ФКН

Департамент больших данных и информационного поиска

Москва 2025

# Лекция 3
# Оценка параметров ММП

Машинное обучение в цифровом продукте

Полякова И.Ю.

# От теории вероятностей к мат.статистике

Данные

Модель данных

$$X_1, \ldots, X_n \xrightarrow{?} P_\theta(X_1, \ldots, X_n)$$

# От теории вероятностей к мат.статистике

$X_1, \ldots, X_n$

**Данные**: СВ с каким-то распределением

$\theta \in R^m$

**Параметры**: делаем предположение, что это распределение задается некоторым набором параметров $\theta$

*Пока работаем в контексте того, что данные – это выборка из одномерной СВ

# ММП (Метод максимального правдоподобия)

*Англ*. Maximum Likelihood (ML)

Или Maximum Likelihood Estimation (MLE)

**Идея:** подобрать такие параметр/ы распределения, при которых вероятность получить наблюдаемую выборку максимальна

# ММП

**Иллюстрация**: важность первого впечатления

Вы зашли в кофейню «Яблоко» и Вы выпили вкусную чашку кофе

**H0:** В этой кофейне всегда вкусный кофе вне зависимости от дня (в-ть выпить вкусную чашку = 1)
**H1:** Конкретно сегодня мне повезло, но в любой другой день кофе был бы невкусным (в-ть выпить вкусную чашку $\rightarrow 0$)

Вернетесь ли второй раз в «Яблоко»?

# ММП: простой пример

$X$, Выборка: 1, 0, 0, 2

$X_i$ - независимы

| $X$ | 0 | 1 | 2 |
|---|---|---|---|
| $p$ | $p$ | $2p$ | $1-3p$ |

$p_{ML}$ - ?

# ММП: простой пример

$$P(X_1=1, X_2=0, X_3=0, X_4=2) =$$

$$= P(X_1=1)\cdot P(X_2=0)\cdot P(X_3=0)\cdot P(X_4=2) =$$

$$= 2p\cdot p\cdot p\cdot(1-3p) = 2p^3 - 6p^4 \to \max_p$$

$$FOC: 6p^2 - 24p^3 = 0$$

$$6p^2(1-4p) = 0$$

$$\left[\begin{array}{l} p=0 \\ p=1/4 \end{array}\right.$$

$$P(p=0) = 0$$

$$\wedge$$

$$P(p=1/4) = 2/128$$

# ММП в общем случае

- Для дискретной СВ

$$L(\boldsymbol{\theta}) = P(X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\theta}) = \prod_{i=1}^{n} P(X_i = x_i | \boldsymbol{\theta})$$

*конечно, при условии, что иксы – независимо и одинаково распределенные СВ

# ММП в общем случае

- Для дискретной СВ
- Для простоты расчетов (вместо произведения – сумма) обычно берут логарифм функции правдоподобия

$$\boldsymbol{ln(L(\theta)) = \sum_{i=1}^{n} ln(P(X_i = x_i|\theta))}$$

# Пример: распределение Бернулли

- Каждый день Вы приходите в кофейню «Яблоко» и берете чашку кофе: в каждый день она либо хорошая, либо плохая
- Вы документируете качество кофе нулем или единицей в каждый день
- Вы уверены, что качество не зависит от дня
- При этом кофе с вероятностью p является хорошим, с вероятностью (1-p) плохим
- Оценим параметр p

*Если захочется повторить выкладки, то проверить себя можно здесь: https://habr.com/ru/articles/830326/

# ММП в общем случае

- Для непрерывной СВ:

*функция распределения заменяется на плотность вероятности*

$$\boldsymbol{L(\theta) = f(X_1 = x_1, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} f(X_i = x_i|\theta)}$$

- Аналогично, обычно работают с логарифмом функции правдоподобия

*конечно, при условии, что иксы – независимо и одинаково распределенные СВ

# Пример: нормальное распределение

- Вы научились различать оттенки вкуса и качества кофе
- И считаете, что в большинство дней качество достаточно среднее
- При этом иногда Вам попадается чашка выше среднего, а еще реже просто превосходная!
- Но иногда и ниже среднего, а еще реже такая, которую невозможно пить
- В общем все указывает на то, что распределение качества кофе по дням можно приблизить нормальным
- Оценим параметры $\mu$ и $\sigma^2$

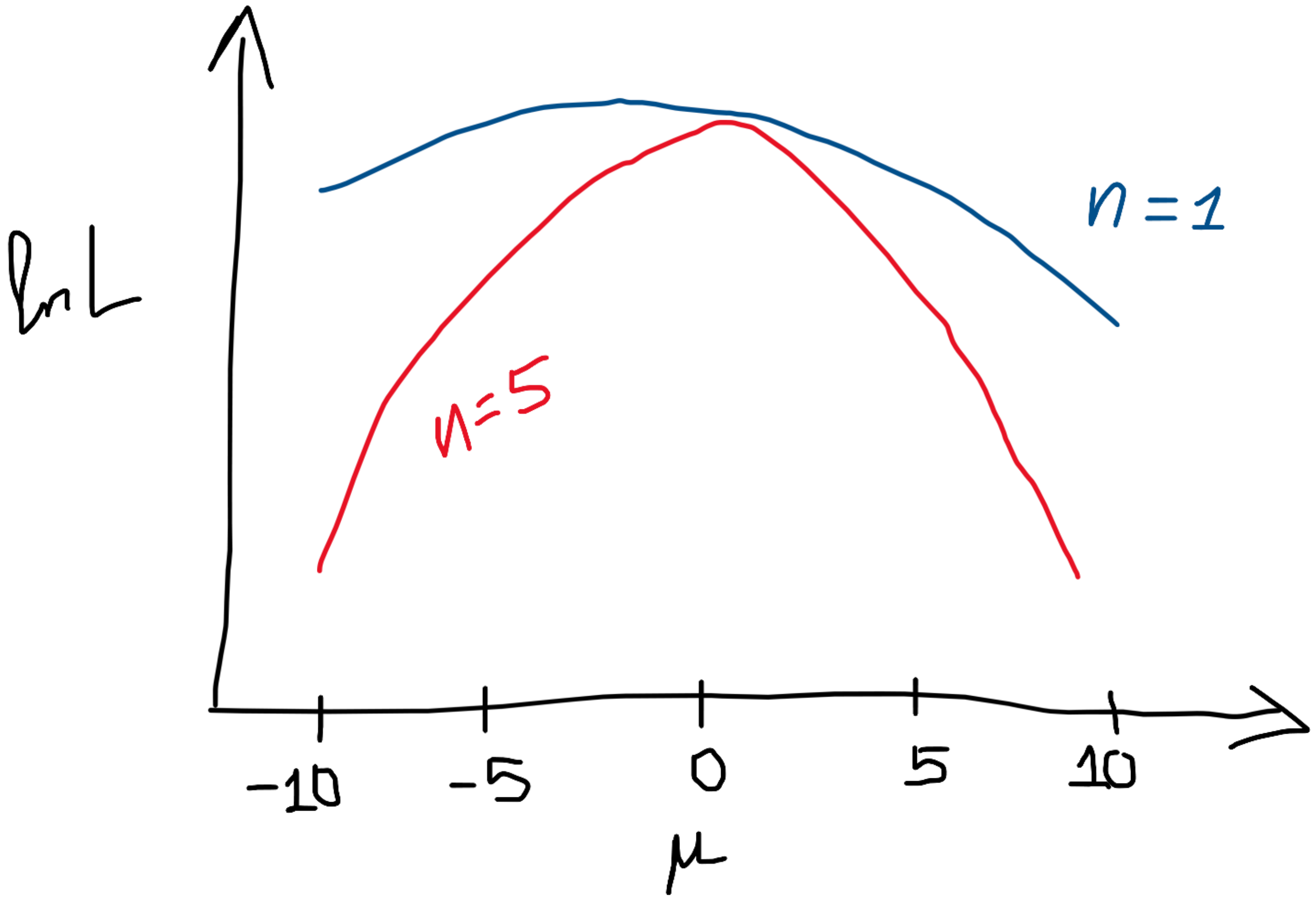
*Если захочется повторить выкладки, то проверить себя можно здесь: https://habr.com/ru/articles/830326/

# Точка максимума

**Нам важно, как себя ведет функция правдоподобия в окрестности точки максимума**

- Если вблизи максимума ф-я достаточно плоская, то по имеющимся наблюдениям достаточно сложно определить искомые параметры
- Те же самые значения можно наблюдать с высокой вероятностью и при совершенно иных параметрах
- При этом, если функция имеет ярко выраженный пик, то мы полагаем, что данные содержат «больше информации» об оцениваемых параметрах

# Точка максимума




$N(\mu, 1)$

- В функции правдоподобия одно слагаемое можно интерпретировать как лог-правдоподобие, вычисленное на основе одного наблюдения

- Чем больше выборка – тем ярче выражен максимум (если выборка репрезентативна)

# Информация Фишера

- Чем выпуклее функция, тем более четко выражен максимум

**Observed information для одного п-ра**: $I_o(\theta) = -\frac{d^2 \ln(L)}{d\theta^2}$

Вторая производная ф-ии правдоподобия по оцениваемому параметру со знаком минус

**Observed information для вектора п-ов:** $I_o(\theta) = -\left(\frac{d^2 \ln(L)}{d\theta_i d\theta_j}\right) = -H$

Матрица вторых производных ф-ии правдоподобия (матрица Гессе)

# Информация Фишера

- Мат.ожидание такой матрицы (по распределению наблюдений) называется **информационной матрицей Фишера**

$$\boldsymbol{E}(\boldsymbol{I}) = \boldsymbol{E}[\boldsymbol{I_o}(\boldsymbol{\theta})] = -\boldsymbol{E}\left(\frac{\boldsymbol{d^2}\ \mathbf{ln}(\boldsymbol{L})}{\boldsymbol{d\theta_i d\theta_j}}\right) = -\boldsymbol{E}(\boldsymbol{H})$$

**(теоретическая/ожидаемая инф. матрица Фишера)**

# Информация Фишера

- Наблюдаемая информация зависит от конкретных значений наблюдений/выборки

- Ожидаемая информация зависит только от закона распределения, и отражает, какой вклад вносит в правдоподобие среднестатистическое наблюдение

# Неравенство Рао-Крамера

- Если функция плотности $f(x_i|\theta)$ удовлетворяет условиям регулярности, то для любой несмещенной оценки $\hat{\theta}$, выполняется неравенство Рао-Крамера:

$$\boldsymbol{Var(\widehat{\theta}) \geq [I(\theta)]^{-1}}$$

Обратная к информации Фишера величина является нижней границей оценки дисперсии оцениваемого параметра

*Подробнее: https://tvims.nsu.ru/chernova/ms/lec/node27.html

# Справка: условия регулярности

**Основные:**

1. $\mathrm{A}: \{\mathrm{x}: \mathrm{f}(\mathrm{x}|\theta) > 0\}$ не зависит от $\theta$

   Область определения x не зависит от $\theta$

2. $\mathrm{f}(\mathrm{x}|\theta)$ непрерывно дифференцируема на всем множестве A
3. Производные ф-ии правдоподобия по $\theta$ могут быть вычислены путем изменения порядка дифференцирования/интегрирования/суммирования
4. Величина $E\left[\left(\frac{d\ln(L)}{d\theta}\right)^2\right]$ - положительна и конечна

**Дополнительные:**

5. $\mathrm{f}(\mathrm{x}|\theta)$ дважды дифференцируема на всем множестве $A$
6. $\mathrm{f}(\mathrm{x}|\theta)$ трижды дифференцируема на всем множестве $A$

Подробнее: https://tvims.nsu.ru/chernova/ms/lec/node27.html

# Свойства ММП-оценок

- Состоятельность

$$\underset{n\to\infty}{p\lim}\,\widehat{\theta} = \theta$$

С ростом объема выборки оценка сходится по вероятности к истинному значению параметра

# Свойства ММП-оценок

- Асимптотическая эффективность

$$\lim_{n \to \infty} Var(\widehat{\theta}) = [I(\theta)]^{-1}$$

Среди всех несмещенных оценок дисперсия оценки ММП является минимально достижимой. Этот минимум задается нижней границей Рао-Крамера

# Свойства ММП-оценок

- Асимптотическая нормальность

$$\hat{\boldsymbol{\theta}} \sim \boldsymbol{N}\left(\boldsymbol{\theta}, [\boldsymbol{I}(\boldsymbol{\theta})]^{-1}\right)$$

Оценка ММП имеет асимптотически нормальное распределение, что позволяет нам строить доверительные интервалы

В нерегулярных случаях оценка ММП может терять эти свойства

# Свойства ММП-оценок

- Инвариативность

Если $\hat{\theta}$ - оценка ММП для $\theta$, тогда если $g(t)$ — непрерывная функция, то $g(\hat{\theta})$ — ММП оценка для $g(\theta)$

Мы можем оценивать, например, $\sigma^2$ в нормальном распределении, извлекать корень и получать оценку для $\sigma$

Аналогично с оценками для $\lambda$ и $\frac{1}{\lambda}$ в экспоненциальном распределении и тд..

# Пример: расчет доверительного интервала для оценок норм. расп-я

Первые производные:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \cdot \sum(x_i - \mu) \qquad \frac{\partial \ln L}{\partial \sigma^2} = \frac{\sum(x_i - \mu)^2}{2\,\sigma^4} - \frac{n}{2\sigma^2}$$

Вторые производные:

$$H = \begin{pmatrix} \dfrac{\partial^2 \ln L}{\partial \mu^2} & \dfrac{\partial^2 \ln L}{\partial \mu\, \partial \sigma^2} \\[2ex] \dfrac{\partial^2 \ln L}{\partial \sigma^2\, \partial \mu} & \dfrac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{pmatrix}$$

# Пример: расчет доверительного интервала для оценок норм. расп-я

Вторые производные:

$$H = \begin{pmatrix} -\dfrac{n}{\sigma^2} & -\dfrac{1}{\sigma^4}\sum(x_i-\mu) \\ -\dfrac{1}{\sigma^4}\sum(x_i-\mu) & -\dfrac{\sum(x_i-\mu)^2}{\sigma^6}+\dfrac{n}{2\,\sigma^4} \end{pmatrix}$$

# Пример: расчет доверительного интервала для оценок норм. расп-я

Найдем мат. ожидание от Гессиана:

$$\mathbb{E}(H) = \begin{pmatrix} -\dfrac{n}{\sigma^2} & 0 \\ 0 & -\dfrac{n}{2\,\sigma^4} \end{pmatrix}$$

# Пример: расчет доверительного интервала для оценок норм. расп-я

Домножим на -1 и найдем обратную матрицу, получим информацию Фишера:

$$J(\theta) = -\,\mathbb{E}(H) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\,\sigma^4} \end{pmatrix}$$

$$[\,J(\theta)\,]^{-1} = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\,\sigma^4} \end{pmatrix}^{-1} = \begin{pmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\,\sigma^4}{n} \end{pmatrix}$$

# Пример: расчет доверительного интервала для оценок норм. расп-я

Получим асимптотическую дисперсию, заменив теоретические параметры на оцененные:

$$\widehat{Var}(\hat{\theta}) = [\,\hat{j}(\theta)\,]^{-1} = \begin{pmatrix} \frac{\hat{\sigma}^2_{ML}}{n} & 0 \\ 0 & \frac{2\,\hat{\sigma}^4_{ML}}{n} \end{pmatrix}$$

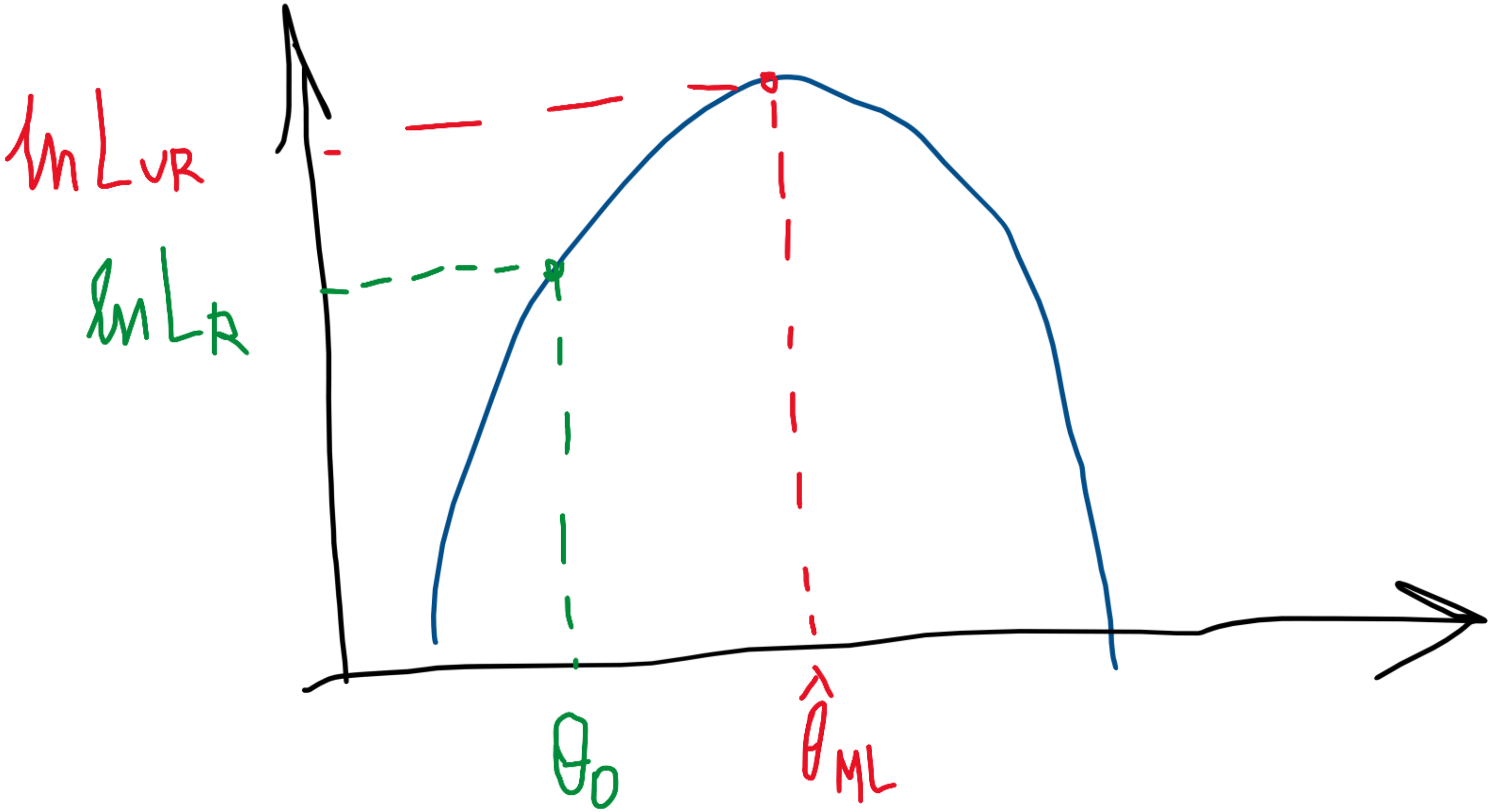
# Пример: расчет доверительного интервала для оценок норм. расп-я

$$\begin{pmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}^2_{ML} \end{pmatrix} \overset{asy}{\sim} N\left[\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \quad \begin{pmatrix} \frac{\hat{\sigma}^2_{ML}}{n} & 0 \\ 0 & \frac{2\,\hat{\sigma}^4_{ML}}{n} \end{pmatrix}\right]$$

Асимптотические доверительные интервалы:

$$\hat{\mu}_{ML} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2_{ML}}{n}} \qquad \hat{\sigma}^2_{ML} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{2\,\hat{\sigma}^4_{ML}}{n}}$$

# Тест отношения правдоподобий




$$H_0: \theta = \theta_0$$

$$H_a: \theta \neq \theta_0$$

# Тест отношения правдоподобий

Тестовая статистика выглядит как:

$$2 \cdot (\ln L_{UR} - \ln L_R) \underset{H_0}{\sim} \chi^2_q$$

$q$ — количество ограничений

$H_0: \theta = \theta_0$

$H_a: \theta \neq \theta_0$

# Тест отношения правдоподобий

Порядок проведения теста:

$H_0$: система из $q$ ограничений на неизвестные параметры

$H_a$: хотя бы одно из ограничений не выполняется

Оцениваем модель без ограничений, находим $\ln L_{UR}$

Оцениваем модель с ограничениями, находим $\ln L_R$

Наблюдаемое значение статистики:

$$LR_{obs} = 2 \cdot (\ln L_{UR} - \ln L_R)$$

Критическое значение статистики:

$$LR_{cr} = \chi_q^2 (1 - \alpha)$$

# Дополнительно

- Б. Демешев : Суть метода максимального правдоподобия; Метод максимального правдоподобия в непрерывном случае
- Выкладки по ММП: https://habr.com/ru/articles/830326/
- Цикл видео по прикладной статистике ФКН (week 12 полностью посвящена ММП): https://www.youtube.com/playlist?list=PLCf-cQCe1FRycw91L5669T4Krk7_sS9WY, в частности в 12-08 рассмотрен ММП для нерегулярного случая
- Н.Чернова, Математическая статистика, 2014: https://tvims.nsu.ru/chernova/ms/lec/node27.html (более строгие выкладки, есть док-ва)
- Тоже про теор. вещи и выводы для ММП: https://github.com/XuMuK1/psmo/blob/master/seminars/Seminar_1.ipynb

ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ