1.1 Перестановки мультимножества

Перестановка мультимножества – это расположение его элементов в ряд. Главное, что нас интересует – это количество таких перестановок. Если дано мультимножество, и известна кратность $n_1,\ldots,n_k$ его элементов, то различных перестановок будет в точности

$$\frac{n!}{n_1!n_2!\ldots n_k!}.$$

Это число называется мультиномиальным коэффициентом и обозначается как

$$\binom{n}{n_1,n_2,\ldots,n_k}.$$

Приведем два полезных разложения этого числа в сумму и произведение. Представление в виде суммы мультиномиальных коэффициентов:

$$\binom{n}{n_1,n_2,\ldots,n_k}=\binom{n-1}{n_1-1,n_2,\ldots,n_k}+\binom{n-1}{n_1,n_2-1,\ldots,n_k}+\ldots+\binom{n-1}{n_1,n_2,\ldots,n_k-1}.$$

Представление в виде произведения биномиальных коэффициентов:

$$\binom{n}{n_1,n_2,\ldots,n_k}=\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\ldots\binom{n_{k-1}+n_k}{n_{k-1}}$$

Справедливость обоих утверждений проверяется непосредственно. Они оба имеют наглядный комбинаторный смысл, о котором будет рассказано позднее.

Скажем несколько слов о происхождении мультиномиальных коэффициентов. Биномиальные коэффициенты возникают при возведении суммы $(x+y)$ в степень n:

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.$$

Аналогично, мультиномиальные коэффициенты возникают при возведении суммы $(x_1+x_2+\ldots+x_m)$ в степень n:

$$(x_1+x_2+\ldots+x_m)^n=\sum_{k_1,k_2,\ldots,k_m}\binom{n}{k_1,k_2,\ldots,k_m}x_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}.$$

1.2 Комбинаторная энтропия.

Определим комбинаторную энтропию для чисел $n_1,n_2,\ldots,n_k$ как

$$H_{comb}=\frac{1}{n}\log_2\frac{n!}{n_1!n_2!\ldots n_k!},\text{ где }n=n_1+n_2+\ldots+n_k.$$

Выясним, какой смысл вкалывается в эту формулу. В ней используется мультиномиальный коэффициент, который есть не что иное как количество перестановок мультимножества или количество строк с заданной статистикой. Все строки с заданной статистикой можно перенумеровать числами от 0 до $\frac{n!}{n_1!n_2!\ldots n_k!}-1$. После этого по числу можно однозначно восстановить строку, а по строке восстановить ее номер. Для передачи номера такой строки в двоичном представлении требуется не более $\log_2\frac{n!}{n_1!n_2!\ldots n_k!}$ бит. Все строки одинаковой длины, поэтому удобно посчитать среднее число бит, которое приходится на один символ строки – поделить количество бит на n. Это и есть комбинаторная энтропия.

1.3 Свойства комбинаторной энтропии

Симметричность.
От перестановки элементов комбинаторная энтропия не меняется.

Максимум.
Значение комбинаторной энтропии максимально, при одинаковых n_i . Чем сильнее n_i отличаются друг от друга (более неравномерное распределение), тем меньше значение комбинаторной энтропии.

Аддитивность.
TODO:

1.4 Энтропия Шеннона

Вычисление значения комбинаторной энтропии (точнее, мультиномиальных коэффициентов) при больших n_i является трудоемким процессом. Однако с помощью формулы Стирлинга для факториала можно построить приближение, которое легко вычислимо:

$$H_{comb} = \frac{1}{n}\log_2\left(\frac{n!}{n_1!n_2!\ldots n_k!}\right) = \frac{1}{n}\log_2\left(\frac{n!}{(p_1 n)!(p_2 n)!\ldots(p_k n)!}\right) \approx$$
$$\approx -\left(p_1\log_2 p_1 + p_2\log_2 p_2 + \ldots + p_k\log_2 p_k\right) = -\sum_{i=1}^{k} p_i\log_2 p_i$$

Имеет место предел:

$$\lim_{n\to\infty} H_{comb} = -\sum_{i=1}^{k} p_i\log_2 p_i \text{ , где } p_i = \lim_{n\to\infty}\frac{n_i}{n}.$$

Такое приближение комбинаторной называется энтропией Шеннона:

$$H = -\sum_{i=1}^{k} p_i\log_2 p_i .$$

Предел справедлив ввиду формулы Стирлинга для представления факториала:

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left\{\frac{1}{12n+1}\right\} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left\{\frac{1}{12n}\right\}$$

Отсюда в частности следует, что

$$\lim_{n\to\infty}\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1$$

$$\frac{n!}{(p_1 N)!(p_2 N)!\ldots(p_k N)!} \to 2^{H(p_1,\ldots,p_k)N}\times(2\pi N)^{(1-n)/2}\left(p_1\ldots p_k\right)^{-1/2}\left(1+O(1/N)\right)$$

Предел на самом деле является пределом снизу, поэтому всегда выполняется:

$$H_{comb} < H .$$

Обычно в литературе сразу приводится формула энтропии Шеннона без каких-либо пояснений. Понимание того, что энтропия Шеннона является пределом комбинаторной энтропии помогает лучше понять ее суть.

Свойства энтропии Шеннона

Непрерывность. Функция $H\left(p_1, p_2, \ldots, p_k\right)$ непрерывна по каждому аргументу.

Симметричность.

$H(p_1, p_2, \ldots, p_k) = H\left(p_{\sigma(1)}, p_{\sigma(2)}, \ldots, p_{\sigma(k)}\right)$, где σ – любая перестановка k элементов.

Максимум достигается, когда $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$, таким образом:

$$H(p_1, p_2, \ldots, p_k) \le H\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$$

Аддитивность.
TODO:

Пример.
Комбинаторная энтропия и энтропия по Шеннону.

Допустим, имеется битовая строка s длины N (четное), в которой число нулей и единиц одинаково и равно $\frac{N}{2}$. В этом случае $H = 1$, потому что распределение равномерное. Рассмотрим комбинаторную энтропию. Всего строк длины N с одинаковым числом нулей и единиц существует $\binom{N}{N/2}$, следовательно, на одну строку необходимо тратить $\log_2 \binom{N}{N/2}$ бит, а в среднем

$$H_{comb} = \frac{1}{N} \log_2 \binom{N}{N/2}.$$

Рассмотрим частный случай при $N = 6$, тогда

$$H_{comb} = \frac{1}{6} \log_2 \binom{6}{6/2} = \frac{1}{6} \log_2 \frac{6!}{3! \cdot 3!} = \frac{\log_2 20}{6} \approx 0.72 .$$

Получаем, что $H_{comb} = 0.72$ меньше, чем $H = 1$. Для кодирования строки s из 6 символов требуется либо 4.32 бита, либо 6 бит при условии, что при декодировании заранее известно, что число нулей равно числу единиц.
Рассмотрим другой случай $N = 10000$, тогда

$$H_{comb} = \frac{1}{10000} \log_2 \binom{10000}{10000/2} = \frac{1}{10000} \log_2 \frac{10000!}{5000! \cdot 5000!} \approx 0.9993 .$$

При больших N комбинаторная энтропия близка к энтропии Шеннона, поэтому выигрыша практически нет.