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 иноформацию 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.

Задача 1. Красная Шапочка должна отнести бабушке пирожки. В какой точно деревне сейчас живет бабушка, Шапка не знает, но выбирает она из трех: А, Б и В. Известно, что внучка отнесла пирожки туда, куда нужно. Посчитайте энтропию, если попав в какую-то деревню, Шапочка никогда не сможет выбраться из неё.

Задача 2. Через несколько недель мама снова попросила Шапочку отнести бабушке пирожки. Правда, за все это время произошло много нового. Во-первых, бабушка перекочевала в другую деревню (какую – неизвестно). Во-вторых, в лесу завелся волк, известно, что он находится где-то в окрестности деревни, но неизвестно, какой именно. Если Шапке по дороге встретится волк, то пирожки бабушка не получит... Известно, что все обошлось и Шапка смогла отыскать бабушку. Посчитайте энтропию, если вновь Шапочка, попав в одну из деревень, остается в ней навсегда.

Задача 3. Перепуганная мама Красной Шапочки решила не рисковать здоровьем дочери и вызвала охотников, чтобы те поймали волка в одной из деревень. Когда все более-менее успокоилось, мама снова отправила дочку к бабушке. Однако, охотники еще не поймали волка, так как не могли его найти. А бабушка снова переехала в другую деревню. Если Шапочка окажется в одной деревне с волком, а охотников рядом не будет, то девочка провалит свою миссию. А если в это время охотники будут в той же деревне, что и волк, то они сразу же прибегут на помощь. Известно, что Шапочка смогла добраться до бабушки. Посчитайте энтропию, и вновь дорога в деревню – дорога в один конец.

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

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:

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

Задача 4. Шапка вновь отправилась на встречу к бабушке с корзинкой пирожков. И, конечно же, внучка не знала, в какой деревне на этот раз осела бабуля. Также Шапка не знала, что около деревень ошивалось целых N волков ($N \leqslant 3$), при этом возле каждой деревни либо был один волк, либо не было волков вообще. На борьбу с хищниками вышел один храбрый охотник, но находится он мог только в одной из трех деревень. Если Шапке не посчастливится, и она встретит волка, а помощь не подоспеет, то бабушка не получит свои пирожки. Известно, что Красной Шапочке снова удалось добраться до бабушки. Попади она в другую деревню, бабушка никогда бы не поела пирожки и, возможно, Шапку съел бы волк. Посчитайте условную энтропию в зависимости от N.

Задача 5. За обедом бабушка решила съесть два пирожка. Известно, что всего Шапочка принесла ей 20 пирожков с капустой и 30 пирожков с вареньем. Известно, что второй пирожок был с капустой. Посчитайте энтропию при условии, что первый пирожок тоже был с капустой.

> **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:

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

Задача 6. Охотники изловили почти всех волков, кроме одного. Красная шапочка снова направилась к бабушке с гостинцами. Из ее похода стало известно, что она смогла дойти до бабушки, но при этом охотники так и не поймали волка. Посчитайте совместную энтропию.

Задача 7. Девочка по возвращению домой поняла, что оставила у бабушки свою красивую красную шапку, поэтому на следующий день направилась обратно. За это время все персонажи успели поменять свое место нахождения. Получилось так, что и бабушка, и волк, и охотник оказались в одной деревне, поэтому волк не поймал Шапку, а внучка и бабушка благополучно встретились. Посчитайте совместную энтропию этих событий.

Все упомянутые выше герои обладают следующими свойствами:

- $H \geqslant 0$
- $H(Y|X) = H(X,Y) - H(X)$ или в более общем случае $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_n)$
- $H(Y|X) \leqslant H(Y)$
- $H(X,Y) = H(X|Y) + H(Y|X) + I(X;Y) = H(X) + H(Y) - I(X;Y)$, 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.

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

Задача 8. Шапочка думала, что в корзинке 20 пирожков с капустой и 20 – с вареньем, но ее мама все перепутала и вместо этого положила 10 с капустой и 30 с вареньем. По дороге к бабушке Шапочка решила съесть один пирожок. Он оказался с капустой. Посчитайте кросс энтропию.

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:

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

- Условная энтропия:

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

- Совместная энтропия:

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

- Взаимная информация:

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

- Кросс-энтропия:

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

- Расстояние Кульбака-Лейблера:

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

Задача 9. Пока Красная Шапочка бежала к бабушке из её корзинки в какой-то момент начали выпадать пирожки. Всего в корзинке их было N штук. Известно, что пирожки упали равномерно на некоторый участок дороги. Потерю всех пирожков Шапка обнаружила лишь по прибытии к бабушке и сразу решила собрать все выпавшие пирожки. Предполагая, что расстояние от дома Шапочки до дома Бабушки равно a, рассчитайте кросс-энтропию, если:

(а) Шапочка знает, что пирожки выпадали равномерно;
(б) Шапочка не знает, что пирожки выпадали равномерно;
(в) Какая из величин больше?

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.

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, 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.

Приведем пример работы алгоритма UMAP. Мы возьмем набор данных об одежде, который включает в себя 70000 черно-белых изображений различной одежды по 10 классам: футболки, брюки, свитеры, платья, кроссовки и т.д. Каждая картинка имеет размер 28x28 пикселей или 784 пикселя всего (то есть изначально у нас имеется 784-мерное пространство). UMAP перевел его в 2-мерное и визуализировал результат (см. ниже).

```
import numpy as np #импортируем библиотеку Numpy, чтобы работать с матрицами
from mnist import MNIST #импортируем библиотеку MNIST с наборами данных
#импортируем библиотеку matplotlib для построения графиков
import matplotlib.pyplot as plt
%matplotlib inline

mndata = MNIST('fashionmnist') #выбираем набор данных с фотографиями одежды
#обычно в машинном обучении выборку делят на обучающую и тестовую, чтобы
#обучить алгоритм и проверить, как он работает, но нам это не понадобится
#сохраняем обучающую выборку
#в переменную train сохраняется выборка с признаками объектов (фотографиями)
#в train_labels - ответы, то есть категории одежды,
#которые изображены на соответствующих фотографиях
train, train_labels = mndata.load_training()
test, test_labels = mndata.load_testing() #сохраняем тестовую выборку
#соединяем обучающую и тестовую выборку с признаками
data = np.array(np.vstack([train, test]), dtype=np.float64) / 255.0
#соединяем обучающую и тестовую выборку с ответами
target = np.hstack([train_labels, test_labels])
#записываем список из наименований одежды
```

```
classes = [
        'T-shirt/top',
        'Trouser',
        'Pullover',
        'Dress',
        'Coat',
        'Sandal',
        'Shirt',
        'Sneaker',
        'Bag',
        'Ankle boot']

import umap #импортируем библиотеку с алгоритмом UMAP

#анализируем набор данных
embedding = umap.UMAP(n_neighbors=10).fit_transform(data)

#рисуем график из получившегося нового распределения данных embedding
fig, ax = plt.subplots(1, figsize=(14, 10))
plt.scatter(*embedding.T, s=0.5, c=target, cmap='Spectral', alpha=1.0)
cbar = plt.colorbar(boundaries=np.arange(11)-0.5)
plt.setp(ax, xticks=[], yticks=[])
cbar.set_ticks(np.arange(10))
cbar.set_ticklabels(classes)
plt.title('Fashion MNIST Embedded via UMAP')
```

Вот что вышло:

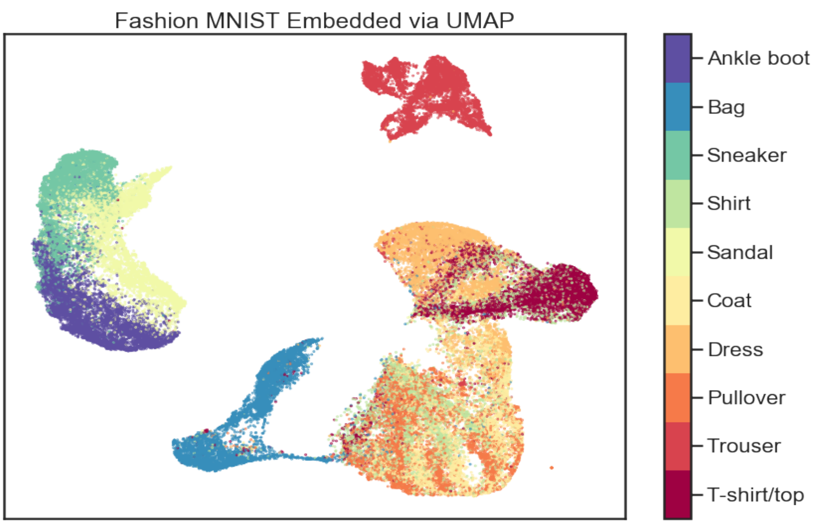

Рис. 1: Алгоритм UMAP

Изначально каждый пиксель являлся признаком объекта (фотографии) и принимал некоторое значение (цвет). Если бы каждая картинка состояла из 2 пикселей, мы бы смогли построить график, где по оси абсцисс отложен цвет 1 пикселя, по оси ординакт цвет 2 пикселя и изобразить точками все объекты.

В нашем случае из-за большого количества пикселей мы не можем так сделать с исходной выборкой. Однако, если мы преобразуем выборку таким образом, что останется всего лишь 2 признака, то мы сможем визуализировать и ее. Только получившиеся 2 признака будут уже не пикселями, а абстрактными признаками, которые алгоритм получает из исходных.

Но получить такие 2 абстрактных признака можно очень многими способами. Но при этом нам необходимо, чтобы получившиеся 2 признака описывали исходную выборку как можно лучше - чтобы при визуализации мы видели не случайно нарисованное изображение, а некоторое отображение начального пространства. Именно тут алгоритм применяет кросс-энтропию. Минимизируя кросс-энтропию, между двумя распределениями (истинным и созданным алгоритмом), мы сокращаем отличия между ними, что позволяет получить максимально приближенный к нужному результат.

Ключи к сердцу Красной Шапочки

Тут можно найти ответы и решения ко всем задачам, которые были представлены выше. Переходите к этому разделу, только если уже всё решили и хотите проверить себя :)

Задача 1.

Так как деревни всего 3 и вероятность Красной Шапочки прийти в любую из них одинаковая, то примем параметр р = 1/3. Тогда, следуя формуле выше, получаем, что:

$$H = -\sum_{i=1}^{3} \frac{1}{3} \log \frac{1}{3}$$

Результатом подсчёта данного выражения будет число 1.58, что при округлении до ближайшего целочисленного даёт значение 2. Оно и верно! Смотрите, можно узнать, в какой деревне живёт бабушка на 100 % всего за два вопроса. Например, 1) Название деревни - это гласная буква? (если да, то ответ уже найден; если нет следует задать еще один вопрос); 2) Название деревни начинается на букву Б? (в любом случае, далее вы уже сможете ответить на вопрос правильно с вероятностью, равной единице).

Ответ: 2

Задача 2.

Теперь давайте подсчитаем количество возможных вариантов размещения трех агентов: Шапки, бабули и волка. Очевидно, что общее число размещений каждого из 3 агентов в 3 местах равно 27. Далее, чтобы узнать вероятность хорошего исхода, необходимо узнать, во скольких случаях бабушка не получает свои пирожки (то есть Шапка и волк оказываются в одной деревне вместе). Причём заметим, они могут оказаться в одной деревне только вдвоём или с бабулей, однако и том, и в другом случае сказка не имеет счастливого конца. Таких вариантов всего 9. Тогда вариантов, в которых бабуля наслаждается пирожками 18 (получено как 27-9). Тогда подсчитать энтропию в данном случае можно по следующей формуле:

$$H = -\sum_{i=1}^{18} \frac{1}{18} \log \frac{1}{18}$$

Результатом подсчёта данного выражения будет число 4.17, что при округлении до ближайшего целочисленного дает значение 5. Это означает, что за 5 вопросов возможно узнать, получила ли бабушка свои пирожки и в какой деревне она находится.

Ответ: 5

Задача 3.

В данном случае наших героев стало уже четверо! Теперь количесвто возможных расположений увеличилось до 81, сотвественно (легко дойти до этого самостоятельно). Теперь выясним количество благоприятных исходов, то есть, когда Шапка, волк и охотники оказались в одной деревне. Для этого нам проще выяснить обратное событие, то есть когда Шапка и волк оказались в одной деревне вместе (как с бабулей, так и без нее), а охотники в это же время были в другом месте. Например, пусть Шапка и волк находятся вместе в 1 деревне, тогда возможных вариантов расположения охотников и бабушки, при которых Шапка не выживает и бабуля не получает свои пирожки - 6 штук. Тогда, распространив случай на три деревни, получаем, что всего печальных исходов 18. Тогда благоприятных исходов 63. Тогда получить ответ можно, воспользовавшишь формулой:

$$H = -\sum_{i=1}^{63} \frac{1}{63} \log \frac{1}{63}$$

Результатом подсчёта данного выражения будет число 5.9, что при округлении до ближайшего целочисленного дает значение 6.

Ответ: 6

Задача 4.

Задача 5.

Задача 6.

Задача 7.

Задача 8.