Энтропия и ее свойства

- Определим энтропию как среднее количество информации, приходящееся на одно сообщение в ансамбле сообщений (или на один символ в отдельном сообщении).
- Иначе говоря, **энтропия – это математическое ожидание количества информации в сообщении.**

- Пусть информационная система может порождать ансамбль (алфавит) сообщений a1, a2,…,am.
- Вероятности каждого сообщения следующие: P(a1), P(a2), …,P(am).
- Вероятности сообщений не одинаковы, то они несут разное количество информации, определяемое формулой Шеннона:

$$I(a_i) = -\log_2 P(a_i).$$

Среднее количество информации или математическое ожидание количество информации

$$H(a) = M[I(a)] = \sum_{i=1}^{m} P(a_i) \cdot I(a_i) = -\sum_{i=1}^{m} P(a_i) \cdot \log_2 P(a_i)$$

- Совершенно аналогично вводится энтропия сообщений:

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

Детерминированность источника означает, что один из возможных символов генерируется источником постоянно (с единичной вероятностью), а остальные – не производятся вовсе

...ва энтропии

- **...ни... ет значение, ...ько в случае ...ованного источника сообщений системы.**
- Доказательство
- Пусть P(ak)=1 , а P(ai)=0 для всех i=1,…,k-1,…

$$h_k = -P(a_k)\log_2 P(a_k) = 0$$

$$h_i = -0 \cdot \log_2 0 = 0 \cdot \infty \, (?)$$

$$\lim_{x\to 0}(-x\log_2 x)=\lim_{x\to 0}x\log_2\frac{1}{x}=\lim_{x\to 0}\frac{\log_2\frac{1}{x}}{\frac{1}{x}}=\left|\frac{1}{x}=z\right|=$$

$$=\lim_{z\to\infty}\frac{\log_2 z}{z}=\lim_{z\to\infty}\frac{(\log_2 z)'}{z'}=\left|(\log_2 z)'=\frac{1}{z\ln 2}\right|=\lim_{z\to\infty}\frac{1}{z\ln 2}=0$$

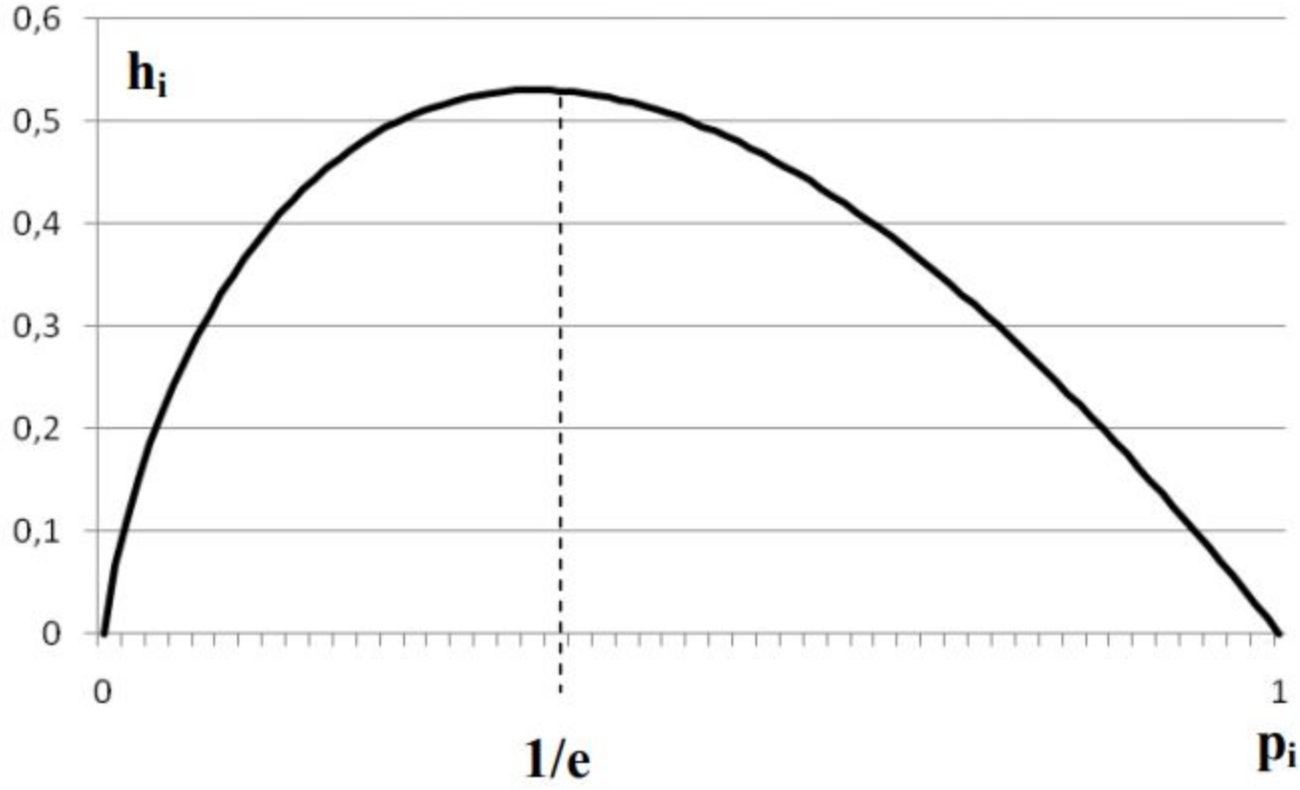
Свойства энтропии

- **2. Энтропия - величина неотрицательная и ограниченная.**
- Если каждое слагаемое **$h_i = -p(a_i)\log_2 p(a_i)$** неотрицательно и ограниченно, то и их сумма также будет неотрицательна и ограниченна.

$$p(a) \geq 0; \quad p(a) \leq 1 \Rightarrow \log p(a) \leq 0 \Rightarrow -p(a)\log p(a) \geq 0$$

$$\frac{\partial h_i}{\partial p_i} = -\frac{\partial p_i}{\partial p_i}\log p_i - p_i\frac{\partial \log p_i}{\partial p_i} = -\log p_i - p_i\frac{1}{p_i \ln 2} = -\log p_i - \log e = 0,$$

отсюда $\mathbf{p_i = 1/e}$.



Свойство энтропии

- **3. Энтропия дискретной системы, имеющей m равновероятных состояний, максимальна и равна log2m.**
- Найдем значение максимальной энтропии. Пусть все символы

$$H_{max} = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \frac{1}{m} = \sum_{i=1}^{m} \frac{1}{m} \log_2 m = \log_2 m$$

- **4.Совместная энтропия независимых источников сообщений равна сумме энтропий.**
- Пусть источник А порождает ансамбль Ma сообщений (a1, a2,…, aMa),
- а источник B порождает ансамбль Mb сообщений (b1, b2,…, bMb), и источники независимы.
- **Общий алфавит источников** представляет собой множество пар вида (ai , bj), общая мощность алфавита равна Ma×Mb. Совместная энтропия композиции двух источников равна

$$H(A,B) = -\sum_{i=1}^{Ma}\sum_{j=1}^{Mb} P(a_i, b_j)\log P(a_i, b_j)$$

Поскольку **A** и **B** независимы, то $\mathbf{P(a_i,b_j) = P(a_i)\cdot P(b_j)}$, а $\mathbf{\log P(a_i,b_j) = \log P(a_i) + \log P(b_j)}$. Отсюда вытекает:

$$\mathbf{H(A,B) = -\sum_{i=1}^{Ma}\sum_{j=1}^{Mb} P(a_i)P(b_j)(\log P(a_i) + \log P(b_j)) =}$$

$$\mathbf{= -\sum_{i=1}^{Ma}\sum_{j=1}^{Mb} P(a_i)P(b_j)\log P(a_i) - \sum_{i=1}^{Ma}\sum_{j=1}^{Mb} P(a_i)P(b_j)\log P(b_j)}$$

$$\mathbf{H(A,B) = -\sum_{i=1}^{Ma} P(a_i)\log P(a_i)\sum_{j=1}^{Mb} P(b_j) - \sum_{j=1}^{Mb} P(b_j)\log P(b_j)\sum_{i=1}^{Ma} P(a_i)}$$

учитывая, что $\mathbf{\sum_{i=1}^{Ma} P(a_i) = 1}$ и $\mathbf{\sum_{j=1}^{Mb} P(b_j) = 1}$, получим

$$\mathbf{H(A,B) = -\sum_{i=1}^{Ma} P(a_i)\log P(a_i) - \sum_{j=1}^{Mb} P(b_j)\log P(b_j) = H(A) + H(B)}$$

Условная энтропия

- Найдем совместную энтропию сложной информационной системы (композиции А, В) в том случае, если их сообщения не являются независимыми, то есть если на содержание сообщения В оказывает влияние сообщение А.

Условная энтропия

- Пусть источник А порождает ансамбль Ma сообщений (a1, a2,…, aMa),
- источник В порождает ансамбль Mb сообщений (b1, b2,…, bMb) и **источники зависимы**.
- Общий алфавит источников представляет собой множество пар вида (ai , bj), общая мощность алфавита: Ma×Mb.

$$\mathbf{H(A,B)=-\sum_{i=1}^{Ma}\sum_{j=1}^{Mb}P(a_i,b_j)\log P(a_i,b_j)}$$

Поскольку **A** и **B** зависимы, то $\mathbf{P(a_i,b_j) = P(a_i)\cdot P(b_j|a_i)}$,

а $\mathbf{\log P(a_i,b_j) = \log P(a_i) + \log P(b_j|a_i)}$. Подставив это в выражение для энтропии сложной системы, получаем:

$$\mathbf{H(A,B)=-\sum_{i=1}^{Ma}\sum_{j=1}^{Mb}P(a_i)P(b_j\,|\,a_i)(\log P(a_i)+\log P(b_j\,|\,a_i))=}$$

$$\mathbf{=-\sum_{i=1}^{Ma}\sum_{j=1}^{Mb}P(a_i)P(b_j\,|\,a_i)\log P(a_i)-\sum_{i=1}^{Ma}\sum_{j=1}^{Mb}P(a_i)P(b_j\,|\,a_i)\log P(b_j\,|\,a_i)}$$

$\mathbf{\sum_{j=1}^{Mb}P(b_j\,|\,a_i)}$, который равен **1**

Во втором слагаемом члены вида $-\sum_{j=1}^{Mb} P(b_j \mid a_i) \log P(b_j \mid a_i) = H(B \mid a_i)$ имеют смысл энтропии источника **B** при условии, что реализовалось сообщение $\mathbf{a_i}$ – будем называть ее **частной условной энтропией**. Если ввести данное понятие и использовать его обозначение, то второе слагаемое будет иметь вид:

$$\sum_{i=1}^{Ma} P(a_i) H(B \mid a_i) = H(B \mid A),$$

или подробнее

$$H(B \mid A) = -\sum_{i=1}^{Ma} P(a_i) \sum_{j=1}^{Mb} P(b_j \mid a_i) \log P(b_j \mid a_i)$$

где **H(B|A)** есть **общая условная энтропия** источника **B** относительно источника **A**. Окончательно получаем для энтропии сложной системы:

$$H(A, B) = H(A) + H(B|A)$$

Свойства условной энтропии

- 1. **Условная энтропия является величиной неотрицательной.**
- Причем $H(B|A) = 0$ только в том случае, если любое сообщение А полностью определяет сообщение В,
- т.е. $H(B|a1) = H(B|a2) = \ldots = H(B|aN) = 0$
- В этом случае $H(A,B) = H(A)$.

$$0 \leq H(B|A) \leq H(B)$$

- **2.Если источники А и В независимы, то H(B|A) = H(B), причем это оказывается наибольшим значением условной энтропии.**
- Другими словами, сообщение источника А не может повысить неопределенность сообщения источника В; оно может либо не оказать никакого влияния (если источники независимы), либо понизить энтропию В.
- **3.H(A, B) ≤ H(A) + H(B), причем равенство реализуется только в том случае, если источники А и В независимы.**

Энтропия источника непрерывных сообщений

- Рассмотрим систему, где качественные признаки состояния изменяются **непрерывно** (непрерывный сигнал).
- **Вероятность нахождения системы в состоянии x (т.е. сигнал принимает значение x) характеризуется плотностью вероятности f(x).**
- Чтобы найти энтропию такого сообщения, разбиваем **диапазон возможного изменения сигнала на дискреты размером Δx.**

$\mathbf{P(x_i) = f(x_i)\cdot\Delta x}$

Тогда энтропия системы вычисляется так:

$$\mathbf{H = -\sum_i f(x_i)\Delta x \log(f(x_i)\Delta x) = -\sum_i f(x_i)\Delta x \cdot (\log f(x_i) + \log \Delta x) =}$$

$$\mathbf{= -\sum_i f(x_i)\log f(x_i)\Delta x - \log \Delta x \cdot \sum_i f(x_i)\Delta x}$$

при малых $\mathbf{\Delta x}$:

$$\mathbf{\sum_i f(x_i)\log f(x_i)\Delta x \approx \int_{-\infty}^{\infty} f(x)\log f(x)dx}$$

А также $\mathbf{\sum_i f(x_i)\Delta x \approx \int_{-\infty}^{\infty} f(x)dx = 1}$

$$\boxed{\mathbf{H = -\int_{\infty}^{\infty} f(x) \log f(x) dx - \log \Delta x}}$$

Положим **Δx=1** (это возможно сделать, выбрав соответствующий масштаб и единицу измерения), тогда $\mathbf{H = H^* = -\int_{-\infty}^{\infty} f(x) \log f(x) dx}$

Величина **H*** называется **приведенной** или **дифференциальной** энтропией.

Количественные характеристики источника сообщений Относительная энтропия

- Соотношение реальных и оптимальных сообщений выражается посредством коэффициента сжатия μ(s) (иное название – относительная энтропия)

$$\mu(s) = H_p(s) / H_0(s) = n_0 / n_p$$

- где $H_p(s)$ и $H_0(s)$ – энтропия реального и идеального источника сообщений соответственно, n_0 и n_p – количество символов оптимального и реального сообщения.

Одно и то же количество информации $\mathbf{I(s)}$ может содержаться в сообщении, состоящим из $\mathbf{n_p}$ символов с энтропией $\mathbf{H_p(s)}$ или из $\mathbf{n_0}$ символов с энтропией $\mathbf{H_0(s)}$

$\mathbf{I(s) = n_p \cdot H_p(s) = n_0 \cdot H_0(s)}$, а так как $\mathbf{H_p(s) \leq H_0(s)}$, то $\mathbf{n_p \geq n_0}$.

Избыточность источника сообщений

- Поскольку реальные источники информации имеют энтропию, меньшую оптимальной, то сообщения таких источников содержат избыточные символы. Коэффициент избыточности φ выража

$$\varphi(s) = \frac{n_p - n_0}{n_p} = 1 - \frac{n_0}{n_p} = 1 - \mu(s)$$

$$\varphi(s) = \frac{H_0 - H_p}{H_0} = 1 - \frac{H_p}{H_0}$$

- !!!Коэффициент избыточности показывает, какая часть реального сообщения является излишней и могла бы не передаваться, если бы источник сообщений был организован оптимально.

Экономичность источников информации

- Существует теоретический оптимум для мощности алфавита. Найдем его.
- !!! При какой мощности алфавита m общая энтропия будет максимальной, если $k \cdot m = const$, где k – количество независимых источников, а m – это мощность алфавита каждого источника? (Под независимыми источниками можно понимать и независимые сигналы одного источника.)

Пусть **k·m = a**. Энтропия композиции независимых источников равна

$$\mathbf{H = \sum_{i=1}^{k} H_i = k \cdot H_i = k \cdot \log m}$$

k = a/m

$$\mathbf{H = \frac{a}{m} \log_2 m}$$

Найдем максимум энтропии, для чего продифференцируем по **m**

$$\mathbf{\frac{\partial H}{\partial m} = \frac{a}{m} \frac{1}{m \ln 2} - \frac{a}{m^2} \log m = \frac{a}{m^2} \log_2 e - \frac{a}{m^2} \log_2 m = 0}$$

$$\mathbf{\frac{a}{m^2} \log_2 e = \frac{a}{m^2} \log_2 m}$$

$\boxed{\mathbf{m = e}}$ Оптимальная мощность алфавита

Производительность источника сообщений

- **Производительностью источника называется количество информации, порождаемое источником в среднем за единицу времени**
- Пусть H – энтропия источника,
- m – мощность алфавита,
- p_i (i=1, 2,…, m) – вероятность появления i-го символа,
- θ_i – длительность генерации i–го символа.

Производительность источника сообщений

- В среднем, один символ генерируется за время $M[\theta]=\sum_{i=1}^{m}\theta_i p_i$.
- На генерацию n символов будет затраче $T=n\cdot M[\theta]$.
- Количество информации

$$I = n\cdot H.$$

- Производительность источника будет вычислена следующим образом:

$$R = \frac{I}{T} = \frac{n \cdot H}{n \cdot M[\theta]} = \frac{-\sum_{i=1}^{m} p_i \log p_i}{\sum_{i=1}^{m} \theta_i p_i}$$

- Если все символы алфавита генерируются за одно и то же время θ,

$$R = \frac{-\sum_{i=1}^{m} p_i \log p_i}{\theta}.$$

Производительность источника сообщений

- Максимальной производительностью обладает источник с максимальной энтропией, которая в соответствии с третьим свойством энтропии равна log2 m:

$$R_{max} = \frac{\log_2 m}{\theta}.$$