Глава 6

Метод максимального правдоподобия

В предыдущих главах мы изучали задачу оценки неизвестных параметров модели при известных, пусть и зашумленных, результатах измерений тех величин, которые модель, собственно, и предсказывает. Такая задача сводится к поиску параметров, при которых модель наилучшим образом предсказывает измеренные значения.

Теперь рассматривается другая задача. Пусть наши измерения — реализации какой-то случайной величины, или случайного вектора. Будем считать, что случайность в данном случае — свойство присущее измеряемой физической величине. Конечно, в самом простом случае, случайность обусловлена шумами и погрешностями измерений. Но это не единственный источник, например, мы можем регистрировать энергию частиц, возникающих при некотором распаде, и даже с учетом конечной точности измерений окажется, что каждая конкретная частица имеет свое значение энергии, а вся совокупность имеет некоторое распределение.

Допустим мы знаем класс распределений, но хотели бы оценить конкретные параметры распределения, либо нам дополнительно известна модель, выражающая параметры распределения через другие параметры, имеющие самостоятельный смысл. *Метод максимального правдоподобия*[1] — один из способов определения параметров распределения случайной величины по известной выборке реализаций этой величины. Его идея состоит в следующем. Представим, что нам известна реализация некоторого случайного вектора $\mathbf{x}$ и известна функция плотности вероятности этого случайного вектора $p(\mathbf{x}|\theta)$, с точностью до некоторых параметров распределения θ, которые мы и хотим оценить. *Функцией правдоподобия*[2] называется плотность вероятности случайного вектора, вычисленная для известной реализации:

$$L(\theta) \equiv p(\mathbf{x}|\theta). \tag{6.1}$$

[1] *Maximum likelihood estimation*
[2] *likelihood*

Предположим, что раз мы увидели реализацию $\mathbf{x}$, то для истинных параметров θ вероятность такой реализации $p(\mathbf{x}|\theta)$ должна быть максимально возможной, среди всех альтернатив.

Конечно, максимальная вероятность исхода не гарантирует, что этот исход будет случаться всегда, но интуитивно понятно, что следуя этому предположению мы будем ошибаться реже всего. Метод максимального правдоподобия состоит в максимизации функции правдоподобия относительно неизвестных параметров распределения θ:

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ln L(\theta). \tag{6.2}$$

На практике вместо максимизации функции правдоподобия $L(\theta)$ удобнее использовать ее логарифм $\ln L(\theta)$. Во-первых, логарифм в силу своей монотонности не изменит положения экстремумов. Во-вторых, $\ln L(\theta)$ упрощает выражения для функций плотностей вероятности, имеющих экспоненциальную зависимость. В-третьих, оказывается удобным рассматривать важный частный случай выборки независимых одинаково распределенных величин. Например, пусть известно N реализаций случайного вектора $\mathbf{x}_i \in \mathrm{R}^{\mathrm{n}}$, тогда

$$\ln L(\theta) = \ln p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N|\theta) = \ln \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i|\theta). \tag{6.3}$$

Рассмотрим некоторые простые примеры, чтобы оценить насколько разумными получаются оценки параметров $\hat{\theta}$. Пусть дана выборка из N одинаково распределенных случайных величин x_i с нормальным распределением, найдем оценки параметров μ и σ^2. Для этого запишем логарифм функции правдоподобия

$$\ln L(\theta) = \sum_{i=1}^{N} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right), \tag{6.4}$$

и, дифференцируя выражение по μ и σ^2, окончательно находим

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{6.5}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 . \tag{6.6}$$

Предполагается, что сначала вычисляется оценка среднего $\hat{\mu}$ из (6.5), а затем она может быть подставлена как число в (6.6). Обратим внимание, что оценка параметра $\hat{\sigma}^2$, получаемая по методу максимального правдоподобия, оказывается смещенной.

Аналогично, пусть дана выборка из N одинаково распределенных случайных величин x_i с экспоненциальным распределением

$$p(x_i|\lambda) = \lambda \exp(-\lambda x_i), \tag{6.7}$$

найдем оценку параметра этого распределения. Логарифм функции правдоподобия будет выглядеть следующим образом:

$$\ln L(\theta) = \sum_{i=1}^{N} (\ln \lambda - \lambda x_i), \tag{6.8}$$

дифференцируя его по λ находим, что

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}. \tag{6.9}$$

Теперь рассмотрим более сложный пример, демонстрирующий, что метод максимального правдоподобия способен приводить к менее тривиальным результатам. Пусть у нас как и в разделе 3 есть линейная модель $A\theta = \mathbf{b}$. Как и прежде, мы считаем, что нам доступны отягощенные погрешностью ϵ измерения $\hat{\mathbf{b}} = \mathbf{b} + \epsilon$. Предположим, что все ϵ_i независимы и распределены нормально с нулевым средним и одинаковой дисперсией σ^2, тогда логарифм функции правдоподобия вектора $\hat{\mathbf{b}}$ выражается следующим образом:

$$\ln L(\hat{\mathbf{b}}) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\hat{\mathbf{b}} - A\theta\|^2, \tag{6.10}$$

где подлежащими оценке параметрами распределения являются константа σ^2 и вектор $\theta \in \mathrm{R}^{\mathrm{m}}$. Дифференцируя по параметрам, получим:

$$A^T A\hat{\theta} = A^T \hat{\mathbf{b}}, \tag{6.11}$$

$$\hat{\sigma}^2 = \frac{1}{n} \|\hat{\mathbf{b}} - A\hat{\theta}\|^2 = \frac{1}{n} \|\hat{\epsilon}\|^2. \tag{6.12}$$

Сравнив результат с формулами (3.10) и (3.17), можно увидеть, что при определенных предположениях метод наименьших квадратов может быть получен из метода максимального правдоподобия как частный случай. Оценка дисперсии по формуле (6.12) имеет смещение в отличии от (3.17).

Рассмотрим несколько полезных примеров, относящихся к многомерному нормальному распределению. Пусть в нашем распоряжении N независимых реализаций $\mathbf{x}_i \in \mathrm{R}^{\mathrm{m}}$, каждая из которых распределена нормально со средним μ и матрицей ковариации $\Sigma(\alpha)$, которая зависит от некоторого параметра α. Логарифм функции правдоподобия записывается как

$$\ln L(\theta) = -\frac{Nm}{2} \ln 2\pi - \frac{N}{2} \ln \det \Sigma(\alpha) - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mu, \Sigma^{-1}(\alpha) (\mathbf{x}_i - \mu)). \tag{6.13}$$

Используя градиент по вектору μ получаем выражение

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i. \tag{6.14}$$

Возьмем производную по параметру α, и после некоторых манипуляций получим:

$$\operatorname{tr}\Sigma^{-1}(\hat{\alpha})\left.\frac{\partial\Sigma(\alpha)}{\partial\alpha}\right|_{\hat{\alpha}}\left(I-\Sigma^{-1}(\hat{\alpha})\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\mu})\otimes(\mathbf{x}_i-\hat{\mu})\right)=0, \qquad (6.15)$$

где символ $\otimes$ обозначает внешнее векторное произведение.

Выражение (6.15) упрощается, если ввести дополнительные предположения. Например, пусть α — поочередно каждый элемент матрицы Σ, тогда

$$\hat{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\mu})\otimes(\mathbf{x}_i-\hat{\mu}). \qquad (6.16)$$

Пусть $\Sigma=\sigma^2 I$, и параметр $\alpha=\sigma^2$, тогда

$$\hat{\sigma}^2=\frac{1}{Nd}\sum_{i}\|\mathbf{x}_i-\mu\|^2. \qquad (6.17)$$

Пусть Σ блочная диагональная матрица:

$$\Sigma=\begin{bmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{bmatrix},$$

тогда можно показать, что уравнение (6.15) справедливо отдельно для Σ_1 и для Σ_2, причем от внешнего произведения нужно взять соответствующий блок нужного размера.

6.1 EM-алгоритм

Не всегда логарифм функции правдоподобия предполагает форму, удобную для аналитического поиска максимума. В частности, так происходит в случае моделей со *скрытыми переменными*. Несмотря на то, что мы могли бы применить один из методов численной оптимизации, существует еще более мощный и изящный метод, предложенный американскими исследователями (см. Dempster, Laird и Rubin 1977), который позволяет применять метод максимального правдоподобия удобно и эффективно в этих случаях.

Скрытой переменной принято называть физическую величину, которая непосредственно в измерениях не регистрируется, но в то же время и не является искомым параметров модели. Например, мы измеряем энергию частиц некоторым детектором: в таком случае измеряемой величиной является значение энергии E[3], предположим еще, что в ходе нашего эксперимента на детектор попадают частицы двух сортов (это могут быть электроны и мюоны; либо, например, фотоэлектроны и термоэлектроны), проблема в том, что теперь мы получаем смесь распределения энергий для двух сортов

[3] Чаще какой-то пропорциональной энергии величины, например заряда или напряжения.

частиц. В этом случае мы можем называть тип частицы скрытой переменной, физически она существует, но в ходе эксперимента не измеряется. Предположим, что нет технической возможности полностью подавить поток того или иного рода частиц, тогда мы привлекаем метод максимального правдоподобия, чтобы определить параметры смешанного распределения.

Хотя скрытая переменная не обязана быть дискретной переменной, EM-алгоритм традиционно демонстрируется на примере смеси двух нормальных распределений. Пусть у нас есть скрытая дискретная случайная переменная z, которая принимает значение 0 с вероятностью τ, и, соответственно, значение 1 с вероятностью $1-\tau$, а также пусть x — непрерывная измеряемая нормально распределенная величина. Соответствующие условные функции плотности вероятности запишутся следующим образом:

$$p(x|\{z=0\},\theta) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right), \tag{6.18}$$

$$p(x|\{z=1\},\theta) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right). \tag{6.19}$$

Пусть мы располагаем выборкой из N независимых реализаций x_i, и интересующий нас набор параметров состоит из пяти величин: τ, μ_1, μ_2, σ_1^2, σ_2^2.

Так как величина z_i является скрытой, для применения метода максимального правдоподобия необходимо вычислить маргинальное распределение для x_i:

$$\begin{aligned} p(x_i|\theta) &= \sum_{j\in\{0,1\}} p(x_i|\{z_i=j\},\theta)p\{z_i=j|\,\theta\} = \\ &= \frac{\tau}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-\tau}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right). \end{aligned} \tag{6.20}$$

Логарифм правдоподобия для всей выборки измерений по прежнему равен следующей сумме:

$$\ln L(\theta) = \sum_{i=1}^{N} \ln p(x_i|\theta). \tag{6.21}$$

Мы видим, что структура этой функции неудобна, она содержит в себе сумму логарифмов от суммы двух слагаемых. Оценки параметров θ не только не будут иметь аналитического вида, но их поиск также будет затруднен на практике с помощью численных методов оптимизации.

Итак, EM-алгоритм (метод) состоит из двух последовательных шагов, повторяющихся итеративно до достижения результата. Первый шаг — *усреднение*[4], второй шаг — *максимизация*[5]. Ключевой идеей метода является рассмотрение среднего логарифма полного правдоподобия (как если бы

[4] expectation
[5] maximization

скрытая переменная регистрировалась в эксперименте), вместо логарифма маргинального правдоподобия, имеющего неудобную форму. В нашем случае, логарифм полного правдоподобия записывается следующим образом:

$$\ln p(\mathbf{x}, \mathbf{z}|\theta) = \sum_{i=1}^{N} \left((1 - z_i) \ln \tau p(x_i | \{z_i = 0\}, \theta) + z_i \ln(1 - \tau) p(x_i | \{z_i = 1\}, \theta) \right), \tag{6.22}$$

где условные функции плотности вероятности задаются формулами (6.18)–(6.19).

Предположим, что некоторая грубая оценка параметров θ' нам известна (в реальности это оценка параметров выполненная на предыдущем шаге), тогда мы можем оценить значение скрытой переменной z_i для каждого измерения, понятно, что это значение будет зависит от измеренного значения x_i. Усредним логарифм полного правдоподобия по всем скрытым переменным:

$$Q(\theta|\theta') = \mathrm{E}_{\mathbf{z}|\mathbf{x},\theta'} \left[\ln p(\mathbf{x}, \mathbf{z}|\theta) \right]. \tag{6.23}$$

Или в рассматриваемом частном случае смеси двух нормальных распределений:

$$\begin{aligned} Q(\theta|\theta') = \sum_{i=1}^{N} \sum_{j \in \{0,1\}} & p(\{z_i = j\} | x_i, \theta') \left((1 - j) \ln \tau p(x_i | \{z_i = 0\}, \theta) \right. \\ & \left. + j \ln(1 - \tau) p(x_i | \{z_i = 1\}, \theta) \right) = \\ = \sum_{i=1}^{N} & \left(p(\{z_i = 0\} | x_i, \theta') \ln \tau p(x_i | \{z_i = 0\}, \theta) \right. \\ & \left. + p(\{z_i = 1\} | x_i, \theta') \ln(1 - \tau) p(x_i | \{z_i = 1\}, \theta) \right). \end{aligned} \tag{6.24}$$

Вместо логарифма суммы получилась сумма логарифмов, а значит с этим выражением можно удобно работать. Найдем новую улучшенную оценку искомых параметров максимизируя средний логарифм правдоподобия:

$$\hat{\theta} = \arg\max_{\theta} Q(\theta|\theta'). \tag{6.25}$$

Выражения вида $p(\{z_i = j\} | x_i, \theta')$, встречающиеся в (6.24), могут быть предварительно доведены до числа с использованием известного соотношения:

$$p(\{z_i = j\} | x_i, \theta') = \frac{p(x_i | \{z_i = j\}, \theta') p\{z_i = j|\theta'\}}{\sum_{k \in \{0,1\}} p(x_i | \{z_i = k\}, \theta') p\{z_i = k|\theta'\}}, \tag{6.26}$$

а значит рассматриваются нами как набор числовых коэффициентов, стоящих перед логарифмами правдоподобий, при дифференцировании (6.24)

по искомым параметрам. Подставляя условные функции плотности вероятности (6.18)–(6.19), получим:

$$p(\{z_i = 0\}\,|x_i, \theta') = \frac{\frac{\tau'}{\sqrt{2\pi\sigma_1'^2}} \exp\left(-\frac{(x_i-\mu_1')^2}{2\sigma_1'^2}\right)}{\frac{\tau'}{\sqrt{2\pi\sigma_1'^2}} \exp\left(-\frac{(x_i-\mu_1')^2}{2\sigma_1'^2}\right) + \frac{1-\tau'}{\sqrt{2\pi\sigma_2'^2}} \exp\left(-\frac{(x_i-\mu_2')^2}{2\sigma_2'^2}\right)}, \tag{6.27}$$

$$p(\{z_i = 1\}\,|x_i, \theta') = \frac{\frac{1-\tau'}{\sqrt{2\pi\sigma_2'^2}} \exp\left(-\frac{(x_i-\mu_2')^2}{2\sigma_2'^2}\right)}{\frac{\tau'}{\sqrt{2\pi\sigma_1'^2}} \exp\left(-\frac{(x_i-\mu_1')^2}{2\sigma_1'^2}\right) + \frac{1-\tau'}{\sqrt{2\pi\sigma_2'^2}} \exp\left(-\frac{(x_i-\mu_2')^2}{2\sigma_2'^2}\right)}. \tag{6.28}$$

Подробное выражение для среднего логарифма полного правдоподобия:

$$Q(\theta|\theta') = \text{const} + \sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta') \ln\tau + \sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta') \ln(1-\tau) +$$

$$+ \sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta') \left(-\frac{1}{2}\ln\sigma_1^2\right) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')(x_i-\mu_1)^2 +$$

$$+ \sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta') \left(-\frac{1}{2}\ln\sigma_2^2\right) - \frac{1}{2\sigma_2^2}\sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta')(x_i-\mu_2)^2. \tag{6.29}$$

И, взяв производные по параметрам, получаем простые выражения:

$$\hat{\tau} = \frac{\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')}{N}, \tag{6.30}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')x_i}{\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')}, \tag{6.31}$$

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')(x_i-\hat{\mu}_1)^2}{\sum_{i=1}^{N} p(\{z_i = 0\}\,|x_i, \theta')}, \tag{6.32}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta')x_i}{\sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta')}, \tag{6.33}$$

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta')(x_i-\hat{\mu}_2)^2}{\sum_{i=1}^{N} p(\{z_i = 1\}\,|x_i, \theta')}. \tag{6.34}$$

Сравнивая выражения (6.31) и (6.33) с (6.5) и выражения (6.32) и (6.34) с (6.6), можно заметить, что ЕМ-алгоритм «заменил» выборочное усреднение на взвешенное усреднение, где веса определяют степень принадлежности измерения x_i к тому или иному классу.

Докажем, что предложенный алгоритм сходится к максимуму маргинального правдоподобия. Для этого усредним следующее соотношение с обоих сторон:

$$\ln p(\mathbf{x}|\theta) = \ln p(\mathbf{x}, \mathbf{z}|\theta) - \ln p(\mathbf{z}|\mathbf{x}, \theta), \tag{6.35}$$

получим

$$\ln p(\mathbf{x}|\theta) = Q(\theta|\theta') - H(\theta|\theta'), \tag{6.36}$$

где введено обозначение

$$H(\theta|\theta') \equiv \sum_j p(\{z_i = j\}\,|x_i, \theta') \ln p(\{z_i = j\}\,|x_i, \theta). \tag{6.37}$$

Рассмотрим приращение логарифма маргинального правдоподобия, которое запишется в следующем виде:

$$\ln p(\mathbf{x}|\theta) - \ln p(\mathbf{x}|\theta') = (Q(\theta|\theta') - Q(\theta'|\theta')) + (H(\theta|\theta') - H(\theta'|\theta'))\,. \tag{6.38}$$

Разность $Q(\theta|\theta') - Q(\theta'|\theta') \geq 0$ по построению алгоритма, а разность H:

$$H(\theta|\theta') - H(\theta'|\theta') = -\sum_j p(\{z_i = j\}\,|x_i, \theta') \ln \frac{p(\{z_i = j\}\,|x_i, \theta)}{p(\{z_i = j\}\,|x_i, \theta')}. \tag{6.39}$$

Напомним, что логарифм — функция, выпуклая на интересующем нас интервале, поэтому действует неравенство Йенсона, и правую часть можно оценить как

$$\begin{aligned}
&\sum_j p(\{z_i = j\}\,|x_i, \theta') \ln \frac{p(\{z_i = j\}\,|x_i, \theta)}{p(\{z_i = j\}\,|x_i, \theta')} \\
&\quad \leq \ln \sum_j \{z_i = j\}\,|x_i, \theta') \frac{p(\{z_i = j\}\,|x_i, \theta)}{p(\{z_i = j\}\,|x_i, \theta')} = \ln \sum_j p(\{z_i = j\}\,|x_i, \theta) \leq 0.
\end{aligned} \tag{6.40}$$

Получается, что

$$\ln p(\mathbf{x}|\theta) - \ln p(\mathbf{x}|\theta') \geq Q(\theta|\theta') - Q(\theta'|\theta'), \tag{6.41}$$

т.е. любое увеличение функции $Q(\theta|\theta')$ приводит к увеличению соответствующего логарифма маргинального правдоподобия, в силу ограниченности последнего алгоритм сходится.