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. Шапочка не помнит, где живет бабушка, но помнит, что равновероятно в одной из трех деревень. Также известно, что в одной из деревень равновероятно находится волк. Если Шапка и волк встретятся, то бабушка не получит пирожки.

1. Посчитайте количество возможных вариантов, в которых бабушка получит пирожки при условии, что волк не поймает Шапку;

2. Посчитайте энтропию события из прошлого пункта.

Еще немножко :)

> **Условная энтропия** — количество бит, необходимое для того, чтобы узнать имеющуюся информацию о случайной величине Y при условии, что случайная величина X 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?

Энтропия (среднее число вопросов, необходимое для того, чтобы понять, кого загадали) рассчитывается так:

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

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

Задача 2. Красная Шапочка должна отнести бабушке пирожки. Шапочка не помнит, где живет бабушка, но помнит, что что равновероятно в одной из трех деревень. Также известно, что около деревень ошивается целых N волков ($N \leqslant 3$), при этом возле каждой деревни либо один волк, либо нет волков. Волков ловит храбрый охотник, он равновероятно находится в одной из трех деревень. Волк поймает Шапку, если рядом не будет охотника.

1. Посчитайте вероятность того, что бабушка получит пирожки при условии того, что волк не поймал Шапку;
2. 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.

Совместная энтропия (среднее число вопросов, необходимое для отгадывания пары людей) рассчитывается так:

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

Задача 3. Красная Шапочка должна отнести бабушке пирожки. Красная Шапочка не помнит, где живет бабушка, но помнит, что в деревне А с вероятностью $1/2$, а в Б и В – с $1/4$. Также известно, что в одной из деревень равновероятно находится волк. Посчитайте совместную энтропию местонахождения бабушки и волка 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:

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

Задача 4. Красная шапочка не помнит, где живет бабушка, но помнит, что либо в А, либо в Б. Также известно, что в одной из деревень А, Б или В равновероятно находится волк. Если Шапка и волк встретятся, то бабушка не получит пирожки. 1. Посчитайте вероятность того, что бабушка получит пирожки при условии, что волк не поймает Шапку; 2. Посчитайте взаимную информацию двух событий: волк не поймал Шапку, и бабушка получила пирожки.

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:

$$D(P\,||\,Q) = H(p) - CH(p,q), \text{ или } CH(p,q) = H(p) + D_{KL}(p\,||\,q)$$

Задача 5. Шапочка думала, что в корзинке 20 пирожков с капустой и 20 — с вареньем, но её 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. 1. Посчитайте кросс энтропию в такой ситуации; 2. Найдите расстояние Кульбака-Лейблера.

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:

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

◇ 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:

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

Задача 6. Красная шапочка идет от дома до бабушки и по дороге садится на пеньки, которые равномерно распределены на отрезке [0, B]. За Шапкой охотится волк, местонахождение которого определяется равномерным распределением на отрезке [0, B]. Известно, что волк съел девочку. Найдите энтропию расстояния, которое Шапке удалось пройти.

Задача 7. Красная шапочка идет от дома до бабушки и по дороге садится на пеньки, которые равномерно распределены на отрезке [0, B]. За ней охотится волк, местонахождение которого определяется равномерным распределением на том же отрезке. Территория от 0 до X охраняется, поэтому если волк и Шапка встретятся до X, девочку спасут. Известно, что волк съел Красную шапочку. Найдите условную энтропию расстояния, которое прошла девочка, при условии X.

Задача 8. Красная шапочка идет от дома до бабушки и по дороге садится на пеньки, которые равномерно распределены на отрезке [0, 10]. За ней охотится волк, местонахождение которого определяется следующей функцией плотности

$$f(X) = \begin{cases} 25x & \text{елси } x \in [0,5] \\ 10 - 0.016x & \text{если } x \in (5,10] \end{cases}$$

Территория от 0 до X (X<10) охраняется, поэтому если волк там окажется, его сразу же поймают. Посчитайте совместную энтропию следующих событий: Шапочка не встретила волка и волка поймали.

Задача 9. Красная шапочка идет от дома до бабушки и по дороге садится на пеньки, которые равномерно распределены на отрезке [0, 10]. За ней охотится волк, местонахождение которого определяется следующей функцией плотности

$$f(X) = \begin{cases} 25x & \text{елси } x \in [0, 5] \\ 10 - 0.016x & \text{если } x \in (5, 10] \end{cases}$$

Территория от 0 до X (X<10) охраняется, поэтому если волк там окажется, его сразу же поймают. Шапка присела на пенек, который находится на расстоянии Y (Y < X). Посчитайте взаимную информацию следующих двух событий: шапочка не встретила волка и волка не поймали.

Задача 10. Красная Шапочка должна отнести бабушке пирожки. 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 падали 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?

Задача 11.

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

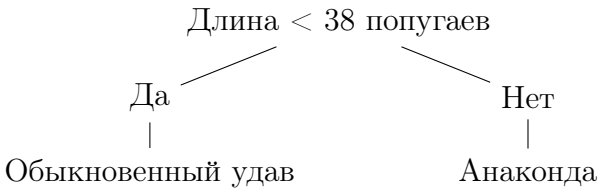
Энтропийное кодирование

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.

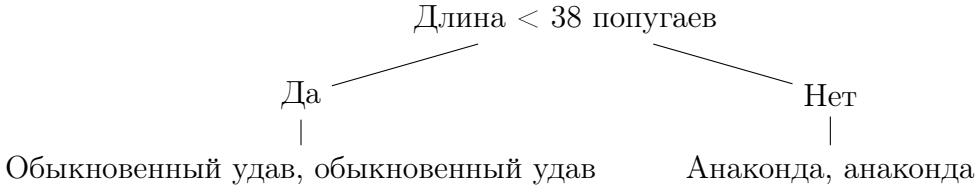
Построение решающих деревьев

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.

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

```
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', 'Shirt', 'Sneaker', 'Bag', 'Ankle boot']
```

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:

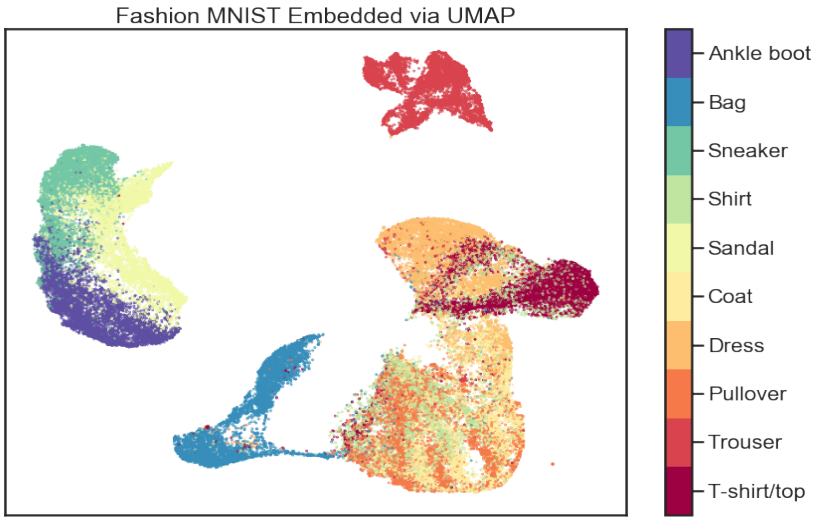


Рис. 1: Алгоритм UMAP

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:

	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. 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 (при большом количестве повторений эксперимента) ей будет достаточно 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: а) 18 вариантов; б) 4.17

Задача 2. 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 3. Волк равновероятно находится в любой из деревень: в А, Б и В. Вероятность каждого из них равна 1/3. Бабушка живет в деревнях с разной вероятностью. Таким образом вероятность того, что они были в деревне А равна 1/6, так как мы перемножили 1/2 и 1/3. Аналогично, вероятность того, что они были в деревнях Б и В равна 1/12.

Итого совместная вероятность:

$$H = -(\frac{1}{6}\log\frac{1}{6} + \frac{1}{12}\log\frac{1}{12} + \frac{1}{12}\log\frac{1}{12})$$

Ответ: $\frac{1}{2} + \frac{1}{3} \cdot \log 3$

Задача 4. Сюрприз! Вам предоставляется невероятная возможность самостоятельно проверить свои силы и решить такую задачу. Варианты решения мы просим присылать в issues на открытый репозиторий на сайте github.com. Первое правильное решение поощряется специальным ценным призом. Ссылка на репозиторий: https://github.com/oomaystrenko/entropy Желаем удачи и с нетерпением ждем Ваших решений!

Задача 5. 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 6.

Задача 7.

Задача 8.

Задача 9.

Задача 10.

Задача 11.