Метод максимального правдоподобия: идея

Задача: на основе имеющихся данных оценить значения параметров модели или распределения (коэффициент при переменной в регрессии, вероятность принадлежность документа к определенной тематике, математическое ожидание или дисперсия количественных показателей).
Идея: определить функцию правдоподобия, которая зависит от оцениваемого параметра, записать её с учётом имеющихся данных и найти такое значение параметра, при котором значение этой функции максимально.

Функция правдоподобия обозначается так:

$$L(\theta|x_1, x_2, \ldots, x_n),$$

где $x_1, x_2, \ldots, x_n$ – имеющиеся данные (в самом простом случае – выборка из n элементов), а θ – оцениваемый параметр.

Функция правдоподобия вычисляется так:

- дискретное распределение

$$L(\theta|x_1, x_2, \ldots, x_n) = P(X = x_1) \cdot P(X = x_2) \cdot \cdots \cdot P(X = x_n) = \prod_{i=1}^{n} P(X = x_i),$$

  где $P(X = x_i)$ – вероятность значения в точке x_i.
- непрерывное распределение

$$L(\theta|x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdot \cdots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i),$$

  где $f(x_i)$ – значение функции плотности в точке x_i.

Когда функция определена, нужно найти такое значение параметра θ, при котором функция достигает своего максимума. Оценку параметра, полученную методом максимального правдоподобия (*maximum likelihood estimation*) часто обозначают так:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta|x_1, x_2, \ldots, x_n),$$

где $\arg\max_{\theta}$ расшифровывается как «значение θ, при котором значение функции максимально».

Полезные напоминания

1. Свойства логарифмов:

   - $\log(a \cdot b) = \log(a) + \log(b)$
   - $\log(\frac{a}{b}) = \log(a) - \log(b)$
   - $\log(a^k) = k \log(a)$

   Основание логарифма может быть любым, здесь для определенности давайте считать, что log – это натуральный логарифм, то же что и ln.
2. Частные производные: производные по конкретной переменной. Если мы берём производную по одной переменной, то все другие переменные можно считать константами. Рассмотрим пример.

   Пример 1. Дана функция от двух переменных:

   $$f(x, y) = x^2 y$$

   Найдем частные производные – производные по x и по y:

   $$\frac{\partial f}{\partial x} = 2xy$$

   $$\frac{\partial f}{\partial y} = x^2.$$

Метод максимального правдоподобия: алгоритм

Алгоритм получения оценок параметров с помощью ММП (MLE):

1. Записать функцию правдоподобия $L(\theta)$ на основе имеющихся данных.
2. Найти натуральный логарифм функции правдоподобия L: $l(\theta) = \log(L)$.
3. Максимизировать функцию $l(\theta)$: взять (частную) производную по параметру θ, приравнять её к нулю и вывести θ. Полученное значение часто обозначают $\hat{\theta}_{MLE}$.

Комментарии:

1. Зачем нужен логарифм и почему нельзя максимизировать исходную функцию $L(\theta)$? Максимизировать L теоретически можно, просто задача сильно усложняется, поскольку функция L представляет собой произведение, а производная произведения считается не так просто и быстро:

   $$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$$

   Если учесть, что в L множителей обычно сильно больше, чем 2, то становится понятно, почему считать производную от исходной L невыгодно: даже для компьютера работа с таким числом множителей может оказаться непосильной (или посильной, но работать всё будет очень медленно). Переход к логарифму

сильно упрощает задачу, так как логарифм произведения – сумма логарифмов, а считать производную суммы легко и приятно:

$$(f(x) + g(x))' = f'(x) + g'(x).$$

При этом операция логарифмирования обладает хорошим свойством: при логарифмировании точки, в которых достигается минимум или максимум, никуда не сдвигаются. Другими словами, если у $f(x)$ максимум достигался в точке x_0, то и у $\log(f(x))$ максимум тоже будет достигаться в точке x_0.

2. Если нам необходимо оценить более одного параметра распределения, то нужно найти частные производные по всем интересующим параметрам, приравнять их к нулю и решать систему уравнений.
3. Иногда на практике (на компьютере) удобно вместо максимума находить минимум функции. Тогда пользуются следующим фактом: значение аргумента, при котором достигается максимум функции f, совпадает со значением, при котором достигается минимум функции, f, взятой с обратным знаком. Если найти аналитическое решение достаточно сложно, оценки находят с помощью специальных алгоритмов (метод градиентного спуска, динамическое программирование).

Пример 2. Дана выборка из бинарного распределения с параметром p (p – вероятность «успеха», вероятность 1):

$$0\ 1\ 0\ 1\ 0\ 0\ 0.$$

Параметр p нам неизвестен. Оценим его на основе выборки выше. Пусть вероятность встретить 1 равна p, тогда вероятность встретить 0 равна $(1-p)$. Если вспомнить, что функция правдоподобия определяется через произведение вероятностей, для нашей выборки получим:

$$L(p|x_1, x_2, \ldots, x_7) = (1-p) \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot = p^2 \cdot (1-p)^5.$$

Возьмём натуральный логарифм:

$$l(p) = \log(L(p)) = \log(p^2 \cdot (1-p)^5) = 2\log(p) + 5\log(1-p).$$

Чтобы найти p, при котором значение $l(p)$ максимально, найдём производную по p (здесь просто, других переменных нет) и приравняем её к нулю.

$$l'(p) = \frac{2}{p} - \frac{5}{1-p} = 0$$
$$\frac{2(1-p) - 5p}{p(1-p)} = 0$$
$$\hat{p}_{MLE} = \frac{2}{7}$$

Результат во многом ожидаемый: это просто доля единиц в выборке!

Пример 3. Теперь рассмотрим общий случай. Есть выборка из N элементов, N_0 раз в ней встречаются нули, N_1 – единицы. Найдем $\hat{p}_{MLE}$ в общем виде.

$$L(p|x_1, x_2, \ldots, x_n) = p^{N_1} \cdot (1-p)^{N_0}.$$

$$l(p) = \log(L(p)) = \log(p^{N_1} \cdot (1-p)^{N_0}) = N_1 \log(p) + N_0 \log(1-p).$$

$$l'(p) = \frac{N_1}{p} - \frac{N_0}{1-p}$$

$$\frac{N_1}{p} - \frac{N_0}{1-p} = 0$$

$$\hat{p}_{MLE} = \frac{N_1}{N_0 + N_1}.$$

Как раз и получили долю единиц среди всех элементов.

Пример 4. Дана выборка из биномиального распределения с параметрами $n = 8$ и p:

2 3 4 8 1 0 2 3

Давайте оценим p методом максимального правдоподобия.

$$L(p|x_1, x_2, \ldots, x_8) = P(X = 2) \cdot P(X = 3) \cdot \ldots \cdot P(X = 3)$$

Давайте сначала найдем все вероятности по отдельности:

$$P(X = 2) = C_8^2 \cdot p^2 \cdot (1-p)^6 (\text{будет учтена два раза, см. выборку})$$

$$P(X = 3) = C_8^3 \cdot p^3 \cdot (1-p)^5 (\text{будет учтена два раза, см. выборку})$$

$$P(X = 4) = C_8^4 \cdot p^4 \cdot (1-p)^4$$

$$P(X = 8) = C_8^8 \cdot p^8 \cdot (1-p)^0$$

$$P(X = 1) = C_8^1 \cdot p^1 \cdot (1-p)^7$$

$$P(X = 0) = C_8^0 \cdot p^0 \cdot (1-p)^8$$

Теперь подставим полученные значения в L, предварительно упростив выражение, сложив степени p и $(1-p)$:

$$L(p|x_1, x_2, \ldots, x_8) = C_8^2 \cdot C_8^3 \cdot \cdots \cdot C_8^3 \cdot p^{23} (1-p)^{41}.$$

Логарифмируем:

$$l(p) = \log(C_8^2 \cdot C_8^3 \cdot \ldots \cdot C_8^3) + 23\log(p) + 41\log(1-p)$$

Возьмем производную l по p:

$$l'(p) = \frac{23}{p} - \frac{41}{1-p}$$

Приравняем к нулю и найдём p:

$$\frac{23}{p} - \frac{41}{1-p} = 0$$

$$p_{MLE} = \frac{23}{64}$$

Получается, что оценка вероятности получить в одном испытании Бернулли «успех» равна сумме всех элементов выборки (общее число успехов), деленной на общее число исходов (8×8).

Пример 5. Рассмотрим случай, когда объём выборки не совпадает с числом испытаний Бернулли ($N \neq n$). Пусть есть выборка из биномиального распределения с $n = 8$ и некоторым p:

2 3 4 8 1 0

Вероятности значений возьмём из предыдущего примера. Перемножим, преобразуем и получим следующую функцию правдоподобия:

$$L(p|x_1, x_2, \ldots, x_6) = C_8^2 \cdot C_8^3 \cdot \cdots \cdot C_8^0 \cdot p^{18}(1-p)^{30}.$$

Логарифмируем:

$$l(p) = \log(C_8^2 \cdot C_8^3 \cdot \ldots \cdot C_8^0) + 18\log(p) + 30\log(1-p)$$

Возьмем производную l по p:

$$l'(p) = \frac{18}{p} - \frac{30}{1-p}$$

Приравняем к нулю и найдём p:

$$\frac{18}{p} - \frac{30}{1-p} = 0$$

$$\hat{p}_{MLE} = \frac{18}{48}.$$

Опять получается, что мы суммируем элементы выборки и делим на общее число исходов (8×6).

Пример 6. Рассмотрим произвольную выборку из биномиального распределения $Binom(n, p)$ объема N:

$$x_1, x_2, \ldots, x_N.$$

Оценим p. Запишем функцию правдоподобия L:

$$L(p|x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} C_n^{x_i} p^{x_i} (1-p)^{n-x_i} = \prod_{i=1}^{N} C_n^{x_i} \prod_{i=1}^{N} p^{x_i} (1-p)^{n-x_i}$$

Логарифмируем:

$$l(p) = \sum_{i=1}^{N} C_n^{x_i} + p^{\sum\limits_{i=1}^{N} x_i} (1-p)^{\sum\limits_{i=1}^{N} (n-x_i)}$$

Найдём производную по p:

$$l'(p) = \sum_{i=1}^{N} x_i \cdot \frac{1}{p} - \sum_{i=1}^{N} (n - x_i) \cdot \frac{1}{1-p}$$

Приравняем к нулю и найдём p:

$$\sum_{i=1}^{N} x_i \cdot \frac{1}{p} - (\sum_{i=1}^{N} n - \sum_{i=1}^{N} x_i) \cdot \frac{1}{1-p} = 0$$

$$\begin{cases} (\sum\limits_{i=1}^{N} x_i) \cdot (1-p) - (\sum\limits_{i=1}^{N} n - \sum\limits_{i=1}^{N} x_i) \cdot p = 0 \\ p(1-p) \neq 0 \end{cases}$$

$$(\sum_{i=1}^{N} x_i) \cdot (1-p) - (\sum_{i=1}^{N} n - \sum_{i=1}^{N} x_i) \cdot p = 0$$

$$\sum_{i=1}^{N} x_i - (\sum_{i=1}^{N} x_i) \cdot p - (\sum_{i=1}^{N} n) \cdot p + (\sum_{i=1}^{N} x_i) \cdot p = 0$$

$$\sum_{i=1}^{N} x_i - (\sum_{i=1}^{N} n) \cdot p = 0$$

$$\sum_{i=1}^{N} x_i - N \cdot n \cdot p = 0$$

$$\hat{p}_{MLE} = \frac{\sum\limits_{i=1}^{N} x_i}{N \cdot n}.$$

Получился общий результат для примеров, рассмотренных выше: сумма элементов выборки, делённая на общее число исходов ($N \times n$).