Конспектировали: Элина Суханов, Сергей Кечин.

1. Энтропия

Упражнение 1. Кот Васька хочет закодировать сообщение. Найти оптимальный бинарный код для кодирования. Вероятности передаваемых букв заданы таблицей.

К	О	Т
0.5	0.25	0.25

Решение. Оптимально закодировать К нулем, Т - 10, О - 11. Тогда ожидаемое количество бит, необходимых для одного сообщения $0.5 \cdot 1 + 0.25 \cdot 2 + 0.25 \cdot 2 = 1.5$

Упражнение 2. Подбрасывается правильная монетка. y_i - количество подбрасываний монетки до первого орла в i серии подбрасываний, например, $y_1 = 2, y_2 = 5, y_3 = 1, \ldots$ Придумать оптимальный двоичный код для y_i.

Решение. С увеличением количества подбрасываний вероятности серий убывают, поэтому оптимальный код выглядит следующим образом

Длина серии	Кодировка	Вероятность серии
1	0	0.5
2	10	0.25
3	110	0.125
...	...	...

Можно заметить, что длина сообщения (l) и его вероятность связаны соотношением $l_i = -\log_2 p_i$. Тогда в общем виде ожидаемое количество бит, необходимых для сообщения в оптимальной кодировке (для дискретных распределений) равно

$$-\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

Энтропия (H) - это математическое ожидание длины оптимально закодированного сообщения с информацией о случайной величине X, т.е.

$$H = -\operatorname{E}(\log_2 p(X))$$

Пример. Сравним энтропию для двух распределений

Значения	A	B
Вероятность	1/2	1/2
Длина	1	1

$$H_1 = 0.5 \cdot 1 + 0.5 \cdot 1 = 1$$

Значения	A	B
Вероятность	15/16	1/16
Длина	$\log_2 16/15$	4

$$H_2 = \frac{15}{16} \cdot \log_2 \frac{16}{15} + \frac{1}{16} \cdot 4 \approx 0.34$$

В первом случае оптимальная кодировка: A - 1, B - 0. Во втором случае в передаваемом сообщении будет часто встречаться несколько A подряд, за счет этого можно сократить объем сообщения, например, так: AAA - 1, AA - 01, A - 001, B - 0001. Поэтому $H_2 < H_1$.

2. Кросс-энтропия

Кросс-энтропия - это энтропия для не оптимальной длины сообщения (используем истиные вероятности, а длины не оптимальные)

$$CE_p(q) = -E_p(\log_2 q)$$

где p - это истиные вероятности, а q - ошибочные вероятности.

Пример. Распределение букв истиного сообщения, которое хочет передать кот Васька задано в таблице. К сожалению, кот закодировал сообщение не оптимально.

Значения	К	О	Т
Истиные вероятности	1/2	1/4	1/4
Ошибочные вероятности	1/4	1/2	1/4

Тогда кросс-энтропия равна

$$CE_p(q) = 1/2 \cdot 2 + 1/4 \cdot 1 + 1/4 \cdot 2 = 1.75$$

KL-дивергенция - это разница между энтропией и кросс-энтропией. KL-дивергенция показывает, сколько в среднем бит мы теряем, при использовании не оптимального кода.

$$KL_p(q) = CE_p(q) - H_p$$

В примере с котом Василием $KL_p(q) = 1.75 - 1.5 = 0.25$

3. Правдоподобие

Имеем выборку: $y_1, \ldots, y_n$
Известная модель мира: $L(y_1, \ldots, y_n|\Theta)$
где Θ - неизвестные параметры модели
Тогда:
$L(y_1, \ldots, y_n|\Theta)$ - вероятность получения данной выборки.
$-\log_2 L(y_1, \ldots, y_n|\Theta)$ - длина оптимального кода для передачи сообщения о данной выборке (при известной модели мира).

Две интерпретации метода максимального правдоподобия.

1. $\log_2 L(y_1, ..., y_n|\Theta) \to \max_\Theta$ - находим такое Θ, при которм данная выборка должна выпадать чаще всего

2. $-\log_2 L(y_1, ..., y_n|\Theta) \to \min_\Theta$ - находим такое Θ, при которм в оптимальной системе кодирования имеющаяся выборка получит самый короткий код.

Мини-упражнение:

Обычно величина $l_i = -\log_2 p_i$ показывает нам, сколько бит потребуется, чтобы закодировать сообщение оптимальным образом. По какому основанию следует брать логарифм $-\log_? p_i$, чтобы получить величину l_i не в битах, а в байтах?

$$-\log_2 x = 8 \text{ бит}$$
$$-\frac{1}{8}\log_2 x = 1 \text{ байт}$$
$$-\log_{256} x = 1 \text{ байт}$$

Упражнение 3. В озере водятся караси, щуки и крокодилы. Распределение вероятностей задано таблицей. Поклевки независисы. Найти Θ методом максимального правдоподобия для выборки карась, карась, щука, крокодил.

В озере	Карась	Щука	Крокодил
Вероятности	2Θ	Θ	$1-3\Theta$

Решение. $L = (2\Theta)^2 \cdot \Theta \cdot (1-3\Theta) = 4\Theta^3 \cdot (1-3\Theta)$
$\ln L = \ln 4 + 3\ln\Theta + \ln(1-3\Theta) \to \max_\Theta$
$\dfrac{\partial \ln L}{\partial \Theta} = \dfrac{3-12\Theta}{\Theta\cdot(1-3\Theta)} = 0 \Rightarrow \hat{\Theta} = 1/4$

4. Теоретические свойства правдоподобия

Score function - это случайная величина, равная градиенту логарифма правдоподобия:

$$s(\Theta) = grad \ln L$$

Score function зависит от выборки и параметров Θ.

Интерпретация: насколько сократится длина передаваемого сообщения о имеющейся выборке, если чуть чуть изменить Θ

Как найти $\mathrm{E}(s(\Theta))$**.** Введем обозначения:

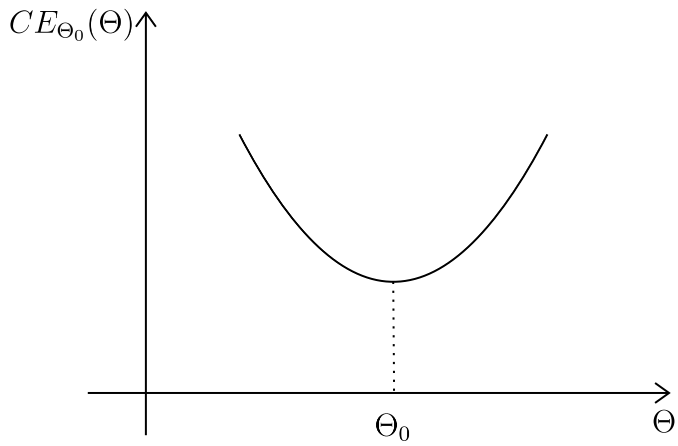
$s(\Theta)$ - градиент логарифма правдоподобия в произвольной точке Θ

$s(\Theta_0)$ - градиент логарифма правдоподобия в точке минимума Θ_0

Тогда:

$$\mathrm{E}_{\Theta_0}(s(\Theta)) = \mathrm{E}_{\Theta_0}(grad\,|_{\Theta=\Theta_0} \ln L(\Theta)) = grad\,|_{\Theta=\Theta_0} E_{\Theta_0}(\ln L(\Theta)) = -grad\,|_{\Theta=\Theta_0} CE_{\Theta_0}(\ln L(\Theta)) = 0$$

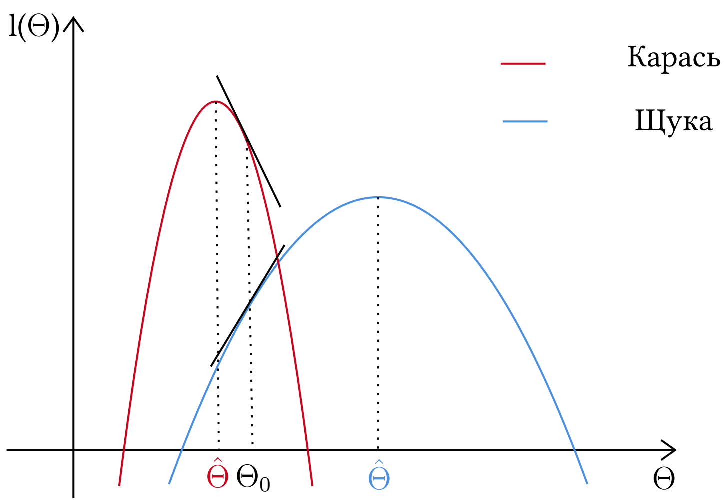
$\mathrm{E}_{\Theta_0}(\ln L(\Theta))$ - это минус кросс-энтропия, а градиент кросс-энтропии в точке истиного параметра Θ_0 равен нулю (в точке Θ_0 мы имеем оптимальный код и не можем больше уменьшать длину сообщения).



Пример. Выборка из озера размера 1.

В озере	Карась	Щука	Крокодил
Правдоподобие выборки (p)	2Θ	Θ	$1-3\Theta$
$\ln p$	$\ln 2\Theta$	$\ln \Theta$	$\ln(1-3\Theta)$
$s(\Theta)$	$\frac{1}{\Theta}$	$\frac{1}{\Theta}$	$-\frac{3}{(1-3\Theta)}$

Интуитивно можем предположить, что $E(s(\Theta)) = 0$, потому что когда-то производная окажется меньше 0, когда-то больше, а в среднем ожидаем, что мы получим 0.



Докажем формально:

$$\mathrm{E}(s(\Theta)) = 2\Theta \cdot \frac{1}{\Theta} + \Theta \cdot \frac{1}{\Theta} + (1-3\Theta) \cdot \frac{-3}{(1-3\Theta)} = 0$$

Пример. Найдем CE для $X \sim Beta(\Theta, 1)$.
$f(x) = const * x^{\Theta-1}(1-x)^{1-1} = const * x^{\Theta-1}$
Найдем константу:

$$\int\limits_0^1 x^{\Theta-1}dx = \frac{1}{\Theta} \Longrightarrow const = \Theta$$

Тогда:

$$f(x) = \Theta * x^{\Theta - 1}$$

Предположим, что $\Theta_0 = 7$, $\Theta = 8$. Тогда:

$$CE_{\Theta_0 = 7}(f(x, \Theta = 8)) = -\operatorname{E}_{\Theta_0}(\ln(8 * x^{8-1}) = -\operatorname{E}_{\Theta_0}(\ln 8 + 7 \ln x) = -\ln 8 - 7 * \int\limits_0^1 \ln x * 7 * x^6 dx$$

Как найти $\operatorname{Var}(s(\Theta))$.

$$s(\Theta) = \begin{bmatrix} \frac{\delta l}{\delta \Theta_1} \\ \vdots \\ \frac{\delta l}{\delta \Theta_k} \end{bmatrix}$$

$$\operatorname{E}(s(\Theta)) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\operatorname{Var}(s(\Theta)) = \begin{bmatrix} \operatorname{Var}(\frac{\delta l}{\delta \Theta_1}) & \\ \operatorname{Cov}(\frac{\delta l}{\delta \Theta_1}, \frac{\delta l}{\delta \Theta_2}) & \ddots \end{bmatrix}$$

Утверждение:

$$\operatorname{Var}(s(\Theta)) = -\operatorname{E}\left(\frac{\delta^2 l}{\delta \Theta \delta \Theta^T}\right)$$

Сравним для Θ_1:

$$\operatorname{Var}\left(\frac{\delta l}{\delta \Theta_1}\right) = \operatorname{E}\left(\left(\frac{\delta l}{\delta \Theta_1}\right)^2\right) \vee -\operatorname{E}\left(\frac{\delta^2 l}{\delta \Theta_1^2}\right)$$

$$\left(\frac{\delta^2 l}{\delta \Theta_1^2}\right) = \left(\frac{\delta \ln L(\Theta)}{\delta \Theta_1}\right)^2 = \left(\frac{1}{L(\Theta)} * \frac{\delta L}{\delta \Theta_1}\right)^2$$

$$\left(\frac{\delta l}{\delta \Theta_1}\right)^2 + \frac{\delta^2 l}{\delta \Theta_1^2} = \frac{1}{L} * \frac{\delta^2 L}{\delta \Theta_1^2}$$

$$\operatorname{E}\left(\left(\frac{\delta l}{\delta \Theta_1}\right)^2 + \frac{\delta^2 l}{\delta \Theta_1^2}\right) = \operatorname{E}\left(\frac{1}{L} * \frac{\delta^2 L}{\delta \Theta_1^2}\right) = \int\limits_\Theta \frac{1}{L} * \frac{\delta^2 L}{\delta \Theta_1^2} * L \, d\Theta = \int\limits_\Theta \frac{\delta^2 L}{\delta \Theta_1^2} d\Theta = \frac{\delta^2 \int\limits_\Theta d\Theta}{\delta \Theta_1^2} = \frac{\delta^2 1}{\delta \Theta_1^2} = 0$$

Получили, что:

$$\operatorname{E}\left(\left(\frac{\delta l}{\delta \Theta_1}\right)^2 + \frac{\delta^2 l}{\delta \Theta_1^2}\right) = 0$$

Домашнее задание

Аналогично доказать:

1. $\mathrm{E}\left(\dfrac{\delta l}{\delta\Theta_1} * \dfrac{\delta l}{\delta\Theta_2} + \dfrac{\delta^2 l}{\delta\Theta_1\delta\Theta_2}\right) = 0$
2. $\mathrm{E}\left(\dfrac{\Theta l}{\delta\Theta_1}\right) = 0$