) = \frac{n}{s}(\bar{x} - a)$$

$$\frac{\partial L_x(\theta)}{\partial s} = -\frac{n}{2s} + \frac{1}{2s^2}\sum_{i=1}^{n}(x_i - a)^2 = \frac{n}{2s^2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - s\right)$$

Градиентный спуск:

$$a_{t+1} = a_t + \eta_1 \frac{\bar{x} - a}{s_t}$$

← ответ для a

$$s_{t+1} = s_t + \eta_2 \frac{1}{s_t^2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - a_t)^2 - s_t\right)$$

← ответ для s

что-то ухудшающее сходимость →

Пусть $f: \mathbb{R}^d \to \mathbb{R}$

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_j}\right)_j \qquad \frac{\partial f(x)}{\partial x_j} = \lim_{\Delta x_j \to 0} \frac{f(x + e_j \Delta x_j) - f(x)}{\Delta x_j}$$

$$\nabla f(x) \sim \underbrace{\underset{\substack{\Delta x_i \\ \|\Delta x\| \leq \varepsilon}}{\operatorname{argmax}} \big(f(x + \Delta x) - f(x)\big)}_{\text{направление наискорейшего роста}}$$

Пусть $P_\theta \in \{ P_\theta \mid \theta \in \Theta \}$

$f(P_\theta)$ – функционал, напр $L_x(\theta)$

Проблема взятия производной:

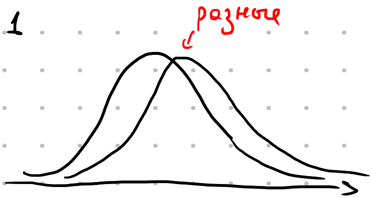
окрестность $\|\Delta\theta\| \le \varepsilon$ не отражает схожесть распределений

(на проблему выше)

Пример 1) $N(\theta_1, \theta_2)$

$\theta = (0, 1)$ $\Delta\theta = (1, 0)$ $\|\Delta\theta\| = 1$

$P_\theta = N(0, 1)$ $P_{\theta+\Delta\theta} = N(1, 1)$



2) $N(\theta_1, \theta_2)$ $\theta = (0, 100)$ $\Delta\theta = (1, 0)$ $\|\Delta\theta\| = 1$

$P_\theta = N(0, 100)$

$P_{\theta+\Delta\theta} = N(1, 100$

Идея: в опр. grad заменим $\|\Delta\theta\| < \varepsilon$ на окр-ть в пр-ве распр

Опр. Натуральный градиент

$$\nabla_N f(P_\theta) \sim \underset{\substack{\Delta\theta_j \\ KL(P_\theta,\, P_{\theta+\Delta\theta}) < \varepsilon}}{\operatorname{argmax}} \left[f(P_{\theta+\Delta\theta}) - f(P_\theta) \right]$$

Распишем Лагранжиан:

$$L = f(P_{\theta+\Delta\theta}) - f(P_\theta) - \lambda \left(KL(P_\theta, P_{\theta+\Delta\theta}) - \varepsilon \right)$$

$$\nabla_{\Delta\theta} L \Big|_{\Delta\theta=0} = \underbrace{\nabla_{\Delta\theta} f(P_{\theta+\Delta\theta}) \Big|_{\Delta\theta=0}}_{\substack{\nabla_\theta f(P_\theta) \\ \text{т.е. обычный} \\ \text{градиент}}} - \lambda \nabla_{\Delta\theta} KL(P_\theta, P_{\theta+\Delta\theta}) \Big|_{\Delta\theta=0}$$

Утв. При $\|\Delta\theta\| \to 0$

инф. матр. Фишера

$$KL(P_\theta, P_{\theta+\Delta\theta}) = \frac{1}{2}\Delta\theta^T \, i(\theta) \, \Delta\theta + o(\|\Delta\theta\|^2)$$

▲ Разложим KL по Тейлору в $\Delta\theta = 0$

$$KL(P_\theta, P_{\theta+\Delta\theta}) = \underbrace{KL(P_\theta, P_\theta)}_{=0} + \underbrace{\nabla_{\Delta\theta} KL(P_\theta, P_{\theta+\Delta\theta})\Big|_{\Delta\theta=0}^T}_{=0\text{, т.к. это произв. в точке минимума}} \cdot \Delta\theta +$$

$$+ \frac{1}{2}\Delta\theta^T \nabla^2_{\Delta\theta} KL(P_\theta, P_{\theta+\Delta\theta})\Big|_{\Delta\theta=0} \cdot \Delta\theta + o(\|\Delta\theta\|^2)$$

$$\nabla^2_{\Delta\theta} KL(P_\theta, P_{\theta+\Delta\theta})\Big|_{\Delta\theta=0} = \int p_\theta(x) \nabla^2_{\Delta\theta} \log \frac{p_\theta(x)}{p_{\theta+\Delta\theta}(x)}\Big|_{\Delta\theta=0} dx =$$

$$= -\int p_\theta(x) \overbrace{\nabla^2_{\Delta\theta} \log p_{\theta+\Delta\theta}(x)\Big|_{\Delta\theta=0}}^{\nabla^2_\theta \log p_\theta(x)} dx = -E_\theta \nabla^2_\theta \log p_\theta(X_1) = i(\theta)$$

Вернёмся к Лагранжиану

$$\nabla_{\Delta\theta} L = \nabla_\theta f(P_\theta) - \lambda\, i(\theta)\, \Delta\theta = 0$$

$$\Rightarrow \quad \Delta\theta = \underbrace{\frac{1}{\lambda}} \, i^{-1}(\theta)\, \nabla_\theta f(P_\theta)$$

на это можно забить, т.к. натуральный градиент используется в град. спуске $\Rightarrow$ там есть learning-rate

$$\text{Опр} \quad \nabla_N f(p_\theta) = i^{-1}(\theta)\, \nabla_\theta f(P_\theta)$$ — натуральный градиент

Применим к примеру с нормальной выборкой:

Из семинара $i(\theta) = \begin{pmatrix} 1/s & 0 \\ 0 & 1/2s^2 \end{pmatrix} \Rightarrow i^{-1}(\theta) = \begin{pmatrix} s & 0 \\ 0 & 2s^2 \end{pmatrix}$

Расписываем градиентный спуск:

$$a_{t+1} = a_t + \eta_1 (\bar{x} - a_t) = \eta_1 \bar{x} + (1-\eta_1) a_t$$

$$s_{t+1} = s_t + \eta_2 \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - a_t)^2 - s_t \right) = \eta_2 \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - a_t)^2 \right) + (1-\eta_2) s_t$$

Логистическая регрессия

$$\nabla l_y(\theta) = X^T (y - s(\theta)) \qquad I(\theta) = X^T V(\theta) X$$

$$\theta_{t+1} = \theta_t + \eta \left(X^T V(\theta_t) X \right)^{-1} X^T (y - s(\theta_t))$$

IRLS

метод 1ого порядка в пр-ве распределений

метод 2ого " порядка в пр-ве признаков