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:

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

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:

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

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

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

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

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

или

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

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 классов. 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.