16. Энтропия, информация

Рассмотрим некоторое случайное событие A, вероятность которого равна p. Каким образом выразить понятие ”количество информации, связанное с событием A”?

Естественно определить это количество информации (обозначим его I) как некоторую функцию от вероятности, $I = f(p)$. Как следует выбрать функцию f, чтобы I обладала теми свойствами, которые мы ожидаем от понятия ”информация ”?

Сформулируем эти свойства в виде двух аксиом:

1) Информация является функцией вероятности $I = f(p)$, удовлетворяющей условию $f(1) = 0$.
2) Если событие является композицией двух независимых событий, то вероятность должна быть равна сумме:

$$f(pq) = f(p) + f(q). \qquad (1)$$

Решением этого уравнения является $f(p) = ln(p)$. Действительно, в Параграфе 4 мы видели, что решение уравнения

$$G(x + y) = G(x)G(y) \qquad (2)$$

с условием $G(0) = 1$ есть $G(x) = e^{cx}$. Положим теперь $G = f^{-1}$ и обозначим $f(p) = x, f(q) = y$, тогда, согласно (18.1),

$$G\big(f(pq)\big) = G\big(f(p) + f(q)\big),$$

откуда $pq = G(x + y)$ и окончательно получаем уравнение (18.2) для функции G. Функцией, обратной к $G(x) = e^{cx}$ является $f(x) = \frac{1}{c}\ln(x)$.

Константа с служит для выбора основания логарифма, или, что то же самое, для выбора единицы измерения информации. Стандартно в качестве основания используют число 2, тогда для равновероятного случайного выбора ($p = \frac{1}{2}$) количество информации равно 1 (одному биту). Но в этом параграфе мы будем писать натуральный логарифм для упрощения некоторых формул.

Часто наряду с информацией рассматривают энтропию, обозначим ее H – меру неопределенности, имея в виду, что полученная в результате статистического эксперимента информация равна уменьшению неопределенности, то есть, энтропии, $\Delta I = -\Delta H$. Определим энтропию, связанную с событием A:

$$H = -\ln(p).$$

Рассмотрим теперь некий статистический эксперимент с N возможными исходами, имеющими соответствующие вероятности $\{p_i, i = 1,2, \dots, N\}$; точно также можно рассуждать о дискретной случайной величине ξ с распределением $\{p_i = P(\xi = x_i), i = 1,2, \dots, N\}$, но сейчас мы интересуемся только вероятностями событий p_i, а не значениями случайной величины. В результате проведения эксперимента может наступить одно из возможных событий A_i для случайной величины ξ,

$$A_i = \{\xi = x_i\},\ p_i = P(A_i).$$

С каждым событием связана соответствующая энтропия $H_i = -\ln(p_i)$, так что мы имеем случайную величину со значениями $-\ln(p_i)$ и соответствующими вероятностями p_i. В качестве энтропии, связанной с этим статистическим экспериментом (или, что тоже самое,

со случайной величиной ξ), естественно принять среднее значение, то есть, математическое ожидание энтропии.

Определение. Энтропия, соответствующая распределению вероятностей $\{p_i, i = 1,2, \dots, N\}$, равна

$$H = -\sum_{i=1}^{N} p_i \ln p_i.$$

Теорема 1. Энтропия обладает следующими свойствами:

1) $H \geq 0$, причем $H = 0$ тогда и только тогда, когда соответствующее распределение вырожденное.
2) Максимум энтропии достигается на равномерном распределении.
3) Если два статистических эксперимента независимы, то их совместная энтропия равна сумме энтропий (аддитивность энтропии).

Доказательство. Рассмотрим функцию

$$f(x) = \ln x + 1 - x\,;$$

она отрицательна и равна нулю лишь при $x = 1$; действительно, вычислив производную,

$$f'(x) = \frac{1}{x} - 1,$$

видим, что $x = 1$ - единственный экстремум и это максимум, поскольку вторая производная

$$f''(x) = -\frac{1}{x^2} < 0.$$

Таким образом, для любого $x \geq 0$,

$$\ln x \leq x - 1. \qquad (3)$$

Рассмотрим невырожденное распределение: пусть $0 < p < 1$, тогда из полученного неравенства выводим для энтропии

$$H = -p\ln p - (1-p)\ln(1-p) = -p\ \ln p - q\ \ln q \geq$$

$$\geq -p(p-1) - q(q-1) = p(1-p) + p(1-p) > 0.$$

Первое утверждение Теоремы доказано.

Для равномерного распределения $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$, следовательно

$$H = \ln N.$$

Пусть H - энтропия произвольного распределения, тогда с помощью того же неравенства (18.3) получаем

$$H - \ln N = \sum_{i=1}^{N} p_i \left(\ln \frac{1}{p_i} - \ln N\right) = \sum_{i=1}^{N} p_i \ln \frac{1}{Np_i} \leq \sum_{i=1}^{N} p_i \left(\frac{1}{Np_i} - 1\right) = 0.$$

Совместную энтропию двух случайных экспериментов обозначим

$$H = -\sum_{i,k} p_{ik} \ln p_{ik}\,,$$

где p_{ik}- совместное распределение вероятностей двух случайных величин, имеющих распределения по отдельности

$$q_k = \sum_i p_{ik} \ \text{ и } \ p_i = \sum_k p_{ik}.$$

Для независимых экспериментов $p_{ik} = p_i q_k$, поэтому энтропия равна сумме,

$$H = -\sum_{i,k} p_{ik} ln\, p_i - \sum_{i,k} p_{ik} ln q_k = -\sum_{i=1} p_i \ln p_i - \sum_{k} q_k ln q_k = H_1 + H_2. \blacksquare$$

Определение. Для непрерывной случайной величины с плотностью распределения $p(x)$ энтропия равна

$$H = -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx.$$

Замечание. Для непрерывных распределений энтропия не обязательно имеет знак положительный, но в качестве меры неопределенности в связи с понятием информации она также широко применяется.

Упражнение 1. Сформулировать и доказать свойство аддитивности энтропии для непрерывных распределений.

Упражнение 2. Как связаны H_ξ и H_η , если $\eta = a + b\xi$ $(b > 0)$, то есть, как меняется энтропия при линейном преобразовании ?

Еще одно уникальное свойство гауссовского распределения связано с понятием энтропии: оказывается, именно гауссовское распределение имеет максимальную энтропию, то есть наибольшую меру неопределенности.

Теорема 2. Среди всех распределений с заданными математическим ожиданием и дисперсией наибольшую энтропию имеет нормальное распределение: обозначим $H(p)$ энтропию распределения $p(x)$,

$$H(p) = -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx$$

и рассмотрим множество распределений, удовлетворяющих условиям

$$\int_{-\infty}^{\infty} x\, p(x) dx = a\,,$$

$$\int_{-\infty}^{\infty} (x-a)^2\, p(x) dx = \sigma^2\,.$$

Максимальное среди всех таких распределений значение энтропии $H(p)$ достигается для нормального распределения $\mathcal{N}(a, \sigma^2)$.

Доказательство. Обозначим p_0 гауссовское распределение

$$p_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

его энтропия $H(p_0)$ равна

$$H(p_0) = ln\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2}.$$

Если $p(x)$ и $q(x)$ – две плотности распределения, то справедливо неравенство

$$-\int_{-\infty}^{\infty} q(x) \ln q(x)\, dx \;\leq -\int_{-\infty}^{\infty} q(x) \ln p(x)\, dx$$

Действительно, с помощью неравенства (18.3) получаем,

$$-\int_{-\infty}^{\infty} q(x)\ln\frac{q(x)}{p(x)}dx = \int_{-\infty}^{\infty} q(x)\ln\frac{\mathrm{p}(x)}{q(x)}dx \le \int_{-\infty}^{\infty} q(x)\left(\frac{\mathrm{p}(x)}{q(x)}-1\right)dx = 0 \quad ,$$

что и требовалось.

Пусть теперь q – любое распределение, удовлетворяющее условиям Теоремы, а $p = p_0$. Тогда для энтропии распределения q получаем неравенство,

$$H(q) = -\int_{-\infty}^{\infty} q(x)\ln q(x)\,dx \le -\int_{-\infty}^{\infty} q(x)\ln p_0(x)\,dx =$$

$$= -\int_{-\infty}^{\infty} q(x)\left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{(x-a)^2}{2\sigma^2}\right]dx = ln\left(\sqrt{2\pi}\sigma\right)+\frac{1}{2} = H(p_0).\ \blacksquare$$

Рассмотрим некоторые конструкции, связанные с понятиями энтропии и информации для совместных распределений случайных величин; будем здесь считать эти распределения непрерывными. Условной энтропией ξ при условии $\eta = y$ называется

$$H(\xi/\eta = y) = -\int_{-\infty}^{\infty} p_{\xi/\eta}(x/y)\cdot\ln(p_{\xi/\eta}(x/y))\,dx$$

Средняя условная энтропия ξ относительно η равна

$$H_\eta(\xi) = E(H(\xi/\eta = y)) = -\int_{-\infty}^{\infty} p_{\xi,\eta}(x,y)\cdot\ln(p_{\xi/\eta}(x/y))\,dx$$

Энтропию двумерного распределения вектора (ξ,η) можно выразить через условную энтропию,

$$H_{\xi,\eta} = H(\xi,\eta) = -\iint p_{\xi,\eta}(x,y)\cdot\ln\left(p_{\xi,\eta}(x,y)\right)dxdy =$$

$$= -\iint p_{\xi,\eta}(x,y)\cdot\left(\ln p_{\eta/\xi}(y/x)+lnp_\xi(y)\right)dx\,dy = H(\xi)+H_\xi(\eta)$$

Количеством информации о величине ξ, содержащейся в величине η, называется

$$I_\eta(\xi) = H(\xi) - H_\eta(\xi)$$

очевидно, $I_\eta(\xi) \ge 0, I_\eta(\xi) = I_\xi(\eta)$.

If we have two separate probability distributions $P(\mathrm{x})$ and $Q(\mathrm{x})$ over the same random variable x, we can measure how different these two distributions are using the **Kullback-Leibler (KL) divergence**:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]. \qquad (3.50)$$

In the case of discrete variables, it is the extra amount of information (measured in bits if we use the base-2 logarithm, but in machine learning we usually use nats and the natural logarithm) needed to send a message containing symbols drawn from probability distribution P, when we use a code that was designed to minimize the length of messages drawn from probability distribution Q.

The KL divergence has many useful properties, most notably being non-negative. The KL divergence is 0 if and only if P and Q are the same distribution in the case of discrete variables, or equal "almost everywhere" in the case of continuous variables. Because the KL divergence is non-negative and measures the difference between two distributions, it is often conceptualized as measuring some sort of distance between these distributions. It is not a true distance measure because it is not symmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$ for some P and Q. This asymmetry means that there are important consequences to the choice of whether to use $D_{\mathrm{KL}}(P\|Q)$ or $D_{\mathrm{KL}}(Q\|P)$. See figure 3.6 for more detail.

A quantity that is closely related to the KL divergence is the **cross-entropy** $H(P, Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$, which is similar to the KL divergence but lacking the term on the left:

$$H(P, Q) = -\mathbb{E}_{\mathrm{x}\sim P} \log Q(x). \tag{3.51}$$

Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.

When computing many of these quantities, it is common to encounter expressions of the form $0 \log 0$. By convention, in the context of information theory, we treat these expressions as $\lim_{x\to 0} x \log x = 0$.

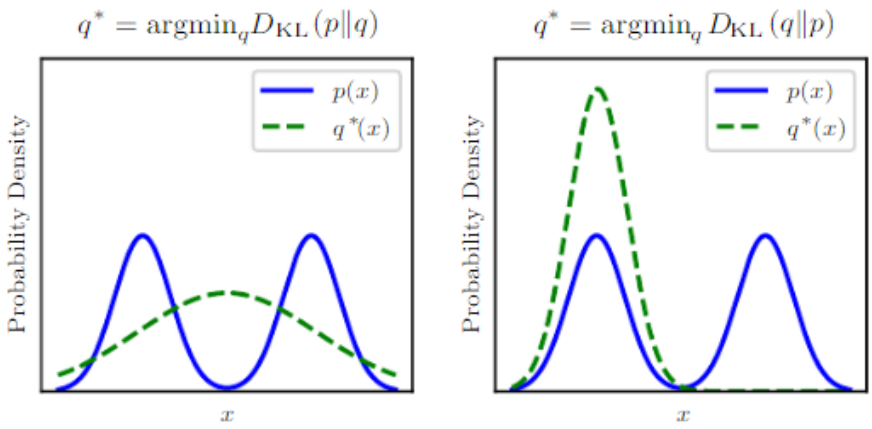

Figure 3.6: The KL divergence is asymmetric. Suppose we have a distribution $p(x)$ and wish to approximate it with another distribution $q(x)$. We have the choice of minimizing either $D_{\text{KL}}(p\|q)$ or $D_{\text{KL}}(q\|p)$. We illustrate the effect of this choice using a mixture of two Gaussians for p, and a single Gaussian for q. The choice of which direction of the KL divergence to use is problem dependent. Some applications require an approximation that usually places high probability anywhere that the true distribution places high probability, while other applications require an approximation that rarely places high probability anywhere that the true distribution places low probability. The choice of the direction of the KL divergence reflects which of these considerations takes priority for each application. *(Left)*The effect of minimizing $D_{\text{KL}}(p\|q)$. In this case, we select a q that has high probability where p has high probability. When p has multiple modes, q chooses to blur the modes together, in order to put high probability mass on all of them. *(Right)*The effect of minimizing $D_{\text{KL}}(q\|p)$. In this case, we select a q that has low probability where p has low probability. When p has multiple modes that are sufficiently widely separated, as in this figure, the KL divergence is minimized by choosing a single mode, to avoid putting probability mass in the low-probability areas between modes of p. Here, we illustrate the outcome when q is chosen to emphasize the left mode. We could also have achieved an equal value of the KL divergence by choosing the right mode. If the modes are not separated by a sufficiently strong low-probability region, then this direction of the KL divergence can still choose to blur the modes.