Энтропия

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

Что это такое?

Вспомним всем известную игру "данетки". Так, чтобы понять, о чем идет речь, мы задаем уточняющие вопросы. Так вот, минимальное число вопросов, необходимое, чтобы выяснить полную инофрмацию об объекте, является *энтропией*.

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.*

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

где p_i – вероятность i-го исхода. Или вероятность того, что в "данетках"угадываемый объект обладает некоторой характеристикой. Например, нужно угадать, какого человека загадали, и с вероятностью 2/3 он моложе 30 лет, с вероятностью 1/3 старше 30 лет. Такое может быть в ситуации, когда молодых людей среди тех, кого могли бы загадать, больше, или загадывающий отдает предпочтение более молодым людям.

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).

Можно объяснить и проще – вспомним вновь игру выше. Вам необходимо узнать, кого загадал человек, ведущий в "данетке". Однако теперь он загадывает не одного человека, а *пару*. Каждый человек в этой паре с равными вероятностями может быть как моложе 30 лет (в 2 случаях из 3), так и старше 30 (в 1 случае из 3). И нам известно, что точно загадали человека моложе 30 лет (одному человеку из этой пары меньше 30 лет). Это и будет наше условие X. А далее мы уже исходя из данной информации должны отгадать, кого же все-таки загадали?

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.

Как и ранее, ведущий загадал пару людей. Однако теперь мы ничего заранее не знаем, кроме вероятностей, с которыми могли загадать людей, обладающих определенными признаками. Иными словами, вероятность, с которой загадали человека моложе 30, вероятность, с которой волосы загаданного человека имеют рыжий оттенок, и так далее.

Формула для рассчета:

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

[1] `https://stackoverflow.com/questions/510412/what-is-the-computer-science-definition-of-entropy`

Все упомянутые выше герои обладают следующими свойствами:

- $H \geqslant 0$
- $H(Y|X) = H(X,Y) - H(X)$ или в более общем случае $H(X_1, \ldots, X_n) = \sum\limits_{i=1}^{n} H(X_i|X_1, \ldots, X_n)$
- $H(Y|X) \leqslant H(Y)$
- $H(X,Y) = H(X|Y) + H(Y|X) + I(X;Y) = H(X) + H(Y) - I(X;Y)$, где $I(X;Y)$ – взаимная информация о случайных величинах X и Y
- $I(X;Y) \leqslant H(X)$

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

Другими словами, **взаимная информация** — это то, что нам известно о загаданной паре. Если ведущий сначала выбрал одного человека в паре, а затем подобрал второго так, чтобы они как-то были похожи друг на друга или, наоборот, максимально отличались, то информацию о паре можно вычислить, узнав всю информацию о первом человеке в этой паре, затем о втором, сложив их и вычтя те сведения, которые осведомляют о признаках сразу обоих людей.

Рассчитывается она так:

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

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.

$$CH(p,q) = -\sum_{i=1}^{n} p_i \log q_i$$

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:

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

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

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

или

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

А что, только для дискретных случайных величин?

Нет! :)

В случае, если Вы работаете с абсолютно непрерывными случайными величинами, энтропия и её родственники расчитываются по следующим формулам:

◇ Самая главная и простая энтропийка:

$$H(X) = -\int\limits_{-\infty}^{+\infty} f(x)\log f(x)dx$$

◇ Условная энтропия:

$$H(Y|X) = -\int\limits_{-\infty}^{+\infty} f(x,y)\log f_{Y|X}(y)dy$$

◇ Совместная энтропия:

$$H(X,Y) = -\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)\log f(x,y)dxdy$$

◇ Взаимная информация:

$$I(X;Y) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

◇ Кросс-энтропия:

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

◇ Расстояние Кульбака-Лейблера:

$$D(P\,||\,Q) = \int\limits_{-\infty}^{+\infty} f(x)\log f(x)dx - \int\limits_{-\infty}^{+\infty} f(x)\log g(x)dx$$