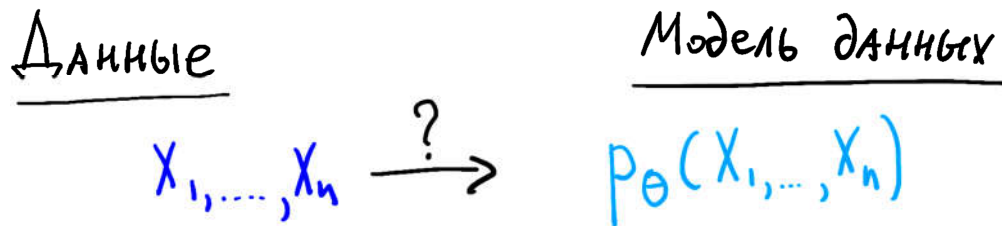


# Метод максимального правдоподобия

В этой главе мы вспомним метод максимального правдоподобия и на примере нескольких задач попробуем ещё сильнее раскрыть его потенциал.

## 1.1 Общая схема

Метод максимального правдоподобия (сокращённо ММП или MLE, maximum likelihood estimation) возникает как естественный инструмент для решения следующей задачи.



Пусть у нас имеются некоторые данные, представленные как случайные величины  $X_1, \dots, X_n$  с каким-то совместным распределением. Мы как исследователи делаем предположение, что это распределение задаётся некоторым набором параметров  $\theta \in \mathbb{R}^m$ , а далее пытаемся понять, какие параметры лучше всего соответствуют данным. Идея метода максимального правдоподобия интуитивно понятна: нужно выбрать такие параметры  $\hat{\theta} \in \mathbb{R}^m$ , что данные в такой модели наиболее вероятны. Если  $X_1, \dots, X_n$  дискретны (например, категориальные признаки), то *наиболее вероятно* значит буквально, что данные имеют наивысшую совместную вероятность; если же они непрерывны, то нужно максимизировать совместную плотность.

На самом деле в виде данных нам доступна реализация выборки  $x_1, \dots, x_n$ , на основе которой можем выписать достаточно короткий алгоритм (запишем для дискретной выборки, в непрерывном случае вероятность просто заменится на плотность).

1. Записать совместную функцию вероятности  $P_\theta$

$$L(\theta) = P_\theta(X_1 = x_1, \dots, X_n = x_n),$$

где зависимость модели от параметра обозначена нижним индексом;

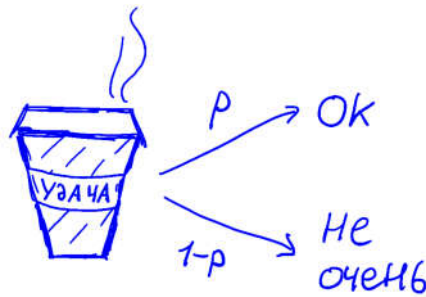
2. Найти параметр  $\hat{\theta}$ , делающий выборку наиболее вероятной:

$$\hat{\theta} := \arg \max_{\theta} L(\theta).$$

Функция  $L$  также часто называется *правдоподобием* ( $L$  - Likelihood).

## 1.2 Как вычислять оценки ММП

**Пример 1.1.** (Оцениваем качество кофе) Алексей часто посещает кофейню рядом со своим офисом. Кофе там, правда, от случая к случаю разный и бывало, что ему совсем не нравилось, а бывало, что вполне хорошо. Между днями он не может сравнить два хороших или два плохих кофе, поэтому он точно может сказать только одно: плохой был кофе в конкретный день или хороший. На основе своих посещений Алексей составил список дней, где подписал, когда и какой кофе ему попадался и предложил следующую вероятностную модель: каждый день независим от остальных и с одинаковой вероятностью  $p$  кофе является хорошим, а с вероятностью  $1 - p$  – плохим (Алексей надеется, что сама кофейня за это время не поменялась). Как на основе наблюдений оценить  $p$ ?



Мы имеем дело с выборкой независимых и одинаково распределённых случайных величин  $X_1, \dots, X_n \sim^{iid} \text{Ber}(p)$ . При известной реализации  $x_1, \dots, x_n$  вероятность данных в силу независимости записывается как

$$L(p) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i) = p^g (1 - p)^b,$$

где  $g$  – количество хороших дней, а  $b$  – плохих. Здесь полезно рассмотреть лог-правдоподобие

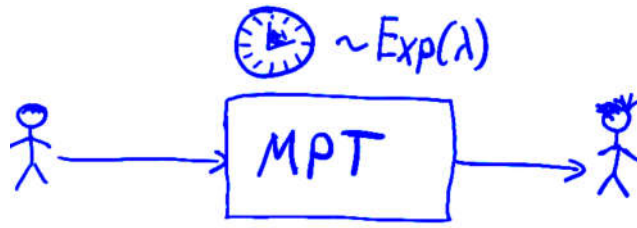
$$l(p) = \ln L(p)$$

и, промаксимизировав его, посчитать оценку

$$\hat{p} := \frac{g}{n}.$$

**Пример 1.2.** В медицинской информатике давно замечено, что времена обслуживания клиентов в кабинетах обследований имеют распределения с тяжёлыми хвостами. Самой первой идеей было предположить, что время обслуживания имеет экспоненциальное распределение с параметром  $\lambda > 0$ , обозначаемое как  $\text{Exp}(\lambda)$  и имеющее плотность

$$f_\lambda(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$



Если есть логи посещений клиентов, то можно посчитать время обслуживания  $T_i$  каждого клиента. В предположении независимости клиентов получим, что  $T_1, \dots, T_n$  – независимая выборка из распределения  $\text{Exp}(\lambda)$ . В этом случае при данных  $t_1, \dots, t_n$

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda t_i}$$

и решение даётся как

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n t_i}.$$

### 1.3 Свойства ММП

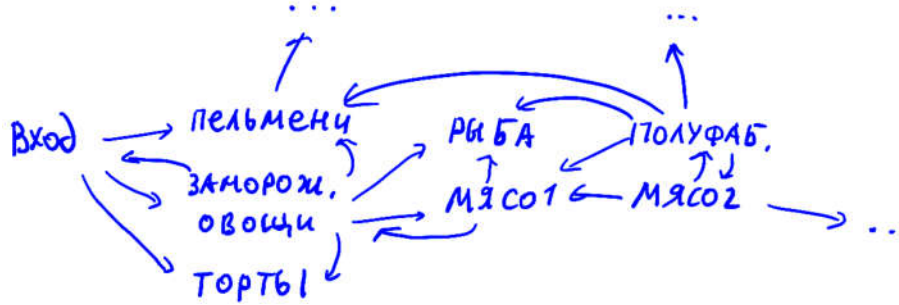
Из курса матстатистики нам известно, что при большом количестве условий регулярности, которые, тем не менее, не сильно ограничивающие, оценка ММП  $\hat{\theta}_n$  связана с истинным параметром  $\theta_0$  и как минимум в случае выборки  $X_1, \dots, X_n$  независимых величин обладают очень полезными для практики свойствами:

1. Асимптотическая несмещённость:  $\mathbb{E}[\theta_n] \rightarrow \theta_0$  при  $n \rightarrow \infty$ ;
2. Состоятельность:  $\theta_n \xrightarrow{\mathbb{P}} \theta_0$  при  $n \rightarrow \infty$ ;
3. Асимптотическая нормальность:  $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow^d \mathcal{N}(0, 1/i(\theta_0))$  при  $n \rightarrow \infty$ , где  $i(\theta_0)$  – информация Фишера для одного наблюдения в модели с параметром  $\theta_0$ .

Однако есть отдельные случаи, когда свойства ММП не выполняются из-за нарушений условий регулярности и о них следует помнить.

**Пример 1.3.** Проверьте свойства для ММП-оценки параметра  $a$  в модели  $X_1, \dots, X_n \sim \mathcal{U}(0, a)$ .

## 1.4 Модель клиента магазина



Клиент каким-то образом перемещается по магазину. Через приложение или систему наблюдения можно составить его траектории, которые бы выглядели как-то так:

<бак><шок><бак><кас><шок><кас><вых> ,

*бак* — означает бакалею, *шок* — шоколад, *кас* — кассу, *вых* — выход. Для схемы размещения отделов и анализа того, как клиенты переходят между ними можно предложить следующую модель клиента:

1. Клиент стартует на входе и все клиенты перемещаются независимо от других;
2. В каждый дискретный момент времени, находясь в отделе  $i$ , он делает выбор пойти в отдел  $j$  с вероятностью  $p_{ij}$ ;
3. Предполагаем условие нормировки  $\forall i \sum_j p_{ij} = 1$ .
4. Нет зависимости от истории:

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = t, X_{t-1} = x_{t-1}, \dots, X_{t-k} = x_{t-k}) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = t).$$

То, что мы описали, называется *цепью Маркова*, а последнее свойство — *Марковским свойством*.

В нашей модели параметрами являются вероятности переходов  $P = (p_{ij})$ ,  $i, j = 1, \dots, m$ . Величины  $X_t^k$  для клиента номер  $k$  являются зависимыми по  $t$ , поэтому функцию правдоподобия  $L(\theta)$  не получится выписать так просто, как в примерах выше...

К счастью, есть Марковское свойство, которое мы можем переформулировать так: положения клиента зависимы, но *переходы*  $(X_t \rightarrow X_{t+1})$  независимы и ещё мы знаем, что вероятности переходов

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = p_{ij}.$$

Рассмотрим для начала всего одного клиента с траекторией  $\tau$ , которая имеет длину  $n$ . Если бы в данных был только он, то функция лог-правдоподобия записалась бы как

$$l(P) = \sum_{j=1}^{n-1} \ln p_{x_j, x_{j+1}} = \sum_{i,j=1}^m n_{ij} \ln p_{ij},$$

где  $n_{ij}$  – количество переходов из  $i$  в  $j$ , наблюдаемое в данных. Мы полагали клиентов независимыми, поэтому когда мы рассмотрим полный датасет логов, то заметим, что концептуально результат не меняется и

$$l(P) = \sum_{i,j=1}^m n_{ij} \ln p_{ij}$$

после суммирования по всем клиентам. Нам теперь нужно только вычислить

$$\hat{P} = \arg \max_{P=(p_{ij})} l(P).$$

Однако остаётся одна небольшая техническая тонкость: все производные  $l(P)$  по параметрам нигде не обращаются в ноль. Это связано с тем, что в задаче мы пока не учли условие нормировки. Можем положить

$$p_{im} = 1 - \sum_{j=1}^{m-1} p_{ij}$$

и дальше остаётся вычислить оптимум.

Итоговый ответ поразительно интуитивно понятный:

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}.$$

В этом есть некоторая модельная сила ММП. Используя свойства оценки, мы можем даже проверять различные гипотезы с помощью тестов правдоподобия.

## 1.5 Модель авторегрессии

Модель авторегрессии  $AR(p)$  – это один из базовых инструментов в анализе временных рядов. Мы наблюдаем реализацию части последовательности  $X_t, X_{t+1}, \dots, X_{t+k}$  зависимых случайных величин в дискретном времени, которая формально бесконечная в обе стороны. Ставя перед собой задачу оценки параметров, про подобные модели мы можем, к примеру, спрашивать

1. Какие значения будут достигаться в будущем? (задача прогноза)
2. Как связаны элементы последовательности между собой и во времени? (задача моделирования)

Модель авторегрессии порядка 1, или  $AR(1)$  задаётся как

$$X_0 = x_0 \in \mathbb{R}, \quad X_{t+1} = \alpha X_t + \varepsilon_t,$$

где  $(\varepsilon_t)$  – последовательность независимых и одинаково распределённых случайных величин, каждая из которых имеет распределение  $\mathcal{N}(0, \sigma^2)$ . Такую последовательность также называют стандартным гауссовским белым шумом. Модель  $AR(p)$  отличается только количеством переменных в истории:

$$X_0 = x_0, \dots, X_{p-1} = x_{p-1}, \quad X_t = \sum_{i=1}^m \alpha_i X_{t-i} + \varepsilon_t.$$

Параметры в этой модели можно оценивать с помощью метода максимального правдоподобия, который в силу гауссовости с точностью до констант совпадает с методом наименьших квадратов. Рассмотрим для простоты случай  $p = 1$ , чтобы понять, как можно построить метод оценки.

Как и выше, наша главная цель – разбить зависимости в данных; напомним, что мы наблюдаем  $X_0, \dots, X_n$ , что не является выборкой независимых одинаково распределённых случайных величин. Но мы знаем, что  $\varepsilon_1, \dots, \varepsilon_n$  является, причём каждая из этих величин распределена по  $\mathcal{N}(0, \sigma^2)$ . Мы даже можем вычислить шум, если бы мы знали  $\alpha$ :

$$\varepsilon_t = X_t - \alpha X_{t-1}.$$

Незнание  $\alpha$  нисколько не мешает записать функцию лог-правдоподобия, которая при данных  $x_0, \dots, x_n$  будет выглядеть как

$$l(\alpha) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^n \frac{(x_i - \alpha x_{i-1})^2}{2\sigma^2},$$

уже знакомая нам по оценке параметров гауссовского распределения. Остаётся прооптимизировать по параметрам  $\alpha, \sigma^2$  и получить

$$\hat{\alpha} = \frac{\sum_{i=1}^n x_{i-1} x_i}{\sum_{i=1}^n x_{i-1}^2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\alpha} x_{i-1})^2.$$