Лекция 1: Данные, информация, энтропия

А. М. Шур

Кафедра алгебры и фундаментальной информатики УрФУ

20 февраля 2017 г.

Данные

Википедия говорит о том, что данные — это совокупность значений количественных и/или качественных переменных величин, сопровождая это картинкой:

Данные

Википедия говорит о том, что данные — это совокупность значений количественных и/или качественных переменных величин, сопровождая это картинкой:

В этом курсе мы говорим об алгоритмах компьютерной обработки данных, а значит

- Важно не происхождение данных, а способ их компьютерного представления;
- Любые данные на физическом уровне — последовательность бит;
- Минимальная "лексическая" (т.е. имеющая смысл) единица данных зависит от типа данных, например

Данные

Википедия говорит о том, что данные — это совокупность значений количественных и/или качественных переменных величин, сопровождая это картинкой:

В этом курсе мы говорим об алгоритмах компьютерной обработки данных, а значит

- Важно не происхождение данных, а способ их компьютерного представления;
- Любые данные на физическом уровне — последовательность бит;
- Минимальная "лексическая" (т.е. имеющая смысл) единица данных зависит от типа данных, например
  - бит для черно-белого изображения (пиксел)
  - пара бит для последовательностей ДНК/РНК (нуклеотид)
  - байт для текста в ASCII-кодировке (символ)
  - пара байт для текста в Unicode (символ)
  - тройка байт для 24-битного цветного изображения (пиксел)
  - 4/8 байт для количественных данных (число)

Сжатие данных

По умолчанию, обрабатываемый блок данных мы называем текстом, обозначаем за T и считаем последовательностью символов (строкой) некоторого конечного алфавита:

- $T \in \Sigma^n$, n (длина T, $|T|$) — велико;
- Σ чаще всего является 256-символьным (байты), иногда бинарным (биты) или, например, размера 2^{32};

Сжатие данных

По умолчанию, обрабатываемый блок данных мы называем текстом, обозначаем за T и считаем последовательностью символов (строкой) некоторого конечного алфавита:

- $T \in \Sigma^n$, n (длина T, $|T|$) — велико;
- Σ чаще всего является 256-символьным (байты), иногда бинарным (биты) или, например, размера 2^{32};

Что такое «сжать T»?

Неформально — заменить T на более короткую строку C, означающую то же самое (в некотором смысле).

Сжатие данных

По умолчанию, обрабатываемый блок данных мы называем текстом, обозначаем за T и считаем последовательностью символов (строкой) некоторого конечного алфавита:

- $T \in \Sigma^n$, n (длина T, $|T|$) — велико;
- Σ чаще всего является 256-символьным (байты), иногда бинарным (биты) или, например, размера 2^{32};

Что такое «сжать T»?

Неформально — заменить T на более короткую строку C, означающую то же самое (в некотором смысле).

Две интерпретации:

- Строгая (сжатие без потерь, lossless compression): T можно однозначно восстановить по C, если известен использованный метод сжатия (сами по себе C и T — данные совершенно различной природы)
- Нестрогая (сжатие с потерями, lossy compression): T и C одинаково воспринимаются человеческими органами чувств (если T — картинка/ звуковая дорожка/ видео, то C имеет тот же тип и пользователь не отличит глазами/ушами T от C)

Сжатие данных

По умолчанию, обрабатываемый блок данных мы называем текстом, обозначаем за T и считаем последовательностью символов (строкой) некоторого конечного алфавита:

- $T \in \Sigma^n$, n (длина T, $|T|$) — велико;
- Σ чаще всего является 256-символьным (байты), иногда бинарным (биты) или, например, размера 2^{32};

Что такое «сжать T»?

Неформально — заменить T на более короткую строку C, означающую то же самое (в некотором смысле).

Две интерпретации:

- Строгая (сжатие без потерь, lossless compression): T можно однозначно восстановить по C, если известен использованный метод сжатия (сами по себе C и T — данные совершенно различной природы)
- Нестрогая (сжатие с потерями, lossy compression): T и C одинаково воспринимаются человеческими органами чувств (если T — картинка/ звуковая дорожка/ видео, то C имеет тот же тип и пользователь не отличит глазами/ушами T от C)

★ Разная философия этих интерпретаций влечет совершенно разную математику. Мы рассматриваем только **сжатие без потерь**.

Сжатие без потерь

Согласно принятой интерпретации,

метод сжатия без потерь — это

функция $enc : \Sigma^* \to \Sigma^*$, преобразующая (кодирующая, сжимающая) произвольный текст T в $enc(T) = C$, которая является обратимой, т.е. биекцией

Сжатие без потерь

Согласно принятой интерпретации,

метод сжатия без потерь — это

функция $enc : \Sigma^* \to \Sigma^*$, преобразующая (кодирующая, сжимающая) произвольный текст T в $enc(T) = C$, которая является обратимой, т.е. биекцией

★ Строк бесконечно много, но строк любой фиксированной длины — конечное число. Поэтому биекция
- либо сохраняет длину любой строки,
- либо одни строки укорачивает, а другие — удлиняет (по принципу Дирихле);

ни то, ни другое не вяжется с представлением о сжатии...

Методов сжатия без потерь не существует??

Сжатие без потерь

Согласно принятой интерпретации,

метод сжатия без потерь — это

функция $enc : \Sigma^* \to \Sigma^*$, преобразующая (кодирующая, сжимающая) произвольный текст T в $enc(T) = C$, которая является обратимой, т.е. биекцией

★ Строк бесконечно много, но строк любой фиксированной длины — конечное число. Поэтому биекция

- либо сохраняет длину любой строки,
- либо одни строки укорачивает, а другие — удлиняет (по принципу Дирихле);

ни то, ни другое не вяжется с представлением о сжатии...

Методов сжатия без потерь не существует??

★ Существуют, поскольку "в реальном мире" T — не любое: компьютерные методы представления данных основаны на простоте и удобстве доступа/анализа/изменения, а отнюдь не на максимальной компактности. Лишь ничтожная часть всех строк длины n могут оказаться на месте T, и нам надо, чтобы эти строки функция enc укорачивала; что функция enc делает со всеми остальными строками — нам не интересно.

Сжатие без потерь

Согласно принятой интерпретации,

метод сжатия без потерь — это

функция $enc : \Sigma^* \to \Sigma^*$, преобразующая (кодирующая, сжимающая) произвольный текст T в $enc(T) = C$, которая является обратимой, т.е. биекцией

★ Строк бесконечно много, но строк любой фиксированной длины — конечное число. Поэтому биекция

- либо сохраняет длину любой строки,
- либо одни строки укорачивает, а другие — удлиняет (по принципу Дирихле);

ни то, ни другое не вяжется с представлением о сжатии...

Методов сжатия без потерь не существует??

★ Существуют, поскольку "в реальном мире" T — не любое: компьютерные методы представления данных основаны на простоте и удобстве доступа/анализа/изменения, а отнюдь не на максимальной компактности. Лишь ничтожная часть всех строк длины n могут оказаться на месте T, и нам надо, чтобы эти строки функция *enc* укорачивала; что функция *enc* делает со всеми остальными строками — нам не интересно.

Позднее мы уточним этот тезис математически...

Информация и ее количество

Информация — абстрактная философская сущность

- связана с данными и знаниями
- может быть передана (коммуницирована)
- может быть приобретена при помощи наблюдения
- может быть закодирована символами/сигналами для передачи/интерпретации
- уменьшает неопределенность, связанную с ситуацией, объектом, etc
- может быть измерена количественно

Информация и ее количество

Информация — абстрактная философская сущность

- связана с данными и знаниями
- может быть передана (коммуницирована)
- может быть приобретена при помощи наблюдения
- может быть закодирована символами/сигналами для передачи/интерпретации
- уменьшает неопределенность, связанную с ситуацией, объектом, etc
- может быть измерена количественно

Количество информации — математическая величина $I(p)$, функция вероятности события, с наступлением (или наблюдением) которого связана информация

- $I(p)$ определена на $(0,1]$ (невозможные события не наступают)
- $I(p)$ строго убывает (чем маловероятнее событие, тем больше информации)
- $I(1)=0$ (наступление достоверного события не содержит информации)
- $\lim_{p\to+0} I(p)=+\infty$ (количество информации в событии не ограничено)
- $I(pq)=I(p)+I(q)$ (информация из независимых источников складывается)
  - для независимых событий A, B имеем $P(AB)=P(A)P(B)$)

★ если еще потребовать, чтобы $I(p)$ была бесконечно дифференцируемой (т.е. “хорошей”), то вариантов не останется: $I(p)=\log_a p$ для некоторого $a<1$

Количество информации

Функции $\log_a p$ при разных $a < 1$ отличаются умножением на константу:

- выбор a — это выбор единицы измерения
- принято $a = 1/2$, т.е. $I(1/2) = 1$
- единичное количество информации (1 бит) содержится в результате броска симметричной монеты
  - как и в наступлении любого события A с $P(A) = 1/2$
- 1 бит информации можно закодировать 1 символом в бинарном алфавите; поэтому бит — это еще и единица объёма данных

Количество информации

Функции $\log_a p$ при разных $a < 1$ отличаются умножением на константу:

- выбор a — это выбор единицы измерения
- принято $a = 1/2$, т.е. $I(1/2) = 1$
- единичное количество информации (1 бит) содержится в результате броска симметричной монеты
  - как и в наступлении любого события A с $P(A) = 1/2$
- 1 бит информации можно закодировать 1 символом в бинарном алфавите; поэтому бит — это еще и единица объёма данных

Обычно пишут $I(p) = -\log p$ (log без индекса по умолчанию двоичный). Иногда для простоты будем писать $I(A)$ вместо $I(P(A))$.

- Если T — текст и $f : \Sigma^* \to \Sigma^*$ — известная функция, то $I(f(T)) \leqslant I(T)$
  - получение $f(T)$ из T не содержит неопределенности; то, что мы получим именно $f(T)$ — достоверное событие
- Если f — биекция, то $I(f(T)) = I(T)$ для любого T

Количество информации

Функции $\log_a p$ при разных $a < 1$ отличаются умножением на константу:

- выбор a — это выбор единицы измерения
- принято $a = 1/2$, т.е. $I(1/2) = 1$
- единичное количество информации (1 бит) содержится в результате броска симметричной монеты
  - как и в наступлении любого события A с $P(A) = 1/2$
- 1 бит информации можно закодировать 1 символом в бинарном алфавите; поэтому бит — это еще и единица объёма данных

Обычно пишут $I(p) = -\log p$ (log без индекса по умолчанию двоичный). Иногда для простоты будем писать $I(A)$ вместо $I(P(A))$.

- Если T — текст и $f : \Sigma^* \to \Sigma^*$ — известная функция, то $I(f(T)) \leqslant I(T)$
  - получение $f(T)$ из T не содержит неопределенности; то, что мы получим именно $f(T)$ — достоверное событие
- Если f — биекция, то $I(f(T)) = I(T)$ для любого T

Более общо,

★ Если $f(T)$ не имеет прообразов кроме T, то $I(f(T)) = I(T)$; если же $f(T) = f(T')$, то знание $f(T)$ оставляет неопределенность, был ли исходный текст равен T, откуда $I(f(T)) < I(T)$.

Информация содержится в наступлении/ненаступлении некоторого события (уменьшение неопределенности)

- ⇒ Свяжем информацию с дискретной случайной величиной $\xi = \{x_{1}|_{p_1}, \ldots, x_k|_{p_k}\}$
- ★ Информация — результат опыта над (подходящей) ξ

Информация и случайные величины

Информация содержится в наступлении/ненаступлении некоторого события (уменьшение неопределенности)

⇒ Свяжем информацию с дискретной случайной величиной $\xi = \{x_{1\,|p_1}, \ldots, x_{k\,|p_k}\}$

★ Информация — результат опыта над (подходящей) ξ

- Если поставить много независимых опытов, количество информации будет суммироваться
- Опыты в дальнейшем называем бросками кубика ξ
- Результат последовательности из n бросков кодируется текстом $T = x_{i_1} x_{i_2} \cdots x_{i_n}$

Информация содержится в наступлении/ненаступлении некоторого события (уменьшение неопределенности)

- ⇒ Свяжем информацию с дискретной случайной величиной $\xi = \{x_1|_{p_1}, \ldots, x_k|_{p_k}\}$
- ★ Информация — результат опыта над (подходящей) ξ
- Если поставить много независимых опытов, количество информации будет суммироваться
- Опыты в дальнейшем называем бросками кубика ξ
- Результат последовательности из n бросков кодируется текстом $T = x_{i_1} x_{i_2} \cdots x_{i_n}$

Это очень простая модель текста, на ней начнем учиться. В более сложных моделях разные броски производятся разными кубиками с одним и тем же набором "граней" $\{x_1, \ldots, x_k\}$; разных кубиков может быть даже потенциально бесконечное число.

Информация содержится в наступлении/ненаступлении некоторого события (уменьшение неопределенности)

$\Rightarrow$ Свяжем информацию с дискретной случайной величиной $\xi = \{x_{1\,|p_1}, \ldots, x_{k\,|p_k}\}$

★ Информация — результат опыта над (подходящей) ξ

- Если поставить много независимых опытов, количество информации будет суммироваться
- Опыты в дальнейшем называем бросками кубика ξ
- Результат последовательности из n бросков кодируется текстом $T = x_{i_1}x_{i_2}\cdots x_{i_n}$

Это очень простая модель текста, на ней начнем учиться. В более сложных моделях разные броски производятся разными кубиками с одним и тем же набором "граней" $\{x_1, \ldots, x_k\}$; разных кубиков может быть даже потенциально бесконечное число.

- В зависимости от того, как выпал кубик ξ, может генерироваться разное количество информации (не все p_i одинаковы)
- Поскольку кубик бросаем много раз, хорошо бы знать среднее количество $\Longrightarrow$

Энтропия

Отметим, что количество информации в одном броске кубика — тоже случайная величина

Отметим, что количество информации в одном броске кубика — тоже случайная величина

Определение

Энтропией дискретной случайной величины ξ называется матожидание количества информации в результате опыта над ξ, обозначаемое $H(\xi)$.

- По формулам матожидания и количества информации, $H(\xi) = -\sum_{i=1}^{k} p_i \log p_i$
- Тонкость возникает, когда $p_i = 0$ для некоторого i и логарифм не существует
    - так бывает: потребуется изучать разные случайные величины, имеющие в качестве исходов разные подмножества одного и того же множества букв
- Так как $\lim_{x\to+0} x \log x = 0$, соответствующее слагаемое в формуле для $H(\xi)$ доопределим его пределом

Энтропия

Отметим, что количество информации в одном броске кубика — тоже случайная величина

Определение

Энтропией дискретной случайной величины ξ называется матожидание количества информации в результате опыта над ξ, обозначаемое $H(\xi)$.

- По формулам матожидания и количества информации, $H(\xi) = -\sum_{i=1}^{k} p_i \log p_i$
- Тонкость возникает, когда $p_i = 0$ для некоторого i и логарифм не существует
  - так бывает: потребуется изучать разные случайные величины, имеющие в качестве исходов разные подмножества одного и того же множества букв
- Так как $\lim_{x\to+0} x \log x = 0$, соответствующее слагаемое в формуле для $H(\xi)$ доопределим его пределом

Теорема (свойства энтропии)

Пусть $\xi = \{x_{1\,|p_1}, \ldots, x_{k\,|p_k}\}$. Тогда

1. $H(\xi) \geqslant 0$, причем $H(\xi) = 0 \iff \xi$ — константа
2. $H(\xi) \leqslant \log k$, причем $H(\xi) = \log k \iff \xi$ равномерно распределена

Свойство (1) очевидно:

- в формуле $H(\xi) = -\sum\limits_{i=1}^{k} p_i \log p_i$ все члены под знаком суммы неположительны и равны нулю только при $p_i = 1$ и $p_i = 0$ (по соглашению о доопределении пределом)

Свойство (1) очевидно:

- в формуле $H(\xi) = -\sum_{i=1}^{k} p_i \log p_i$ все члены под знаком суммы неположительны и равны нулю только при $p_i = 1$ и $p_i = 0$ (по соглашению о доопределении пределом)

Для свойства (2) нужно немного матана:

Лемма 1

Для любого $x > 0$ имеем $\ln x \leqslant x - 1$, причем равенство достигается только при $x = 1$.

График логарифма вогнут (функция $(\ln x)' = 1/x$ убывает), а значит, лежит ниже любой своей касательной, в том числе касательной $y = x - 1$ в точке $x = 1$. □

Доказательство свойств энтропии (продолжение)

Лемма 2

Для любых случайных величин $\xi = \{x_{1|p_1}, \ldots, x_{k|p_k}\}$, $\eta = \{x_{1|q_1}, \ldots, x_{k|q_k}\}$ имеем

$$H(\xi) = -\sum_{i=1}^{k} p_i \log p_i \leqslant -\sum_{i=1}^{k} p_i \log q_i,$$

причем равенство достигается только при $\eta = \xi$.

Доказательство свойств энтропии (продолжение)

Лемма 2

Для любых случайных величин $\xi = \{x_{1\,|\,p_1}, \ldots, x_{k\,|\,p_k}\}$, $\eta = \{x_{1\,|\,q_1}, \ldots, x_{k\,|\,q_k}\}$ имеем

$$H(\xi) = -\sum_{i=1}^{k} p_i \log p_i \leqslant -\sum_{i=1}^{k} p_i \log q_i,$$

причем равенство достигается только при $\eta = \xi$.

$$-\sum_{i=1}^{k} p_i \log q_i + \sum_{i=1}^{k} p_i \log p_i = -\sum_{i=1}^{k} p_i \log \frac{q_i}{p_i} = -(\log e) \sum_{i=1}^{k} p_i \ln \frac{q_i}{p_i} \geqslant_{[\text{Лемма}\,1]}$$

$$-(\log e) \sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) = \log e \left(-\sum_{i=1}^{k} q_i + \sum_{i=1}^{k} p_i\right) = \log e(-1+1) = 0.$$

Равенство возможно только при равенстве в лемме 1, т.е. при $q_i = p_i$ для всех i. $\square$

Доказательство свойств энтропии (продолжение)

Лемма 2

Для любых случайных величин $\xi = \{x_{1|p_1}, \ldots, x_{k|p_k}\}$, $\eta = \{x_{1|q_1}, \ldots, x_{k|q_k}\}$ имеем

$$H(\xi) = -\sum_{i=1}^{k} p_i \log p_i \leqslant -\sum_{i=1}^{k} p_i \log q_i,$$

причем равенство достигается только при $\eta = \xi$.

$-\sum\limits_{i=1}^{k} p_i \log q_i + \sum\limits_{i=1}^{k} p_i \log p_i = -\sum\limits_{i=1}^{k} p_i \log \frac{q_i}{p_i} = -(\log e) \sum\limits_{i=1}^{k} p_i \ln \frac{q_i}{p_i} \geqslant_{[\text{Лемма}\mathbf{1}]}$

$-(\log e) \sum\limits_{i=1}^{k} p_i \Big(\frac{q_i}{p_i} - 1 \Big) = \log e \Big(-\sum\limits_{i=1}^{k} q_i + \sum\limits_{i=1}^{k} p_i \Big) = \log e(-1+1) = 0.$

Равенство возможно только при равенстве в лемме 1, т.е. при $q_i = p_i$ для всех i. □

Взяв в качестве η равномерно распределенную величину ($q_i = 1/k$ для всех i), по лемме 2 получим $H(\xi) \leqslant -\sum\limits_{i=1}^{k} p_i \log \frac{1}{k} = \log k \cdot \sum\limits_{i=1}^{k} p_i = \log k$, причем равенство имеет место только при равенстве в лемме 2, т.е. при $\eta = \xi$. Значит, ξ тоже равномерно распределена. Свойство (2) доказано. □

Энтропия случайного вектора

Теорема

Для любых случайных величин $\xi = \{x_{1|p_1}, \ldots, x_{k|p_k}\}$, $\eta = \{y_{1|q_1}, \ldots, y_{m|q_m}\}$ энтропия $H(\xi,\eta)$ случайного вектора (ξ,η) удовлетворяет неравенству $H(\xi,\eta) \leqslant H(\xi) + H(\eta)$, причем равенство достигается тогда и только тогда, когда η и ξ независимы.

Теорема

Для любых случайных величин $\xi = \{x_1|_{p_1}, \ldots, x_k|_{p_k}\}$, $\eta = \{y_1|_{q_1}, \ldots, y_m|_{q_m}\}$ энтропия $H(\xi,\eta)$ случайного вектора (ξ,η) удовлетворяет неравенству $H(\xi,\eta) \leqslant H(\xi) + H(\eta)$, причем равенство достигается тогда и только тогда, когда η и ξ независимы.

Пусть для любого $i = 1, \ldots, k$, $j = 1, \ldots, m$

- случайный вектор (ξ,η) принимает значение (x_i, y_j) с вероятностью r_{ij}
- вспомогательная случайная величина ζ принимает (x_i, y_j) с вероятностью $p_i q_j$
    - определение ζ корректно, так как $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} p_i q_j = 1$

Теорема

Для любых случайных величин $\xi = \{x_{1|p_1}, \dots, x_{k|p_k}\}$, $\eta = \{y_{1|q_1}, \dots, y_{m|q_m}\}$ энтропия $H(\xi,\eta)$ случайного вектора (ξ,η) удовлетворяет неравенству $H(\xi,\eta) \leqslant H(\xi) + H(\eta)$, причем равенство достигается тогда и только тогда, когда η и ξ независимы.

Пусть для любого $i = 1,\dots,k$, $j = 1,\dots,m$

- случайный вектор (ξ,η) принимает значение (x_i, y_j) с вероятностью r_{ij}
- вспомогательная случайная величина ζ принимает (x_i, y_j) с вероятностью $p_i q_j$
  - определение ζ корректно, так как $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} p_i q_j = 1$

Заметим, что $p_i = \sum\limits_{j=1}^{m} r_{ij}$, $q_j = \sum\limits_{i=1}^{k} r_{ij}$. Имеем

$$H(\xi,\eta) = -\sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}\log r_{ij} \leqslant_{[\text{Лемма2}]} -\sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}\log p_i q_j =$$

$$-\sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}\log p_i - \sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}\log q_j = -\sum_{i=1}^{k} p_i\log p_i - \sum_{j=1}^{m} q_j\log q_j = H(\xi) + H(\eta).$$

При этом равенство выполнено тогда же, когда и в лемме 2, т.е. при $\zeta = (\xi,\eta)$. Но $r_{ij} = p_i q_j$ для любых i,j соответствует независимости ξ и η. $\square$

Теоремы Шеннона

Пусть $\xi = \{x_{1\,|\,p_1}, \ldots, x_{k\,|\,p_k}\}$ — кубик, $T = x_{i_1} x_{i_2} \cdots x_{i_n}$ — сгенерированный им текст:

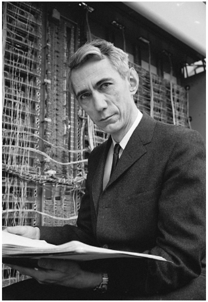
- $P(T) = p_{i_1} p_{i_2} \cdots p_{i_n}$, поскольку броски кубика независимы
- не все тексты равновероятны
- количество вхождений каждого x_i в сгенерированный текст — случайная величина с биномиальным распределением и матожиданием $p_i n$
- ★ наибольшую вероятность имеют те тексты, в которых каждый символ x_i встречается согласно матожиданию, т.е. около $p_i n$ раз

Теоремы Шеннона

Пусть $\xi = \{x_{1\,|\,p_1}, \ldots, x_{k\,|\,p_k}\}$ — кубик, $T = x_{i_1} x_{i_2} \cdots x_{i_n}$ — сгенерированный им текст:

- $P(T) = p_{i_1} p_{i_2} \cdots p_{i_n}$, поскольку броски кубика независимы
- не все тексты равновероятны
- количество вхождений каждого x_i в сгенерированный текст — случайная величина с биномиальным распределением и матожиданием $p_i n$
- ★ наибольшую вероятность имеют те тексты, в которых каждый символ x_i встречается согласно матожиданию, т.е. около $p_i n$ раз

Внимательным разглядыванием предельных теорем теории вероятностей можно вывести гораздо более сильные результаты — теоремы Шеннона



Клод Шеннон внимательно разглядывает тебя, %USERNAME

$\Longrightarrow$

Теоремы Шеннона

Пусть $\xi = \{x_{1\,|p_1}, \ldots, x_{k\,|p_k}\}$ — кубик, $T = x_{i_1}x_{i_2}\cdots x_{i_n}$ — сгенерированный им текст.

Первая теорема Шеннона

Для любых $\varepsilon, \delta > 0$ найдется константа $n_0 = n_0(\varepsilon, \delta)$ такая, что для любого $n \geqslant n_0$ все тексты длины n можно разбить на два класса A и B:

- $\sum_{T\in A} P(T) < \varepsilon$ (текст почти никогда не принадлежит A);
- $\left|\frac{-\log P(T)}{n} - H(\xi)\right| < \delta$ для любого $T \in B$ (все тексты из B имеют близкую вероятность, определяемую энтропией ξ).

Вторая теорема Шеннона

Пусть все тексты T_1, T_2, T_{k^n} длины n упорядочены по убыванию вероятности, а $n(q)$ минимальное число такое, что $\sum_{i=1}^{n(q)} P(T_i) > q$. Тогда $\lim_{n\to\infty} \frac{\log n(q)}{n} = H(\xi)$ для любого $q \in (0, 1)$.

Теоремы Шеннона

Пусть $\xi = \{x_{1\,|p_1}, \ldots, x_{k\,|p_k}\}$ — кубик, $T = x_{i_1} x_{i_2} \cdots x_{i_n}$ — сгенерированный им текст.

Первая теорема Шеннона

Для любых $\varepsilon, \delta > 0$ найдется константа $n_0 = n_0(\varepsilon, \delta)$ такая, что для любого $n \geqslant n_0$ все тексты длины n можно разбить на два класса A и B:

- $\sum_{T \in A} P(T) < \varepsilon$ (текст почти никогда не принадлежит A);
- $\left| \frac{-\log P(T)}{n} - H(\xi) \right| < \delta$ для любого $T \in B$ (все тексты из B имеют близкую вероятность, определяемую энтропией ξ).

Вторая теорема Шеннона

Пусть все тексты T_1, T_2, T_{k^n} длины n упорядочены по убыванию вероятности, а $n(q)$ минимальное число такое, что $\sum_{i=1}^{n(q)} P(T_i) > q$. Тогда $\lim_{n\to\infty} \frac{\log n(q)}{n} = H(\xi)$ для любого $q \in (0, 1)$.

★ согласно теоремам, при больших n можно считать, что существует лишь $2^{H(\xi)n}$ текстов из k^n возможных ($2^{H(\xi)} = k$ лишь при максимальной энтропии) и каждый текст имеет одну и ту же вероятность $2^{-H(\xi)n}$

Кодирование текста

Рассмотрим ξ с вероятностями $p_1 = \cdots = p_k$. Пусть $k = 2^j$; как правильно отобразить (закодировать) $x_1, \ldots, x_k$ в бинарный алфавит, чтобы

- матожидание длины кодирующей строки (“кодового слова”) было минимальным
- все кодовые слова были различными; более того,

★ любая бинарная строка, являющаяся конкатенацией кодовых слов, допускала единственное разбиение на такие слова?

Кодирование текста

Рассмотрим ξ с вероятностями $p_1 = \cdots = p_k$. Пусть $k = 2^j$; как правильно отобразить (закодировать) $x_1, \ldots, x_k$ в бинарный алфавит, чтобы

- матожидание длины кодирующей строки ("кодового слова") было минимальным
- все кодовые слова были различными; более того,
- ★ любая бинарная строка, являющаяся конкатенацией кодовых слов, допускала единственное разбиение на такие слова?

Существует очевидное решение: каждому x_i сопоставим уникальную бинарную строку длины j. Тогда матожидание длины кодового слова равно j $(= H(\xi))$. Можно ли добиться лучших результатов?

Кодирование текста

Рассмотрим ξ с вероятностями $p_1 = \cdots = p_k$. Пусть $k = 2^j$; как правильно отобразить (закодировать) $x_1, \ldots, x_k$ в бинарный алфавит, чтобы

- матожидание длины кодирующей строки ("кодового слова") было минимальным
- все кодовые слова были различными; более того,
- ★ любая бинарная строка, являющаяся конкатенацией кодовых слов, допускала единственное разбиение на такие слова?

Существует очевидное решение: каждому x_i сопоставим уникальную бинарную строку длины j. Тогда матожидание длины кодового слова равно j $(= H(\xi))$. Можно ли добиться лучших результатов?

- Алгебраическая теория кодов (см. Лекцию 3 про неравенство Крафта–Макмиллана) говорит, что нет (единственность разбиения!)
- ★ по теоремам Шеннона, тексты длины n, порожденные любой случайной величиной ξ, распределены почти равномерно (на множестве B, см. первую теорему); значит, любой способ кодирования ξ не может преодолеть "энтропийный предел": матожидание длины кодового слова не будет меньше $H(\xi)$

Рассмотрим ξ с вероятностями $p_1 = \cdots = p_k$. Пусть $k = 2^j$; как правильно отобразить (закодировать) $x_1, \ldots, x_k$ в бинарный алфавит, чтобы

- матожидание длины кодирующей строки ("кодового слова") было минимальным
- все кодовые слова были различными; более того,
- ★ любая бинарная строка, являющаяся конкатенацией кодовых слов, допускала единственное разбиение на такие слова?

Существует очевидное решение: каждому x_i сопоставим уникальную бинарную строку длины j. Тогда матожидание длины кодового слова равно j $(= H(\xi))$. Можно ли добиться лучших результатов?

- Алгебраическая теория кодов (см. Лекцию 3 про неравенство Крафта–Макмиллана) говорит, что нет (единственность разбиения!)
- ★ по теоремам Шеннона, тексты длины n, порожденные любой случайной величиной ξ, распределены почти равномерно (на множестве B, см. первую теорему); значит, любой способ кодирования ξ не может преодолеть "энтропийный предел": матожидание длины кодового слова не будет меньше $H(\xi)$

Как ни странно, существует сверхэффективный способ кодирования, при котором любой текст длины n, порожденный любой случайной величиной ξ, кодируется бинарным словом длины $\lceil H(\xi) \cdot n \rceil$. Он обсуждается в Лекции 5.

Задача сжатия данных

Дан текст T длины n (на физическом уровне текст — это последовательность бит, но что мы считаем символами, из которых состоит текст — наше дело). Требуется

1. Найти случайную величину ξ, такую что
   - T — "типичный" текст, порожденный ξ ($T \in B$ в терминологии первой теоремы Шеннона)
   - чем меньше энтропия ξ, тем лучше
2. Закодировать T как результат последовательности бросков кубика ξ

Задача сжатия данных

Дан текст T длины n (на физическом уровне текст — это последовательность бит, но что мы считаем символами, из которых состоит текст — наше дело). Требуется

1. Найти случайную величину ξ, такую что
   - T — "типичный" текст, порожденный ξ ($T \in B$ в терминологии первой теоремы Шеннона)
   - чем меньше энтропия ξ, тем лучше
2. Закодировать T как результат последовательности бросков кубика ξ

Первая задача называется моделированием, а вторая — кодированием (дискретной информации). Наша цель —

- научиться оптимально решать задачу кодирования, в том числе при различных ограничениях
- научиться "хорошо" решать задачу моделирования, дать обзор существующих классов моделей и соответствующих алгоритмов