

ЭНТРОПИЯ

Марина Аюшеева, Яна Коротова, Олеся Майстренко, Елизавета Махнева, Дарья Писарева

Что это такое?

В теории информации энтропия – *степень неопределенности, связанная со случайной величиной.*¹

Также энтропию можно определить как *наименьшее среднее число бит, необходимое для кодирования некоторой информации.*

$$H = - \sum_{i=1}^n p_i \log p_i \quad \text{или} \quad H = - \int_{-\infty}^{+\infty} f(x) \log f(x) dx,$$

где $f(x)$ – функция плотности, p_i – вероятность i -го исхода.

Еще немножко :)

Условная энтропия — количество бит, необходимое для того, чтобы закодировать имеющуюся информацию о случайной величине Y при условии, что случайная величина X принимает определенное значение (или просто известна).

Рассчитывается так:

$$H(Y|X) = - \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)}$$

Совместная энтропия — степень неопределенности, связанная со множеством случайных величин.

Формула для расчета:

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

Все упомянутые выше герои обладают следующими свойствами:

- ◇ $H \geq 0$
- ◇ $H(Y|X) = H(X, Y) - H(X)$ или в более общем случае $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i|X_1, \dots, X_n)$
- ◇ $H(Y|X) \leq H(Y)$
- ◇ $H(X, Y) = H(X|Y) + H(Y|X) + I(X; Y) = H(X) + H(Y) - I(X; Y)$, где $I(X; Y)$ – взаимная информация о случайных величинах X и Y
- ◇ $I(X; Y) \leq H(X)$

¹<https://stackoverflow.com/questions/510412/what-is-the-computer-science-definition-of-entropy>

| **Взаимная информация** — мера взаимной зависимости двух случайных величин. |

Рассчитывается она так:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Чуть-чуть истории...

В 1948 году, исследуя проблему рациональной передачи информации через зашумлённый коммуникационный канал, **Клод Шеннон** предложил революционный вероятностный подход к пониманию коммуникаций и создал первую, истинно математическую, теорию энтропии.

Его сенсационные идеи быстро послужили основой разработки двух основных направлений: *теории информации*, которая использует понятие вероятности для изучения статистических характеристик данных и коммуникационных систем, и *теории кодирования*, в которой используются главным образом алгебраические и геометрические инструменты для разработки эффективных кодов.

Понятие энтропии, как меры случайности, введено Шенноном в его статье «*Математическая теория связи*» (англ. A Mathematical Theory of Communication), опубликованной в двух частях в Bell System Technical Journal в 1948 году.

В случае равновероятных событий (частный случай), остается зависимость только от количества рассматриваемых вариантов, и формула Шеннона значительно упрощается и совпадает с *формулой Хартли*, которая впервые была предложена американским инженером **Ральфом Хартли в 1928 году**, как один из научных подходов к оценке сообщений:

$$I = -\log p = \log N,$$

где I — количество передаваемой информации, p — вероятность события, N — возможное количество различных (равновероятных) сообщений.

А есть еще кросс энтропия!

| **Кросс энтропия** — минимальное среднее количество бит, необходимое для того, чтобы закодировать некоторую информацию, если схема кодирования базируется на некотором распределении q , а не истинном, p . |

$$CH(p, q) = - \int_{-\infty}^{+\infty} p(x) \log q(x) dx$$

Также кросс энтропию можно определить через *расстояние Кульбака – Лейблера*. Для начала стоит узнать, что это:

| **Расстояние Кульбака – Лейблера** — степень отдаленности друг от друга двух вероятностных распределений (называется также *относительная энтропия*). |

Рассчитывается для дискретного случая так:

$$D(P \parallel Q) = \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i \log q_i$$

Если имеем дело с абсолютно непрерывными распределениями, тогда формула для расчета выглядит так:

$$D(P \parallel Q) = \int_{-\infty}^{+\infty} f(x) \log f(x) dx - \int_{-\infty}^{+\infty} f(x) \log g(x) dx$$

Нетрудно заметить, что расстояние Кульбака – Лейблера равно разности энтропии и кросс-энтропии:

$$D(P \parallel Q) = H(p) - CH(p, q),$$

или

$$CH(p, q) = H(p) + D_{KL}(p \parallel q)$$

Применение энтропии и ее родственников

◇ Энтропийное кодирование

Как говорилось ранее, энтропия показывает наименьшее среднее число бит, необходимое для кодирования некоторой информации. Данное свойство используется, как ни странно, при кодировании информации.

Например, код Шеннона-Фано. С целью минимизации энтропии и, соответственно, оптимизации кода элементы с большой вероятностью появления кодируются меньшим числом символом. Таким образом, производится сжатие объема информации, что позволяет передавать большее количество информации, затрачивая меньший объем памяти.

◇ Построение решающих деревьев

Решающие деревья - метод, использующийся в машинном обучении и работающий по принципу принятия решений человеком. Каждое ветвление представляет собой разделение выборки на 2 части по порогу некоторого признака. Например, признак - длина, пороговое значение - 2,5. Все объекты, длина которых превышает 2,5, отделяются от объектов с длиной меньше 2,5 и дальнейший анализ проходят отдельно.

В данном методе расчет энтропии помогает определить оптимальный порог для каждого узла решения. А именно, подбирается такое разделение выборки, при котором сумма энтропий получившихся выборок минимальна среди возможных вариантов разбиений.

Это позволяет получать после разбиения выборки, содержащие наименее разнообразные по содержанию классы. Соответственно, признак и пороговое значение подбираются наиболее оптимально - алгоритм успешно отделяет объекты, принадлежащие одному классу.

◇ Применение в алгоритмах t-SNE и UMAP

В анализе данных часто возникает необходимость в снижении размерности, и в таких случаях на помощь приходят знания об энтропии, изученной в курсе теории вероятностей. Речь, конечно, идет не об энтропии как таковой, а об алгоритмах, которые базируются на теории.

При создании пространства меньшей размерности, t-SNE и UMAP используют кросс-энтропию как показатель эффективности перенесения свойств объектов. Чем меньше кросс-энтропия, тем ближе к истинному оказалось подобранное распределение.