Глава 1. Мотивация

Начнём с вопроса, что такое информация? Давайте рассмотрим пример:

- Вопрос: “Температура в Москве сейчас выше 15 градусов?”
  На него возможны ответы либо “да” , либо “нет”.
- Вопрос: “Президент Российской федерации поговорил с определенным человеком в Москве. С кем?”
  На него возможно ответить более чем 10 миллионами способов.

Очевидно, что второй вопрос даёт нам гораздо больше информации, чем первый.

Количество возможных ответов связано с “информацией”.

Посмотрим на другой пример:

- Вы бросаете игральную кость один раз. Есть 6 возможных исходов. Вы записываете полученный исход и говорите о нём другу. Таким образом вы передали своему другу определенный объем информации.
- Вы бросаете игральную кость три раза. Опять же вы записываете все полученные исходы и рассказываете о них другу. Очевидно, что в этом случае вы передали своему другу в три раза больше информации.

“Информация” должна обладать аддитивностью.

Заметим, что во второй ситуации возможны 6^3 исходов, что в 36 раз больше, чем в первой ситуации. Но количество информации возросло всего в три раза. Как с этим быть? Довольно логичным кажется использовать логарифм от количества исходов для того, чтобы измерить количество информации. Именно это в 1928 году предложил американский ученый-электронщик Ральф Хартли:

Определение 1.1. *Мы определим меру информации:*

$$I(U) \triangleq \log_b r, \tag{1}$$

где r это количество всевозможных исходов для случайного сообщения U.

Используя это определение, легко убедиться, что оно удовлетворяет свойству аддитивности:

$$I(U_1, U_2, \ldots, U_n) = \log_b r^n = n \cdot \log_b r = nI(U_1), \tag{2}$$

Хартли также корректно отмечал, что основание логарифма b не имеет большого значения. Оно лишь определяет какие единицы измерения используются. Для некоторых особых значения b есть собственные названия такие единиц измерения:

- $b = 2\,(\log_2)$ – бит;
- $b = e\,(\ln)$ – нат (натуральный логарифм);
- $b = 10\,(\log_{10})$ – Хартли.

С определением данным Хартли есть фундаментальная проблема – согласно нему минимальное ненулевое количество информации это $\log_2 2 = 1$ бит. Может показаться, что это небольшой объем информации, но представим, что мы хотим записать номера всех 7‘621‘000‘000 людей на планете. Согласно данному определению, нам понадобится $\log_2(7‘621‘000‘000) \approx 32.8$ битов. То есть, имея информации всего в 33 раза больше, чем 1 бит, можно раздать каждому человеку уникальный телефонный номер.

Чтобы еще лучше разобраться, в чем проблема, представим что у нас есть два мешка: в первом лежит 2 черных шара и 2 белых; а во втором лежат 3 черных шара и 1 белый. Давайте случайно вытаскивать шар из мешка, и пусть U будет цветом вынутого шара. В каждом мешке есть шары двух цветов, таким образом $I(U_A) = I(U_B) = \log_2 2 = 1$ бит. Но очевидно, что вытаскивая из второго мешка черный шар, мы получаем меньше информации, так как мы изначально ожидаем такого исхода.

Хорошая мера “информации” должна учитывать вероятности возможных исходов.

Впервые к такому выводу пришел американский математик Клод Элвуд Шеннон в 1948 году в статье “A Mathematical Theory of Communication”.

Определение 1.2. *Шенноновская мера информация является “усреднен-*

ной информацией Хартли":

$$\sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{r} p_i \log_2 p_i, \tag{3}$$

где p_i обозначает вероятность i-го возможного исхода.

Глава 2. Энтропия

2.1. Определение

Теперь мы формально определим Шенноновскую меру "информации". В силу связи с похожими концептами в разных разделах физики, Шеннон назвал эту меру *энтропией*.

Определение 2.1. *Энтропия дискретной величины U, которая принимает значения из множества $\mathcal{U}$ (алфавит) определяется как:*

$$H(U) \triangleq -\sum_{u \in supp(P_U)} P_U(u) \log_b P_U(u), \tag{4}$$

где $P_U(\cdot)$ обозначает функцию вероятности случайной величины U, и где носитель P_U определен как:

$$supp(P_U) \triangleq \{u \in \mathcal{U} : P_U(u) > 0\}, \tag{5}$$

Другая часто используемая форма записи:

$$H(U) = E_U[-\log_b P_U(U)], \tag{6}$$

Заметим, что $lim_{t \to 0} t \log_b t = 0$, поэтому во многих случаях мы не будем упоминать носитель при суммировании по $P_U(u)$, подразумевая, что мы исключили все u с нулевой вероятностью.

Также важно отметить, что энтропия случайной величины U никак не зависит от возможных значений U, а только зависит от вероятностей этих значений.

Определение 2.2. *Условная энтропия случайной величины X при условии события $Y = y$ определяется как:*

$$\begin{aligned} H(X|Y = y) &\triangleq -\sum_{x \in supp(P_{X|Y}(\cdot|y))} P_{X|Y}(x|y) \log P_{X|Y}(x|y) \\ &= E[-\log P_{X|Y}(X|Y)|Y = y], \end{aligned} \tag{7}$$

Заметим, что определение идентично предыдущему с единственно разницей, что всё обусловлено на событие $Y = y$.

Определение 2.3. *Условная энтропия случайной величины X при условии случайной величины Y определяется как:*

$$\begin{aligned} H(X|Y) &\triangleq -\sum_{y \in supp(P_Y)} P_Y(y) \cdot H(X|Y=y) \\ &= E_Y[H(X|Y=y)] \\ &= -\sum_{(x,y) \in supp(P_{X,Y})} P_{X,Y}(x,y) \log P_{X|Y}(x|y) \\ &= E[-\log P_{X|Y}(X|Y)], \end{aligned} \tag{8}$$

Заметим, что определение идентично предыдущему с единственно разницей, что всё обусловлено на событие $Y = y$.

2.2. Аксиоматическое определение

Может показаться странным, почему энтропия имеет именно такое определение. Однако в своей исходной статье Шеннон показал, что данное определение энтропии может быть получено естественным путем, приняв за основу определенную систему аксиом. Обозначим вероятностное распределение над m буквами как $P = (p_1, \ldots, p_m)$ и рассмотрим функционал $H_m(p_1, \ldots, p_m)$. Если H_m удовлетворяет аксиомам:

1. Инвариантность относительно перестановок.
2. Расширяемость: $H_m(p_1, \ldots, p_{m-1}, 0) = H_{m-1}(p_1, \ldots, p_{m-1})$.
3. Нормализация: $H_2(\frac{1}{2}, \frac{1}{2}) = \log 2$.
4. Субаддитивность: $H(X, Y) \leq H(X) + H(Y)$.
5. Аддитивность: $H(X, Y) = H(X) + H(Y)$, если X ортоганален Y.
6. Непрерывность: $H_2(p, 1-p) \to 0$ при $p \to 0$.

тогда $H_m(p_1, \ldots, p_m) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$ единственно возможный вариант.

2.3. Свойства

Энтропия Шеннона обладает несколькими важными свойствами:

1. (Неотрицательность). $H(X) \geq 0$, причем равенство возможно тогда и только тогда, когда случайная величина X константна.

2. (Равномерное распределение максимизирует энтропию). Для конечного множества $\mathcal{X}$, $H(X) \leq \log |\mathcal{X}|$, причем равенство возможно тогда и только тогда, когда случайная величина X имеет равномерное распределение над $\mathcal{X}$.
3. (Инвариантность относительно перестановки). $H(X) = H(f(X))$ для любой биекции f.
4. (Малое цепное правило). $H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$
5. (Полное цепное правило). $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \leq \sum_{i=1}^{n} H(X_i)$
6. (Обусловленность снижает энтропию). $H(X|Y) \leq H(X)$, причем равенство возможно тогда и только тогда, когда случайные величины X и Y независимы.

Докажем эти свойства:

1. Матожидание положительной функции также положительно.
2. Минус логарифм является выпуклой функцией на интервале $(0, 1)$, поэтому это следует из неравенства Йенсена.
3. H зависит только от значений P_X, но не от позиции аргументов.
4. $$\begin{aligned} H(X, Y) &= E[\log \frac{1}{P_{X,Y}(X, Y)}] = E[\log \frac{1}{P_X(X) \cdot P_{Y|X}(Y|X)}] \\ &= E[\log \frac{1}{P_X(X)}] + E[\log \frac{1}{P_{Y|X}(Y|X)}] \\ &= H(X) + H(Y|X), \end{aligned} \tag{9}$$
5. Отображение $x \mapsto (x, f(x))$ является биекцией, поэтому:
$$H(X) = H(X, f(X)) = H(f(X)) + H(X|f(X)) \geq H(f(X)), \tag{10}$$

6.

$$
\begin{aligned}
H(X|Y) - H(X) &= E[\log \frac{P_X(X)}{P_{X|Y}(X|Y)}] \\
&= E[\log \frac{P_X(X) \cdot P_Y(Y)}{P_{X|Y}(X|Y) \cdot P_Y(Y)}] \\
&= E[\log \frac{P_X(X) P_Y(Y)}{P_{X,Y}(X,Y)}] \\
&= \sum_{(x,y) \in supp(P_{X,Y})} P_{X,Y}(x,y) \log \frac{P_X(x) P_Y(y)}{P_{X,Y}(x,y)} \\
&\leq \sum_{(x,y) \in supp(P_{X,Y})} P_{X,Y}(x,y) \Big(\frac{P_X(x) P_Y(y)}{P_{X,Y}(x,y)} - 1 \Big) \cdot \log e \\
&= \sum_{(x,y) \in supp(P_{X,Y})} (P_X(x) P_Y(y) - P_{X,Y}(x,y)) \cdot \log e \\
&= \left(\sum_{(x,y) \in supp(P_{X,Y})} P_X(x) P_Y(y) - 1 \right) \cdot \log e \\
&\leq \left(\sum_{x \in X, y \in Y} P_X(x) P_Y(y) - 1 \right) \cdot \log e \\
&= \left(\sum_{x \in X} P_X(x) \sum y \in Y P_Y(y) - 1 \right) \cdot \log e \\
&= (1-1) \log e = 0,
\end{aligned}
\tag{11}
$$

Важно, что при этом $H(X|Y = y)$ может быть как меньше, так и больше $H(X)$.

Глава 3. Совместная информация

3.1. Определение

Наконец мы добрались до понятия информации. Представьте, что у нас есть случайная величина X с энтропией $H(X)$. Как измерить количество информации, которое даёт другая случайная величина Y об X? Логично будет замерить энтропию X до и после того, как мы узнали об Y!

Определение 3.1. *Совместная информация между случайными величинами X и Y определяется как:*

$$I(X;Y) \triangleq H(X) - H(X|Y), \tag{12}$$

Заметим, что это именно совместная информация, а не информация об X, которую даёт Y. Это легко увидеть, дважды воспользовавшись цепным правилом:

$$\begin{aligned} &H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \\ &\Rightarrow H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &\Rightarrow I(X;Y) = I(Y;X), \end{aligned} \tag{13}$$

Аналогично тому, как мы определяли условную энтропию, можно определить и условную совместную информацию. Например:

Определение 3.2.

$$\begin{aligned} I(X;Y|Z) &\triangleq E_Z[I(X;Y|Z=z)] \\ &= \sum_z P_Z(z)(H(X|Z=z) - H(X|Y,Z=z)) \\ &= H(X|Z) - H(X|Y,Z), \end{aligned} \tag{14}$$

3.2. Свойства

Многие свойства совместной информации следуют из свойств энтропии.

Теорема 1. *Пусть X и Y являются случайными величинами с совместной информацией $I(X;Y)$. Тогда:*

$$0 \le I(X;Y) \le \min\{H(X), H(Y)\}, \tag{15}$$

Равенство в левой части достигается тогда и только тогда, когда $P_{X,Y} = P_X \cdot P_Y$. Равенство в правой части достигается тогда и только тогда, когда X определяет Y, либо наоборот.

Доказательство 1. *Так как обусловленность уменьшает энтропию, то:*

$$I(X;Y) = H(Y) - H(Y|X) \geq H(Y) - H(Y) = 0, \tag{16}$$

причем равенство возможно только когда $H(Y|X) = H(Y)$. Чтобы доказать правую часть, воспользуемся неотрицательностью энтропии:

$$\left.\begin{aligned} I(X;Y) = H(X) - H(X|Y) \leq H(X) \\ I(X;Y) = H(Y) - H(Y|X) \leq H(Y) \end{aligned}\right\} \Rightarrow I(X;Y) \leq \min\{H(X), H(Y)\} \tag{17}$$

причем равенство возможно только когда $H(X|Y) = 0$, либо $H(Y|X) = 0$, т.е. либо Y задаёт X, либо наоборот.

Заметим, что совместная информация случайной величины с самой собой будет в точности её энтропия:

$$I(X;X) = H(X) - H(X|X) = H(X), \tag{18}$$