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 информацию 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. Любую из деревень Шапка выбирает равновероятно. Известно, что внучка отнесла пирожки туда, куда нужно. А также вернуться на место выбора деревни - невозможно. Посчитайте энтропию местонахождения бабушки.

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), но равновероятно. Во-вторых, в лесу завелся волк, известно, что он находится где-то в окрестности деревни, но неизвестно, какой именно, и также равновероятно в любой из трех. Если Шапке по дороге встретится волк, то пирожки бабушка не получит. . . Вернуться на перепутье Шапка не может. Известно, что все обошлось и Шапка смогла отыскать бабушку. Посчитайте энтропию местонахождения бабушки.

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 (равновероятно в любую другую). 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. Вернуться на место выбора деревни Красная Шапка не может. Известно, что Шапочка смогла добраться до бабушки. Посчитайте энтропию местонахождения бабушки.

[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 двух случаях из трех), так и старше 30 (в одном случае из трех). 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 (в любой из возможных - равновероятно). 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(опять же равновероятно). 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 местонахождения бабушки в зависимости от N.

> **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 5. 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 в такой ситуации.

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.

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

Задача 6. 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 в такой ситуации.

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:

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

◇ Расстояние Кульбака-Лейблера:

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

Задача 7. 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

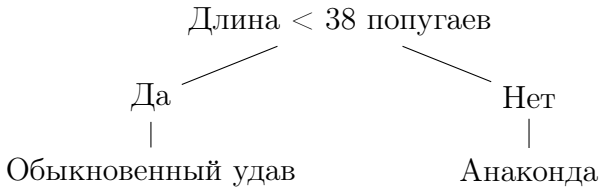
◇ **Энтропийное кодирование**

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.

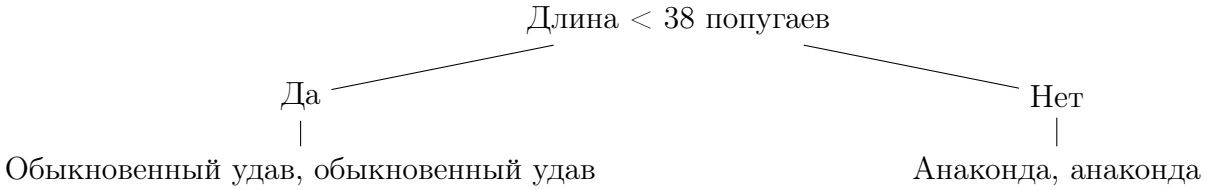
◇ **Построение решающих деревьев**

Решающие деревья – метод, использующийся в машинном обучении и работающий по принципу принятия решений человеком. Каждое ветвление представляет собой разделение выборки на две части по порогу некоторого признака. Например, признак – длина, пороговое значение – 38. Все объекты, длина которых превышает 38, отделяются от объектов с длиной меньше 38 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.

Например, у нас есть выборка объектов с одним признаком, длина: 22 попугая (обыкновенный удав), 46 попугаев (анаконда), 40 попугаев (анаконда), 31 попугай (обыкновенный удав). Мы выбираем порог: 38 или 44 попугаев? Попробуем разделить выборку по 38 попугаям:



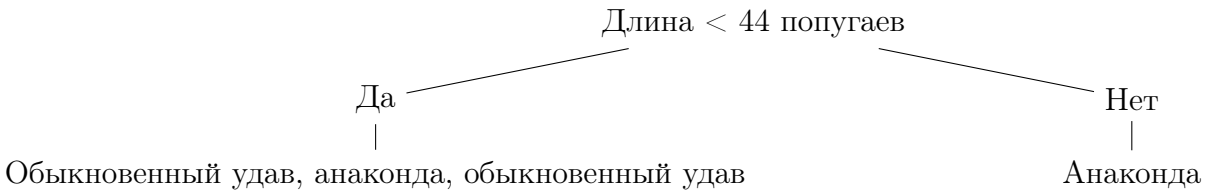
При расчете энтропии $0 \cdot \log_2 0$ считается равным 0, несмотря на $\log_2 0$. За вероятность принимается вероятность встретить данный класс в новой выборке.

Энтропия левой части: $-(1 \cdot \log_2 1 + 0 \cdot \log_2 0) = 0$.

Энтропия правой части: $-(1 \cdot \log_2 1 + 0 \cdot \log_2 0) = 0$.

Суммарная энтропия получилась: 0.

Попробуем разделить выборку по 44 попугаям:



Энтропия левой части: $-(\frac{1}{3} \cdot \log_2 \frac{1}{3} + \frac{2}{3} \cdot \log_2 \frac{2}{3}) \approx 0.92$.

Энтропия правой части: $-(1 \cdot \log_2 1 + 0 \cdot \log_2 0) = 0$.

Суммарная энтропия получилась: 0.92.

В первом случае мы идеально разделили выборку при энтропии, равной нулю. Во втором случае нам удалось отделить одну анаконду, но не удалось отделить классы. Так и энтропия в первом случае оказалась меньше, чем во втором. Причем ее равенство нулю необязательно – любое значение меньше 0.92 показало бы, что первый случай более оптимален. Здесь же, в виду неотрицательности энтропии, однозначно можно сказать, что критерий "длина $<$ 38 попугаев"дает оптимальный результат.

Энтропия позволяет получать после разбиения выборки, наименее разнообразные по содержанию классов (менее хаотичные). Соответственно, признак и пороговое значение подбираются наиболее оптимально – алгоритм успешно отделяет объекты, принадлежащие к 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. 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.

Изначально каждый пиксель является признаком объекта (фотографии) и принимает некоторое значение (цвет). Если бы каждая картинка состояла из двух пикселей, мы бы смогли построить график, где по оси абсцисс отложен цвет одного пикселя, по оси ординат цвет второго пикселя, и изобразить точками все объекты.

У нас каждая картинка состоит из 784 пикселей – 784-мерное пространство, поэтому изобразить его проблематично. Но если мы преобразуем выборку таким образом, что останется всего лишь два признака, то мы сможем визуализировать ее. Получившиеся два признака будут уже не пикселями, а абстрактными признаками, которые алгоритм получает из исходных.

Получить два новых признака можно очень многими способами. Но они должны описывать исходную выборку как можно лучше - чтобы при визуализации мы видели не случайно нарисованное изображение, а отображение начального пространства. Именно тут алгоритм применяет кросс-энтропию. Минимизируя кросс-энтропию между исходным и

новым распределением, мы сокращаем отличия между ними, что позволяет получить максимально приближенное отображение данных на плоскости.

Реализуем описанный алгоритм.

Внимание! Библиотека UMAP требует предварительной установки[2]

Импортируем нужные библиотеки:

```
import numpy as np # работа с матрицами
from mnist import MNIST # наборы данных
import matplotlib.pyplot as plt # построение графиков
%matplotlib inline
import umap # алгоритм UMAP
```

Загружаем набор данных с фотографиями одежды.

```
mndata = MNIST('fashionmnist')
train, train_labels = mndata.load_training()
test, test_labels = mndata.load_testing()
data = np.array(np.vstack([train, test]), dtype=np.float64) / 255.0
target = np.hstack([train_labels, test_labels])
```

Создаем список из наименований одежды.

```
classes = ['T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal', '
```

Запускаем UMAP.

```
embedding = umap.UMAP(n_neighbors=10).fit_transform(data)
```

Визуализируем результат.

```
plt.figure(figsize=(14, 10))
sns.scatterplot(*embedding.T, hue=target, s=4, palette = 'Spectral',
legend='full', alpha=1.0, edgecolor="none")
plt.legend(classes, loc=1, fontsize='large')
```

Вот что вышло:

Это была задача для 784-мерного пространства. И UMAP оказался неким черным ящиком, который перевел его в двумерное пространство непонятным образом. Давайте рассмотрим задачу попроще, чтобы понять, как работает алгоритм.

Пример. В КЭБ717 учатся 3 прекрасные девушки (наши объекты): Даша, Яна и Лиза. Даша – блондинка с голубыми глазами, у Яны темные волосы и карие глаза, у Лизы русые волосы и карие глаза. Попробуем отобразить объекты в одномерном пространстве, то есть с одним признаком.

Решение. Для начала переведем признаки в числовой вид. 1 признак – цвет глаз: карие(1), голубые(0). 2 признак – цвет волос: светлые(0), русые(1), темные(2). Теперь у нас есть три объекта:

[2]Почитать про установку: `https://umap-learn.readthedocs.io/en/latest/`

	Цвет глаз	Цвет волос
D	0	0
Y	1	2
L	1	1

Теперь найдем расстояния между объектами. По умолчанию, в UMAP стоит евклидова метрика. В нашем случае она также подходит, так как наши признаки могут быть интерпретированы как координаты точек в пространстве.

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 три и вероятность Красной Шапочки прийти в любую из них одинаковая, то примем параметр p = 1/3. Получаем, что:

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

Результатом подсчёта данного выражения будет число 1.58. Это означает, что в среднем (при большом количестве повторений эксперимента) нам понадобится 1.58 вопросов, чтобы верно назвать деревню. При округлении до ближайшего целочисленного даёт значение два. 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).
В целом можем сказать, что в среднем нам хватит двух вопросов, чтобы угадать деревню, а вот одного вопроса может хватить не всегда.

Ответ: 1.58

Задача 2.

Теперь подсчитаем количество возможных вариантов размещения трех агентов: Шапки, бабули и волка. Общее число размещений каждого из трех агентов в трех местах равно 27. Далее, чтобы узнать вероятность хорошего исхода, необходимо знать, во скольких случаях бабушка не получает свои пирожки (то есть Шапка и волк оказываются в одной деревне вместе). Они могут оказаться в одной деревне только вдвоём или с бабулей, но и том, и в другом случае сказка не имеет счастливого конца. Таких вариантов девять. Тогда вариантов, в которых бабуля наслаждается пирожками 18. Энтропия равна:

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

Результатом подсчёта данного выражения будет число 4.17. Это означает, что в среднем ей будет достатчоно 4.17 вопросов, чтобы угадать деревню и доставить пирожки. При округлении до ближайшего целочисленного дает значение пять. Это означает, что в среднем за пять вопросов возможно узнать, получила ли бабушка свои пирожки и в какой деревне она находится, а вот за четыре вопроса это будет сделать сложнее.

Ответ: 4.17

Задача 3.

Наших героев стало уже четверо! Теперь количесвто возможных расположений увеличилось до 81. 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 одной деревне, тогда возможных вариантов расположения охотников и бабушки, при которых Шапка не выживает и бабуля не получает свои пирожки - шесть штук. Тогда, распространив случай на три деревни, получаем, что всего печальных исходов 18. Благоприятных исходов 63. Энтропия в этом случае:

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

Результатом подсчёта данного выражения будет число 5.9.

Ответ: 5.9

Задача 4. Всего вариантов шесть (три деревни и Волка не было, волк был, но его поймали). Так как все три деревни равновероятны, рассмотрим одну из них. Вероятность того, что волка не было:

$$\frac{3-N}{3}$$

Вероятность того, что он был и его поймали:

$$\frac{N}{3} \cdot \frac{1}{3}$$

Одно из событий наступило гарантированно.

$$\frac{3-N}{3} + \frac{N}{9} = \frac{9-2N}{9}$$

Тогда вероятность P(Волк был в деревне A и его поймали):

$$\frac{N}{9} \cdot \frac{9}{(9-2N)\cdot 3} = \frac{N}{(9-2N)\cdot 3}$$

P(Волка не было в деревне A):

$$\frac{(3-N)}{3} \cdot \frac{9}{(9-2N)\cdot 3} = \frac{3-N}{9-2N}$$

Аналогично для других деревень.
Получаем энтропию:

$$H = 3 \cdot \frac{N}{(9-2N)\cdot 3} \cdot \log \frac{(9-2N)\cdot 3}{N} + 3 \cdot \frac{3-N}{9-2N} \cdot \log \frac{9-2N}{3-N}$$

Ответ: см. в решении

Задача 5. Охотники могут поймать волка тремя способами: в А, Б и В. Вероятность каждого из них равна 1/3. Шапочка нашла бабушку тремя способами аналогично (волк ей не мешает, так как его поймали). Вероятность каждого также 1/3 Вероятность совместного исхода 1/9. Итого совместная вероятность:

$$3 \cdot 3 \cdot \frac{1}{9} \cdot \log 9 = \log 9$$

Ответ: log9

Задача 6. Съесть пирожок с капустой можно 20 вариантами (а шапочка считает, что 10). Итого, 10 пирожков с капустой и КШ считает, что они с капустой, а 10 с капустой и КШ не считает их с капустой. Тогда:

$$H = 10 \cdot \frac{1}{20} \cdot \log 10 = \frac{\log 10}{2}$$

Ответ: $\frac{\log 10}{2}$

Задача 7.

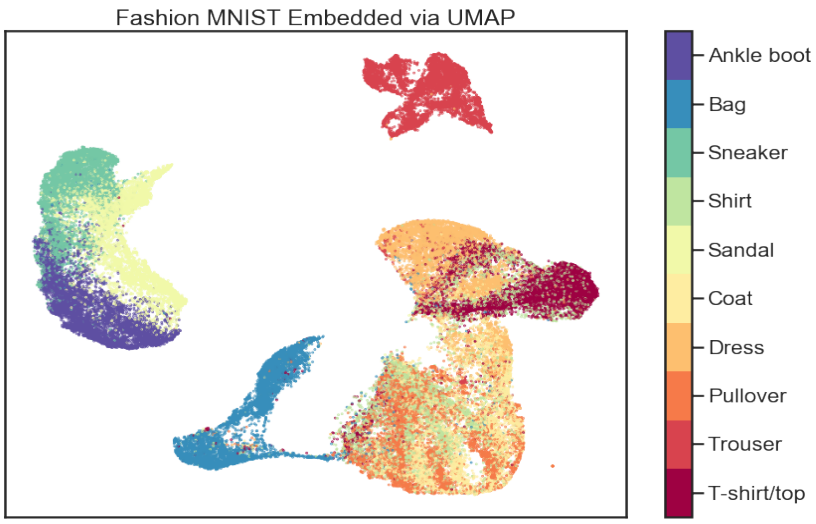


Рис. 1: Алгоритм UMAP