Лекция 5 (от 30.09)

2.6. Экспоненциальный класс распределений

> **Определение:** Семейство распределений $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ принадлежит *экспоненциальному классу*, если плотность $p_\theta(x)$ имеет вид
>
> $$p_\theta(x) = \frac{g(x)}{h(\theta)} e^{a(\theta)^T u(x)},$$
>
> где $g(x) > 0,\ u(x)$ — произвольные борелевские функции,
> $h(\theta) = \int\limits_{\mathscr{X}} g(x) e^{a(\theta)^T u(x)} dx$ — нормировочная константа.
> Если $a(\theta) = \theta$, будем говорить что *параметризация естественная*.

Пример: $\mathcal{P} = \{\mathcal{N}(a, \sigma^2 | a \in \mathbb{R}, \sigma > 0\}$. Перейдем к естественным параметрам:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{xa}{\sigma^2} - \frac{a^2}{2\sigma^2}\right).$$

Введем параметры $\theta = (\theta_1, \theta_2)$: $\theta_1 = -\dfrac{1}{2\sigma^2},\ \theta_2 = \dfrac{a}{\sigma^2}$.

$$p(x) = \sqrt{-\frac{\theta_1}{\pi}} e^{\theta_1 x^2 + \theta_2 x + \frac{\theta_2^2}{4\theta_1}}.$$

$u(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}, a(\theta) = \theta,\ g(x) = 1,\ h(\theta) = \sqrt{-\dfrac{\theta_1}{\pi}} e^{\frac{\theta_2^2}{4\theta_1}}$.
Найдем достаточные статистики для семейства $\mathcal{P}$:

$$p_\theta(x_1, \ldots, x_n) = h^{-n}(\theta) \prod_{i=1}^n g(x_i) e^{a(\theta)^T \sum\limits_{i=1}^n u(x_i)}.$$

По критерию факторизации Неймана-Фишера $S(X) = \sum u(X_i)$ — достаточная статистика.

Замечание: $S(X)$ — статистика фиксированной размерности.

> **Теорема:** Пусть $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ — семейство распределений т.ч. плотность $p_\theta(x)$ непрерывно дифференцируема по x и носитель не зависит от θ. Пусть также $S(X)$ — достаточная статистика фиксированной размерности m. Тогад семейство $\mathcal{P}$ принадлежит экспоненциальному классу.

Следствие: Если плотность достаточно хорошая, то только семейства из экспоненциального класса допускают сжатие данных с помощью достаточных статистик.

Примеры:

1. $\mathcal{P} = \{\text{Коши со сдвигом}\}$ не лежит в экспоненциальном классе $\implies$ нет достаточных статистик фиксированного размера.
2. $\mathcal{P} = \{U[0, \theta]\}$ — носитель зависит от θ. Однако достаточная статистика фикс. размера существует: $S(X) = X_{(n)}$.

Далее потребуем некоторые условия:

1. Параметризация естественная
2. $g(x),\ u(x)$ непрерывны
3. Условие равномерной сходимости интеграла по параметру:

$$\forall s\ \forall j \leqslant k\ \exists \varphi(x) : \forall \theta \in \Theta\ |g(x) u_s^j(x) e^{\theta u(x)}| \leqslant \varphi(x),$$

и при этом $\int\limits_{\mathscr{X}} \varphi(x) dx$ сходится.

Следствия:

1. $h(\theta)$ непрерывно дифференцируема k раз
2. $p_\theta(x)$ непрерывно дифференцируема k раз по θ
3. Можно менять местами $\frac{\partial}{\partial\theta}$ и $\int$

> **Утверждение 1:**
>
> 1. $$E_\theta u(X) = \nabla \ln h(\theta) = \left(\frac{\partial}{\partial\theta}\ln h(\theta)\right)_j$$
> 2. $$D_\theta u(X) = \nabla^2 \ln h(\theta) = \left(\frac{\partial^2}{\partial\theta^2}\ln h(\theta)\right)_{jk}$$

$\triangle$ (1):

$$\frac{\partial h(\theta)}{\partial\theta_j} = \frac{\partial}{\partial\theta}\int\limits_{\mathscr{X}} g(x)e^{\theta^T u(x)}dx = \{\text{следствие 3}\} = \int\limits_{\mathscr{X}} u_j(x)g(x)e^{\theta^T u(x)}dx = h(\theta)\int\limits_{\mathscr{X}} \frac{u_j(x)}{h(\theta)})g(x)e^{\theta^T u(x)}dx = h(\theta)E_\theta u_j(X_1).$$

$$E_\theta u_j(X_1) = \frac{\partial h(\theta)/\partial\theta_j}{h(\theta)} = \frac{\partial \ln h(\theta)}{\partial\theta} \qquad \square.$$

Утверждение: Если Θ — выпуклое множество, то ОМП существует и единственна.

$\triangle$ $\nabla\nabla \ln h(\theta) = D_\theta u(X_1) \geqslant 0 \implies \ln h(\theta)$ выпукла.

$$l_X(\theta) = \underbrace{\sum \ln g(X_i)}_{\text{не зависит от } \theta} \underbrace{-n\overbrace{\ln h(\theta)}^{\text{выпукла}}}_{\text{вогнута}} + \underbrace{\theta\sum u(X_i)}_{\text{линейна по } \theta} \implies l_X(\theta) \text{ вогнута.}$$

Значит, максимум существует и единственный $\quad\square$.

Утверждение: Если Θ — выпуклое открытое множество, то выполнены условия L5-L9.

$\triangle$ L5-L7 выполнены из следствий 1-3

L8: $\frac{\partial \ln p_\theta(x)}{\partial\theta} = \frac{\partial}{\partial\theta}(\ln g(x) - n\ln h(\theta) + \theta u(x)) = \frac{\partial h(\theta)}{h(\theta)} + u(x)$

$i(\theta) = E_\theta(\frac{\partial \ln p_\theta(X_1)}{\partial\theta})^2$ по утверждению 1 существует и конечна

L9 следует из того, что $\frac{\partial^2 \ln p_\theta(X_1)}{\partial\theta^2}$ не зависит от θ $\quad\square$.

2.7. Сравнение оценок

Ранее было:
$X_1, \ldots, X_n \sim Exp(\theta)$.

$\hat\theta_1 = 1/\overline{X},\ \hat\theta_2 = -\ln\overline{I\{X>1\}}$ — (сильно) состоятельная, а. н. оценка θ. Хотим построить оценку для $\tau(\theta) \in \mathbb{R}^d$.

> **Определение:** Функция $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, которая характеризует степень отклонения оценки от $\tau(\theta)$, называется *функцией потерь (loss function)*.

Примеры:

1. $L(x,y) = (x-y)^2$ — квадратичная функция потерь
2. $L(x,y) = |x-y|$ — абсолютная функция потерь
3. $L(x,y) = \log(1+|x-y|)$
   многомерный случай:
4. $L(x,y) = (x-y)^T A(x-y)$, A — симметричная, полож. определенная матрица
   если $A = I_d : L(x,y) = \sum\limits_{j=1}^{d}(x_j-y_j)^2$.

Пусть $\hat\theta$ — оценка $\tau(\theta)$, θ — истинное значение параметра. Тогда $L(\hat\theta,\theta)$ — штраф при оценивании $\tau(\theta)$ оценкой $\hat\theta$.
Проблема: штраф случаен

> **Определение:** Функция риска
> $$R_{\hat\theta,\tau}(\theta) = E_\theta L(\hat\theta,\tau(\theta)).$$

Примеры:

- $\mathrm{MSE}_{\hat\theta,\tau}(\theta) = E_\theta(\hat\theta-\tau(\theta))^2$ — *среднеквадратичная ошибка*
- $\mathrm{MAE}_{\hat\theta,\tau}(\theta) = E_\theta|\hat\theta-\tau(\theta)|$ — *средняя абсолютная ошибка*

Замечание: если $\tau(\theta) = \theta$, то индекс τ опускаем.

Задача: $X_1, \dots, X_n$ — выборка. $\widehat{\theta}_1 = X_1,\ \widehat{\theta}_2 = \overline{X}$ — оценки $\tau(\theta) = E_\theta(X_1)$. Посчитать MSE.

$\triangle$ $\quad \mathrm{MSE}_{\widehat{\theta}_1,\tau}(\theta) = E_\theta(X_1 - E_\theta X_1)^2 = D_\theta X_1$
$\quad \mathrm{MSE}_{\widehat{\theta}_2,\tau}(\theta) = E_\theta(\overline{X} - E_\theta \overline{X})^2 = D_\theta \overline{X} = \frac{1}{n} D_\theta X_1.$

Вывод: усреднение уменьшает среднеквадратичный риск в n раз. $\square$

Подходы к сравнению оценок:

1. Равномерный

- $\widehat{\theta}_1$ *не хуже* $\widehat{\theta}_2$, если $\forall \theta R_{\widehat{\theta}_1,\tau(\theta)} \leqslant R_{\widehat{\theta}_2,\tau(\theta)}$.
- $\widehat{\theta}_1$ *лучше* $\widehat{\theta}_2$, если, кроме того, $\exists \theta : R_{\widehat{\theta}_1,\tau(\theta)} < R_{\widehat{\theta}_2,\tau(\theta)}$.
- Пусть $\mathscr{K}$ — множество оценок. $\widehat{\theta}$ — *наилучшая в* $\mathscr{K}$, если она лучше всех оценок из $\mathscr{K}$.
- Если $L(x, y) = (x - y)^2$, то подход называется *среднеквадратичным*.

Утверждение: Наилучшей оценки может не существовать.
$\triangle$ $\quad \mathscr{K} = \{\widehat{\theta}_1 \equiv 1, \widehat{\theta}_2 \equiv 2\}$
$\quad \mathrm{MSE}_{\widehat{\theta}_1}(\theta) = E_\theta(\theta - 1)^2 = (\theta - 1)^2$
$\quad \mathrm{MSE}_{\widehat{\theta}_2}(\theta) = E_\theta(\theta - 2)^2 = (\theta - 2)^2$
Если $\theta < 1.5$, то $\mathrm{MSE}_{\widehat{\theta}_1}(\theta) < \mathrm{MSE}_{\widehat{\theta}_2}(\theta)$; если $\theta > 1.5$, то $\mathrm{MSE}_{\widehat{\theta}_2}(\theta) < \mathrm{MSE}_{\widehat{\theta}_1}(\theta)$ $\quad \square$.

> **Утверждение:** Справедливо *bias-variance разложение:*
> $$\underbrace{\mathrm{MSE}_{\widehat{\theta},\tau}(\theta)}_{error} = \underbrace{D_\theta \widehat{\theta}}_{variance} + \underbrace{(E_\theta \widehat{\theta} - \theta)^2}_{bias^2}.$$

$\triangle$ $\quad \mathrm{MSE}_{\widehat{\theta},\tau}(\theta) = E_\theta(\widehat{\theta} - \tau(\theta))^2 = E_\theta((\widehat{\theta} - E_\theta \widehat{\theta}) + (E_\theta \widehat{\theta} - \tau(\theta))^2 = E_\theta(\widehat{\theta} - E_\theta(\widehat{\theta}))^2 + 2E_\theta(\widehat{\theta} - E_\theta \widehat{\theta})(E_\theta \widehat{\theta} - \tau(\theta)) + (E_\theta(\widehat{\theta}) - \tau(\theta))^2$
Второе слагаемое равно нулю, следовательно, получаем требуемое. $\square$

Следствие: Среди все несмещенных оценок наилучшей будет та, у которой меньше дисперсия.

2. Байесовский

Пусть Q — некоторое распределение на Θ. Тогда $\widehat{\theta}_1$ не хуже $\widehat{\theta}_2$, если $E_Q R_{\widehat{\theta}_1}(\theta) \leqslant E_Q R_{\widehat{\theta}_2}(\theta)$.

3. Минимаксный

$\widehat{\theta}_1$ не хуже $\widehat{\theta}_2$, если $\sup\limits_{\theta \in \Theta} R_{\widehat{\theta}_1}(\theta) \leqslant \sup\limits_{\theta \in \Theta} R_{\widehat{\theta}_2}(\theta)$.

4. Асимптотический (для а.н.о)

Пусть $\widehat{\theta}_1, \widehat{\theta}_2$ — а.н.о. $\tau(\theta)$ с асимпт. дисперсией σ_1^2 и σ_2^2. Тогда

- $\widehat{\theta}_1$ не хуже $\widehat{\theta}_2$, если $\sigma_1(\theta) \leqslant \sigma_2(\theta)\ \forall \theta \in \Theta$.
- $\widehat{\theta}_1$ лучше $\widehat{\theta}_2$, если, кроме того $\exists \theta \in \Theta :\ \sigma_1(\theta) < \sigma_2(\theta)$.
- *Относительная асимптотическая эффективность*: $\mathrm{ARE}^\tau_{\widehat{\theta}_1,\widehat{\theta}_2}(\theta) = \dfrac{\sigma_2^2}{\sigma_1^2}$ показывает, насколько $\widehat{\theta}_1$ лучше $\widehat{\theta}_2$.

$\widehat{\theta}_1$ не хуже $\widehat{\theta}_2$, если $\mathrm{ARE}^\tau_{\widehat{\theta}_1,\widehat{\theta}_2}(\theta) \geqslant 1 \forall \theta \in \Theta$.

> **Определение:** Оценка $\widehat{\theta}$ называется *асимптотически эффективной оценкой* $\tau(\theta)$, если она имеет наименьшую асимптотическую дисперсию среди всех а.н.о. $\tau(\theta)$ с непрерывной а. д.

Утверждение: Если выполнены условия L1-L9, то ОМП асимптотически эффективна.

Пример: $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$.

- ОМП: $\widehat{\theta}_1 = \overline{X}$ — а.н.о θ с а.д. $\sigma_1^2 = 1$.
- Теор. о выборочной медиане: $\widehat{\theta}_2 = \widehat{\mu}$ — а.н.о θ с а.д. $\sigma_2^2 = \frac{2\pi}{4} = \frac{\pi}{2}$.

$\mathrm{ARE}_{\overline{X},\widehat{\mu}}(\theta) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)} = \frac{\pi}{2} \approx 1.57.$