

Теория информации

Михаил Михайлов

2026

Оглавление

1	Введение в теорию информации	3
1.1	Введение в теорию информации	4
1.1.1	Теория информации: предмет, мотивация и исторический контекст	4
1.1.2	Модель коммуникации Шеннона–Уивера	5
1.1.3	Собственная информация	7
1.2	Энтропия	9
1.2.1	Понятие об опыте. Энтропия Шеннона	9
1.2.2	Неравенство Гиббса. Оценки энтропии	9
	Приложение. О различных смыслах энтропии	10
	Приложение. Неравенство Йенсена	12
1.3	Условная и совместная энтропия	13
1.3.1	Совместная и условная собственная информация	13
1.3.2	Условная энтропия.	14
1.3.3	Совместная энтропия	16

Введение в теорию информации

Меня больше всего беспокоило, как это назвать. Я думал назвать это «информацией», но это слово использовалось слишком часто, поэтому я решил назвать это «неопределённостью». Когда я обсуждал это с Джоном фон Нейманом, у него возникла идея получше. Фон Нейман сказал мне: «Вам следует называть это энтропией по двум причинам. Во-первых, ваша функция неопределенности использовалась в статистической механике под этим именем, поэтому у неё уже есть имя. Во-вторых, что более важно, никто на самом деле не знает, что такое энтропия, поэтому в дебатах у вас всегда будет преимущество»

Клод Шеннон, «отец» теории информации.

Оглавление

1.1	Введение в теорию информации	4
1.1.1	Теория информации: предмет, мотивация и исторический контекст	4
1.1.2	Модель коммуникации Шеннона–Уивера	5
1.1.3	Собственная информация	7
1.2	Энтропия	9
1.2.1	Понятие об опыте. Энтропия Шеннона	9
1.2.2	Неравенство Гиббса. Оценки энтропии	9
	Приложение. О различных смыслах энтропии	10
	Приложение. Неравенство Йенсена	12
1.3	Условная и совместная энтропия	13
1.3.1	Совместная и условная собственная информация	13
1.3.2	Условная энтропия.	14
1.3.3	Совместная энтропия	16

1.1. Введение в теорию информации

Оглавление

1.1.1	Теория информации: предмет, мотивация и исторический контекст	4
1.1.2	Модель коммуникации Шеннона–Уивера	5
1.1.3	Собственная информация	7

1.1.1. Теория информации: предмет, мотивация и исторический контекст

Теория информации — это раздел математики, возникший в середине XX века в ответ на инженерные и научные задачи передачи сообщений. Её цель состоит в том, чтобы количественно описать информацию и установить фундаментальные ограничения на её представление, сжатие и передачу. Классические вопросы, на которые пытается ответить теория информации, можно сформулировать следующим образом:

- как измерить количество информации в сообщении;
- каковы предельные возможности сжатия данных без потерь;
- как надёжно передавать сообщения по зашумлённым каналам;
- как связаны информация, случайность и неопределённость.

Но что есть информация? Слово *информация* происходит от латинского *informatio*, что означало «формирование образа, представления». Информация может передаваться, способна наделять знанием и обуславливать действия, обладать количественным измерением. Изначально оно описывало процесс превращения знания в структуру, которую можно осознать.

Современная теория информации берёт своё начало в работе К. Шеннона 1948 года, где была предложена математическая модель связи, позволившая отделить физические характеристики канала от абстрактной структуры сообщений. Существенным шагом было осознание того, что смысл передаваемых сообщений не играет роли при анализе предельных возможностей передачи. Именно такой отказ сделал возможным строгий количественный анализ.

Центральная идея теории информации состоит в том, что информация связана с *неопределённостью* исхода. Чтобы прояснить эту связь, полезно рассматривать информацию как меру изменения нашего знания о системе. До получения сообщения мы допускаем несколько возможных состояний мира; после получения сообщения множество допустимых состояний сокращается. Чем сильнее это сокращение, тем более информативным является сообщение.

Пример. Рассмотрим подбрасывание честной монеты. До эксперимента возможны два равновероятных исхода. После наблюдения результата неопределённость исчезает полностью. В этом смысле сообщение о результате подбрасывания монеты устраняет одну условную единицу неопределённости.

Если исход был практически предопределён заранее, то соответствующее сообщение не приводит к заметному изменению нашего знания.

Пример. Сообщение «завтра взойдёт солнце» практически не несёт информации, так как вероятность этого события до получения сообщения была близка к единице.

Иногда про информацию думают как про меру «удивительности» сообщения.

Пример. В городе “П” солнечная погода может быть только в августе, а во все остальные месяца погода пасмурная каждый день. Сообщение «Сегодня в городе “П” пасмурно» не несет никакой информации. В тоже время, более удивительное (поскольку более редкое) сообщение «Сегодня в городе “П” солнечно» сразу позволяет нам заключить что сегодня — август.

Такая интерпретация информации тесно связана с вероятностным описанием источника сообщений. Наши ожидания относительно возможных исходов естественным образом выражаются через вероятности. Маловероятные события оказываются более неожиданными и, следовательно, более информативными. Эта идея лежит в основе количественного определения информации, предложенного Шенноном.

Существует и другой, более алгоритмический взгляд на информацию. Интуитивно сообщение можно считать информативным, если его трудно описать короткой инструкцией. Например, последовательность из тысячи нулей легко описывается фразой «тысяча нулей подряд», тогда как случайная на вид двоичная строка той же длины не допускает столь компактного описания и требует перечисления почти всех символов. В этом смысле информация связана с минимальной длиной описания объекта.

Пример. Результат подбрасывания честной монеты можно передать одним битом. Результат броска честного кубика требует большего числа битов, так как число возможных исходов выше и ни один из них не выделен заранее.

Хотя в рамках курса мы не будем формализовывать этот алгоритмический подход, он тесно связан с вероятностной теорией информации и во многом дополняет её. В обоих случаях информация отражает отсутствие простоты, предсказуемости или сжатого описания.

1.1.2. Модель коммуникации Шеннона–Уивера

Первым, кто предложил рассматривать вопросы, связанные с информацией, независимо от её семантического содержания, был Клод Шеннон. Вместе с Уивером он занимался анализом процессов передачи информации, стремясь отделить математические свойства сообщений от их смысла. Для формального исследования передачи информации они предложили универсальную модель коммуникации, которая до сих пор служит фундаментом теории информации. Схема этой модели представлена на рис. 1.1.



Рис. 1.1: Модель Шеннона–Уивера

Модель включает пять основных компонентов, которые описывают полный цикл передачи информации:

- **Источник информации** — объект, генерирующий последовательность символов или сообщений. Источник может быть текстовым, аудио- или видеосигналом, результатом случайного эксперимента или любой другой дискретной последовательностью или функцией от времени.
- **Передатчик (кодирующее устройство)** — устройство или алгоритм, преобразующий сообщение в форму, пригодную для передачи по каналу. Кодирование здесь не связано с шифрованием, оно служит для представления информации так, чтобы её можно было надёжно и эффективно передавать.
- **Канал/источник шума** — среда передачи сообщений, которая может вносить ошибки, искажения или задержки. Каналом может быть например оптоволоконный кабель или электромагнитное поле по которому распространяются радиоволны.
- **Приёмник (декодирующее устройство)** — устройство, восстанавливающее исходное сообщение из полученного сигнала. Основная задача декодера — точно восстановить исходное сообщение.
- **Получатель** — конечный адресат, для которого предназначено сообщение. В терминах модели именно получатель фиксирует информацию и измеряет её эффект на уменьшение неопределённости.

Смысл этой модели в том, что информацию можно измерять количественно, независимо от её содержания. Выход источника рассматривается как случайная величина с определённым распределением вероятностей. Каждое сообщение уменьшает неопределённость относительно исхода, а структура модели позволяет оценить, сколько информации реально передаётся и как её оптимально кодировать.

Пример. Рассмотрим источник, генерирующий буквы A и B с равными вероятностями (у каждого из символов вероятность 0.5). Передатчик кодирует символ A как 0, B как 1. Если канал идеален, приёмник восстанавливает сообщение без ошибок, и каждая буква передаёт одну битовую единицу информации. Если канал зашумлён, некоторые биты могут быть изменены случайным образом. Приёмник получает сигнал с неопределённостью, и задача кодера и декодера состоит в минимизации потерь информации и вероятности ошибки с помощью специальных кодов.

Следующий важный элемент модели — это понятие **сигнала**. Сигнал представляет собой физическую или математическую форму, в которой информация передаётся по каналу. Основное различие делается между дискретными и аналоговыми сигналами. Дискретные сигналы принимают конечное число значений. Примером служит последовательность битов в цифровом сообщении. Аналоговые сигналы непрерывны во времени и амплитуде, как звуковая волна или радиосигнал. Их передача требует дискретизации и квантования, чтобы можно было использовать цифровые методы обработки и кодирования.

Замечание. Квантование и дискретизация — это два процесса преобразования аналогового сигнала для цифровой обработки и передачи.

Дискретизация заключается в разбиении непрерывного по времени сигнала на отдельные моменты (выборки), фиксируя его значения через равные интервалы времени. Этот шаг делает сигнал «дискретным по времени» и позволяет работать с ним как с последовательностью отдельных измерений.

Квантование означает аппроксимацию непрерывных значений амплитуды каждой выборки конечным числом уровней. В результате амплитуда, которая изначально может принимать любое значение, заменяется ближайшим доступным уровнем, делая сигнал дискретным по амплитуде.

Более детальная классификация сигналов учитывает дискретность по времени и уровню амплитуды (квантование). Различают следующие виды сигналов:

- **Непрерывные (аналоговые) сигналы** — непрерывны по времени и амплитуде. Пример: чистый аналоговый радиосигнал. Для передачи таких сигналов используют аналоговые каналы или дискретизацию для цифровой обработки.
- **Дискретно-непрерывные сигналы** — дискретны по времени, амплитуда непрерывна. Пример: показания датчика, снятые через равные интервалы времени. Требует квантования для передачи по цифровому каналу.
- **Непрерывно-квантованные сигналы** — непрерывны по времени, амплитуда дискретизована. Пример: аналоговый сигнал с периодической выборкой и квантованием для цифровой передачи.
- **Дискретно-квантованные (цифровые) сигналы** — дискретны по времени и амплитуде. Пример: цифровой аудиопоток, последовательность битов в сети. Такие сигналы «легко» анализировать и кодировать.

На рис. 1.2 изображены различные формы представления одного и того же аналогового сигнала.

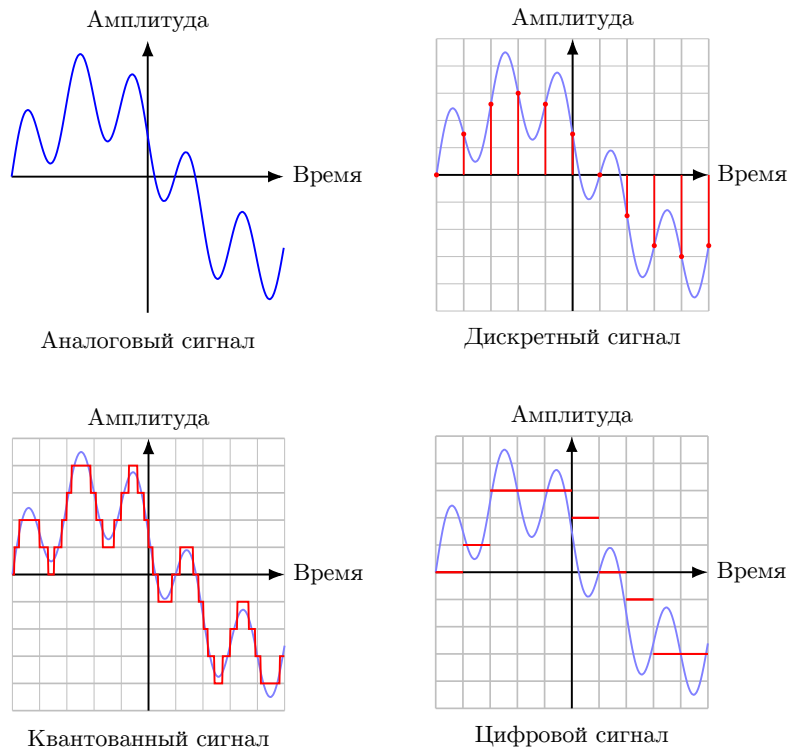


Рис. 1.2: Виды сигналов

Пример. Несколько примеров сигналов.

- Сигнал микрофона, подаваемый напрямую на усилитель без оцифровки, является аналоговым сигналом: и время, и амплитуда изменяются непрерывно.
- Температурный датчик, снимающий показания каждую секунду, создаёт дискретно-непрерывный сигнал. Значения температуры меняются непрерывно, но фиксируются через равные промежутки времени.
- Температурный датчик, передающий показания непрерывно, но с округлением до доли градуса, наоборот является примером источника квантованного сигнала.
- Цифровая аудиозапись в формате WAV является примером цифрового сигнала: звук оцифровывается с определённой частотой дискретизации и квантуется по амплитуде.

В рамках модели Шеннона–Уивера также можно выделить основные классы задач передачи сообщений:

- **Сжатие без потерь** — минимизация объёма передачи без потери информации. Используется, когда необходимо восстановить исходное сообщение точно.
- **Сжатие с потерями** — уменьшение объёма передачи с допустимой потерей деталей, например, для аудио- и видеосигналов.
- **Избыточное кодирование** — добавление дополнительных битов или структур, позволяющих обнаруживать и исправлять ошибки, возникающие в канале.

Давайте пока в качестве «toy model» рассмотрим случай цифрового сигнала и на его основе познакомимся с базовыми идеями теории информации.

1.1.3. Собственная информация

Пусть по каналу связи передаётся одно из заранее заданных сообщений s_1, \dots, s_k . Предполагается, что сообщение s_i передаётся с вероятностью p_i . Каждому событию, состоящему в получении того или иного сообщения, мы хотим сопоставить числовую величину, называемую *собственной информацией* этого события.

Интуитивно ясно, что редкое сообщение должно нести больше информации, чем частое: получение почти гарантированного исхода мало что сообщает получателю, тогда как маловероятный исход существенно сокращает неопределённость. Эти соображения формализуются в виде следующих аксиом.

Аксиомы собственной информации. Пусть A — случайное событие. Функция $\mathcal{I}(A)$, измеряющая информацию, должна удовлетворять условиям:

1. **Неотрицательность:** $\mathcal{I}(A) \geq 0$ для любого события A с положительной вероятностью;
2. **Монотонность:** при увеличении вероятности $\mathbb{P}(A)$ величина $\mathcal{I}(A)$ убывает;
3. **Аддитивность:** для независимых событий A и B

$$\mathcal{I}(A \cap B) = \mathcal{I}(A) + \mathcal{I}(B).$$

Эти требования отражают базовые свойства информации: она не может быть отрицательной, уменьшается для более ожидаемых событий и суммируется при независимом объединении источников неопределённости.

Замечание. При этом еще одно требование, уже не столь естественное — чтобы из равенства $\mathbb{P}(A) = \mathbb{P}(B)$ следовало равенство $\mathcal{I}(A) = \mathcal{I}(B)$. Иначе говоря, информация зависит только от вероятности события. Может показаться странным, что мы оцениваем информацию, не учитывая семантику событий. Однако если в примере из начала положить $s_1 = "0"$, $s_2 = "1"$ (т.е. по каналу передаётся один бит), то в ситуации, когда символы равновероятны, сообщение “0” будет настолько «удивительным», насколько сообщение “1”.

Следующий результат показывает, что выбранные аксиомы практически однозначно определяют вид функции собственной информации.

Единственность функции собственной информации

Теорема 1.1.1 (Единственность функции собственной информации). *Существует единственная (с точностью до умножения на неотрицательную константу) непрерывная функция $\mathcal{I}(A)$, зависящая только от вероятности события $\mathbb{P}(A)$ и удовлетворяющая аксиомам 1–3.*

Для доказательства этой теоремы потребуется вспомогательная лемма.

Лемма 1.1.2. Пусть $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ непрерывная и для любых $x, y \in \mathbb{R}_+$ выполнено равенство:

$$f(x + y) = f(x) + f(y) \tag{1.1}$$

Тогда $f(x) = cx$ для некоторого $c \in \mathbb{R}$.

Доказательство. Заметим, что в силу 1.1 для любого $x \in \mathbb{R}_+$ и для любого $n \in \mathbb{N}$ имеем:

$$f(nx) = f(x) + f((n-1) \cdot x) = \dots = n \cdot f(x).$$

Применяя это равенство к $x = \frac{y}{m}$, где $y \in \mathbb{R}_+$ и для любого $m \in \mathbb{N}$ получаем что:

$$f(y) = f\left(m \cdot \frac{y}{m}\right) = mf\left(\frac{y}{m}\right) \Rightarrow f\left(\frac{y}{m}\right) = \frac{1}{m}f(y)$$

Таким образом для любых $n, m \in \mathbb{N}$ и для любого $x \in \mathbb{R}$:

$$f\left(\frac{n}{m}x\right) = \frac{n}{m}f(x)$$

Положим $c = f(1)$. Докажем что $f(x) = cx$. Пусть $\{q_i\}_{i=1}^{+\infty}$, $q_i = \frac{n_i}{m_i}$, $n_i, m_i \in \mathbb{N}$ — последовательность рациональных чисел, такая что $q_i \xrightarrow{n \rightarrow +\infty} x$. По непрерывности f :

$$f(q_i) \xrightarrow{n \rightarrow +\infty} f(x)$$

С другой стороны, в силу доказанного ранее, $f(q_i) = q_i f(1) = c \cdot q_i$. Откуда

$$cx = \lim_{i \rightarrow +\infty} (c \cdot q_i) = f(x)$$

□

Доказательство теоремы. Аддитивность (аксиома 3) для независимых событий означает, что для любых $p, q \in (0, 1]$ должно выполняться

$$\mathcal{I}(pq) = \mathcal{I}(p) + \mathcal{I}(q),$$

где $\mathcal{I}(p)$ обозначает информацию события вероятности p . Рассмотрим $f(t) = \mathcal{I}(\exp(-t))$:

$$f(t_1 + t_2) = f(t_1) + f(t_2)$$

И значит $f(t) = ct$ по лемме 1.1.2. Тогда, поскольку \mathcal{I} монотонна (аксиома 2)

$$\mathcal{I}(\exp(-t)) = ct \Rightarrow \mathcal{I}^{-1}(ct) = \exp(-t) \Rightarrow \mathcal{I}^{-1}(t) = \exp\left(-\frac{t}{c}\right) \Rightarrow \mathcal{I}(p) = -c \ln p$$

□

Замечание. На самом деле, непрерывность в формулировке теоремы 1.1.1 (и леммы 1.1.2) избыточна и достаточна только монотонность.

Таким образом, аксиоматический подход естественным образом приводит к логарифмической шкале измерения информации.

Определение 1.1.3 (Собственная информация события). Собственной информацией события A с вероятностью $\mathbb{P}(A)$ называется величина

$$\mathcal{I}_m(A) = -\log_m \mathbb{P}(A),$$

где $m > 1$ — положительная константа.

Замечание. Выбор основания логарифма определяет единицы измерения информации: при основании 2 информация измеряется в битах, при основании e — в натах. Далее, если не оговорено иное, будем использовать логарифмы по основанию 2 и обозначать $\mathcal{I} = \mathcal{I}_2$.

Рассмотрим несколько простых ситуаций, иллюстрирующих введённое определение.

Пример (Честная монета). Подбрасывается честная монета. Возможны два исхода: орёл O и решка R , причём

$$\mathbb{P}(O) = \mathbb{P}(R) = 0.5.$$

Собственная информация каждого исхода равна

$$\mathcal{I}(O) = \mathcal{I}(R) = -\log_2 0.5 = 1 \text{ бит.}$$

Получение любого исхода полностью устраняет неопределённость относительно результата броска.

Пример (Нечестная монета). Пусть монета выпадает орлом с вероятностью 0.8 и решкой с вероятностью 0.2. Тогда

$$\mathcal{I}(O) = -\log_2 0.8 \approx 0.32 \text{ бита,} \quad \mathcal{I}(R) = -\log_2 0.2 \approx 2.32 \text{ бита.}$$

Редкий исход (решка) несёт существенно больше информации, чем ожидаемый.

1.2. Энтропия

Оглавление

1.2.1	Понятие об опыте. Энтропия Шеннона	9
1.2.2	Неравенство Гиббса. Оценки энтропии	9
	Приложение. О различных смыслах энтропии	10
	Приложение. Неравенство Йенсена	12

Ранее мы рассматривали собственную информацию отдельных событий. Теперь перейдём к полным системам событий, описывающим исходы некоторого случайного опыта/эксперимента.

1.2.1. Понятие об опыте. Энтропия Шеннона

Определение (Случайный опыт). *Под случайным опытом понимается конечная система непересекающихся событий*

$$\mathcal{A} = \{A_1, \dots, A_n\}, \quad A_i \cap A_j = \emptyset, \quad i \neq j, \quad \bigcup_{i=1}^n A_i = \Omega,$$

где Ω — пространство элементарных исходов, $\mathbb{P}(A_i) > 0, i = 1, \dots, n$. События A_i называются результатами или исходами такого опыта.

Замечание. Данное определение не является конвенциональным, хотя встречается в литературе. Мы используем его только в рамках данного курса, чтобы не писать каждый раз «полная система событий».

Для полной системы событий естественно определить следующую величину:

Определение 1.2.1 (Энтропия случайного опыта). *Энтропией* случайного опыта $\mathcal{A} = \{A_1, \dots, A_n\}$ называется средняя собственная информация результата опыта:

$$H_m(\mathcal{A}) = \sum_{i=1}^n \mathbb{P}(A_i) \cdot \mathcal{I}_m(A_i) = - \sum_{i=1}^n p_i \log_m p_i.$$

Замечание. Как и ранее, если нижний индекс опускается, подразумевается логарифм по основанию 2, $H(\cdot) = H_2(\cdot)$

Пример (Честная монета). Случайный опыт $\mathcal{A} = \{O, R\}$, $\mathbb{P}(O) = \mathbb{P}(R) = 0.5$:

$$H(\mathcal{A}) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1 \text{ бит.}$$

Пример (Нечестная монета). Случайный опыт $\mathcal{A} = \{O, R\}$, $\mathbb{P}(O) = 0.8$, $\mathbb{P}(R) = 0.2$:

$$H(\mathcal{A}) = -(0.8 \log_2 0.8 + 0.2 \log_2 0.2) \approx 0.72 \text{ бита.}$$

Нотация. $H_m(p_1, \dots, p_n)$ будет обозначать энтропию Шеннона для некоторого опыта, результаты которого имеют вероятности p_1, \dots, p_n .

1.2.2. Неравенство Гиббса. Оценки энтропии

Неравенство Гиббса

Лемма 1.2.2 (Неравенство Гиббса). *Для любых двух наборов чисел p_1, \dots, p_n и q_1, \dots, q_n в полуинтервале $(0; 1]$, таких что:*

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n q_i \leq 1,$$

и для любого вещественного $m > 1$ выполнено неравенство:

$$- \sum_{i=1}^n p_i \log_m p_i \leq - \sum_{i=1}^n p_i \log_m q_i \quad (1.2)$$

Причем равенство достигается тогда и только тогда, когда $q_i = p_i$ для всех $i = 1, \dots, n$

пропорциональна $-\log_2$ его вероятности. Это в некотором смысле объясняет, как же информация может быть нецелым числом: в среднем на один символ можно «экономить» часть бита, хотя отдельный символ требует целого бита.

Однако привязка к теории кодирования несколько неудобна, поскольку ограничивает наши способности к обобщению; в частности, на непрерывный случай. Оказывается, что энтропия имеет физический смысл. В первую очередь, она является прямым аналогом и обобщением термодинамической энтропии. Чтобы понять её суть представим себе следующее. Пускай есть некоторая кристаллическая решетка, находящаяся в состоянии термодинамического равновесия, в частности, имеющая некоторую стабилизированную температуру. Будем считать что пока есть только один параметр — температура. Тогда значение температуры определяет *макросостояние* системы. С другой стороны, температура обеспечивается за счет колебаний частиц формирующих решетку; каждая отдельная частица решетки колеблется с частотой из некоторого дискретного набора; полное описание такой системы вида «частица»–«частота колебания» составляет *микросостояние* системы. Одному макросостоянию отвечает множество разных микросостояний, реализующих его. Энтропия Больцмана определяется как

$$S = k_B \ln W,$$

где W — число микросостояний, реализующих данное макросостояние. Однако такое определение корректно для ситуации термодинамического равновесия, когда все микросостояния равновероятны. Обобщение на ситуацию произвольных вероятностей микросостояний было предложено Гиббсом:

$$S_{\text{Гиббса}} = -k_B \sum_i p_i \ln p_i,$$

где p_i — вероятность i -ого микросостояния. При измерении в битах (делении на $k_B \ln 2$) энтропия Гиббса даёт формулу Шеннона:

$$H = \frac{S_{\text{Гиббса}}}{k_B \ln 2}.$$

В этом контексте энтропия Шеннона измеряет *априорный* «объём» нашего незнания о точном микроскопическом состоянии системы при известных макроскопических параметрах. Чем больше возможных микросостояний (конфигураций) совместимо с нашими наблюдениями, тем выше неопределённость (энтропия) и тем больше информации мы получим, узнав точное состояние системы. Внезапно, эта точка зрения довольно тесно перекликается с теорией вероятности. Мы можем считать, что и «микросостояния» — это элементы полной системы событий \mathcal{A} , описывающей различные уровни «энергии», иначе говоря, различные микросостояния системы. Тогда, с точностью до константы, энтропия Шеннона и энтропия Гиббса совпадают.

В этом смысле, энтропию можно понимать как меру дезорганизации, сложности или информационной ёмкости системы. Физическая система с низкой энтропией (например, кристаллическая решётка при абсолютном нуле или выстроенный в линию газ) является высокоорганизованной и предсказуемой: зная положение одной молекулы, можно с высокой уверенностью предсказать положение соседней. Сообщение о её состоянии будет коротким. Система с высокой энтропией (газ в равновесии, хаотичная структура) максимально неупорядочена и непредсказуема — для точного описания её конфигурации потребуется огромное количество информации. Таким образом, физический процесс увеличения термодинамической энтропии в изолированной системе соответствует стиранию различий и переходу к более «типичному», а значит, и более вероятному с информационной точки зрения состоянию, которое описывается большим числом бит.

Приложение. Неравенство Йенсена

Теорема (Неравенство Йенсена). Пусть $f(x): E \rightarrow \mathbb{R}$, $E = \langle a; b \rangle \subset \mathbb{R}$ — выпуклая функция. Тогда для любых точек $x_1, \dots, x_n \in E$, и любых чисел $p_1, \dots, p_n \in (0; 1)$ таких, что $\sum_{i=1}^n p_i = 1$, выполнено неравенство:

$$f\left(\sum_{i=1}^n p_i x_i\right) \geq \sum_{i=1}^n p_i f(x_i) \quad (1.3)$$

Причем равенство достигается тогда и только тогда, когда:

1. либо все точки x_i равны некоторому $x_0 \in E$;
2. либо функция f линейна² на некотором отрезке E' содержащем все точки x_i .

Доказательство. Для случая $n = 2$ неравенство (1.3) есть определение выпуклой функции. Покажем что, если для каких-то $p_1, p_2 \in (0; 1)$, $p_1 + p_2 = 1$ неравенство обратилось в равенство, то f будет линейна на соответствующем отрезке³. Не умаляя общности, пусть $x_1 < x_2$, $p_1 = p$, $p_2 = 1 - p$, $p \in (0; 1)$. И пусть:

$$f(px_1 + (1 - p)x_2) = pf(x_1) + (1 - p)f(x_2)$$

Рассмотрим функцию

$$g(x) = (x - x_1) \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

Имеем:

1. $g(x)$ линейная функция из $E' = [x_1, x_2]$ в \mathbb{R} ;
2. $g(x_1) = f(x_1)$, $g(x_2) = f(x_2)$ и для всех точек $x \in E'$ $f(x) \geq g(x)$
3. $d(x) = f(x) - g(x)$ — выпуклая функция.

Пункты 2 и 3 легко проверить по определению выпуклой функции. Итак, обозначим $x^* = px_1 + (1 - p)x_2 \in E'$. Тогда имеем что $d(x_1) = d(x_2) = d(x^*) = 0$. С другой стороны, если $f(x)$ нелинейна, то $f(x)$ строго больше $g(x)$ хотя бы в одной точке. Не умаляя общности, пусть это точка $\hat{x} \in (x_1; x^*)$. Тогда для некоторого $\lambda \in (0; 1)$ верно что $x^* = \lambda \hat{x} + (1 - \lambda)x_2$. Но тогда, в силу выпуклости d :

$$0 = d(x^*) = d(\lambda \hat{x} + (1 - \lambda)x_2) \leq \lambda d(\hat{x}) + (1 - \lambda)d(x_2) = \lambda \cdot d(\hat{x}) > 0.$$

Противоречие и значит $f(x) = g(x)$ для всех $x \in E'$.

Покажем теперь неравенство для произвольного n . Для этого воспользуемся математической индукцией с уже доказанной базой $n = 2$. Считая, что неравенство доказано для любых наборов точек размера $n - 1$, рассмотрим произвольный набор из n точек $x_1 \leq x_2 \leq \dots \leq x_n$. Положим

$$\begin{aligned} y_1 &= x_1, & q_1 &= p_1 \\ y_2 &= \frac{p_2 x_2 + \dots + p_n x_n}{p_2 + \dots + p_n}, & q_2 &= 1 - q_1 \end{aligned}$$

По определению выпуклости имеем что

$$f\left(\sum_{i=1}^n p_i x_i\right) = f(q_1 y_1 + q_2 y_2) \geq q_1 f(y_1) + q_2 f(y_2) = p_1 f(x_1) + (p_2 + \dots + p_n) f\left(\frac{1}{(p_2 + \dots + p_n)} \sum_{i=2}^n p_i x_i\right)$$

Применяя ко второму слагаемому индукционное предположение

$$\geq p_1 f(x_1) + \sum_{i=2}^n p_i f(x_i) = \sum_{i=1}^n p_i f(x_i)$$

Что и требовалось доказать. Случай равенства разбирается аналогично $n = 2$ — все неравенства выше обращаются в равенства и из этого получается что f линейна на $[x_1, x_n]$. \square

²Везде и далее более правильно говорить аффинная: т.е. $f(x) = ax + b$ для некоторых $a, b \in \mathbb{R}$.

³Случай равенства $x_1 = x_2$ в силу его тривиальности

1.3. Условная и совместная энтропия

Оглавление

1.3.1	Совместная и условная собственная информация	13
1.3.2	Условная энтропия.	14
1.3.3	Совместная энтропия	16

1.3.1. Совместная и условная собственная информация

В реальных задачах часто приходится иметь дело не с одиночными событиями, а с их комбинациями и зависимостями. Для этого вводятся понятия совместной и условной информации.

Определение 1.3.1. Пусть A и B — два события, $\mathbb{P}(A \cap B) > 0$. Совместной информацией событий A и B называется собственная информация пересечения:

$$\mathcal{I}_m(A, B) = \mathcal{I}_m(A \cap B).$$

Условной информацией события A при условии события B называется величина:

$$\mathcal{I}_m(A | B) = -\log \mathbb{P}(A | B).$$

Нотация. $\mathcal{I}_m(A_1, \dots, A_n) = \mathcal{I}_m(A_1 \cap \dots \cap A_n)$

Нотация. $\mathcal{I}_m(A | B_1, \dots, B_n) = \mathcal{I}_m(A | B_1 \cap \dots \cap B_n)$

Предложение. Пусть A и B — два события, $\mathbb{P}(A \cap B) > 0$.

$$\mathcal{I}_m(A | B) = \mathcal{I}_m(A) + \mathcal{I}_m(B | A) = \mathcal{I}_m(B) + \mathcal{I}_m(A | B).$$

Доказательство. Не умаляя общности $m = 2$. Распишем по определению:

$$\mathcal{I}(A \cap B) = \log \mathbb{P}(A \cap B) = \log(\mathbb{P}(A) \cdot \mathbb{P}(A | B)) = \log \mathbb{P}(A) + \log \mathbb{P}(A | B) = \mathcal{I}(B) + \mathcal{I}(A | B)$$

□

Пример (Два независимых броска монеты). Пусть O_1 — событие выпадения орла при первом броске честной монеты, а O_2 — аналогичное событие для второго броска. Имеем

$$\mathbb{P}(O_1) = \mathbb{P}(O_2) = 0.5,$$

причём события независимы. Тогда

$$\mathcal{I}(O_1, O_2) = \mathcal{I}(O_1) + \mathcal{I}(O_2) = 1 + 1 = 2 \text{ бита.}$$

Пример (Зависимые события). Рассмотрим следующую модель: первый бросок монеты честный, а второй полностью определяется первым. Если в первом броске выпал орёл, то во втором обязательно выпадает решка, и наоборот. Тогда

$$\mathbb{P}(O_1) = \mathbb{P}(R_1) = 0.5.$$

Собственная информация первого броска равна $\mathcal{I}(O_1) = 1$ бит, а условная информация второго броска при известном первом исходе равна

$$\mathcal{I}(R_2 | O_1) = -\log 1 = 0.$$

Следовательно,

$$\mathcal{I}(R_2 | O_1) = 1 \text{ бит.}$$

Пример (Последовательность символов). Пусть источник независимо генерирует символы a, b, c с вероятностями

$$\mathbb{P}(a) = 0.5, \quad \mathbb{P}(b) = 0.25, \quad \mathbb{P}(c) = 0.25.$$

Для последовательности abc

$$\mathbb{P}(abc) = 0.03125, \quad \mathcal{I}(a, b, c) = 5 \text{ бит.}$$

Если получены только первые два символа,

$$\mathcal{I}(a, b) = 3 \text{ бита.}$$

1.3.2. Условная энтропия.

Здесь и далее, речь идет о нескольких опытах сразу. Пусть имеются два опыта

$$\mathcal{A} = \{A_1, \dots, A_{n_a}\} = \{A_{i_a}\}_{i_a=1}^{n_a}, \quad \mathcal{B} = \{B_1, \dots, B_{n_b}\} = \{B_{i_b}\}_{i_b=1}^{n_b},$$

Для простоты обозначений, мы будем считать что с каждым опытом связан свой индекс. В частности, для опыта \mathcal{A} вероятности исходов $\mathbb{P}(A_{i_a}), i_a = 1, \dots, n_a$ мы будем обозначать за p_{i_a} , а для опыта \mathcal{B} — вероятности $\mathbb{P}(B_{i_b}), i_b = 1, \dots, n_b$ обозначим за p_{i_b} .

Замечание. Из-за такого злоупотребления обозначениями (abuse of notation), формально запись p_1 обозначает непонятно что. Но таких обозначений далее не будет.

Определение 1.3.2 (Условная энтропия). Пусть $p_{i_b|i_a} = \mathbb{P}(B_{i_b} | A_{i_a})$. Условная энтропия опыта \mathcal{B} относительно опыта \mathcal{A} определяется как

$$H_m(\mathcal{B} | \mathcal{A}) = \sum_{i_a=1}^{n_a} \mathbb{P}(A_{i_a}) \sum_{i_b=1}^{n_b} \mathbb{P}(B_{i_b} | A_{i_a}) \mathcal{I}_m(B_{i_b} | A_{i_a}) = - \sum_{i_a=1}^{n_a} p_{i_a} \sum_{i_b=1}^{n_b} p_{i_b|i_a} \log_m p_{i_b|i_a}.$$

Предложение 1.3.3. Для любых двух опытов \mathcal{A} и \mathcal{B} выполнено:

$$0 \leq H_m(\mathcal{B} | \mathcal{A}) \leq H_m(\mathcal{B})$$

Причем:

- равенство $H_m(\mathcal{B} | \mathcal{A}) = H_m(\mathcal{B})$ достигается тогда и только тогда, когда любые два события $A \in \mathcal{A}, B \in \mathcal{B}$ независимы;
- равенство $H_m(\mathcal{B} | \mathcal{A}) = 0$ достигается тогда и только тогда, когда для любых событий $A \in \mathcal{A}, B \in \mathcal{B}$, выполнено $\mathbb{P}(B | A) \in \{0, 1\}$.

Доказательство. Не умаляя общности $m = 2$.

Нижняя граница. Следует из того что каждое слагаемое неотрицательно. Рассмотрим случай равенства: имеем что

$$\text{Для любых } i_a = 1, \dots, n_a, i_b = 1, \dots, n_b : p_{i_a} p_{i_b|i_a} \log p_{i_b|i_a} = 0$$

Откуда возможны три случая:

1. $p_{i_a} = 0$;
2. $p_{i_b|i_a} = 0$;
3. $\log p_{i_b|i_a} = 0$, т.е. $p_{i_b|i_a} = 1$.

Первый случай невозможен по определению опыта. Откуда получаем что $p_{i_b|i_a} \in \{0, 1\}$, что и требовалось.

Верхняя граница. Воспользуемся неравенством Гиббса (лемма 1.2.2)

$$\begin{aligned} -H(\mathcal{B} | \mathcal{A}) &= \sum_{i_a=1}^{n_a} p_{i_a} \sum_{i_b=1}^{n_b} p_{i_b|i_a} \log p_{i_b|i_a} \geq \sum_{i_a=1}^{n_a} p_{i_a} \sum_{i_b=1}^{n_b} p_{i_b|i_a} \log p_{i_b} = \\ &= \sum_{i_a=1}^{n_a} \sum_{i_b=1}^{n_b} \underbrace{p_{i_a} p_{i_b|i_a}}_{p_{i_b} p_{i_a|i_b}} \log p_{i_b} = \sum_{i_b=1}^{n_b} p_{i_b} \log p_{i_b} \underbrace{\sum_{i_a=1}^{n_a} p_{i_a|i_b}}_1 = -H(\mathcal{B}) \end{aligned}$$

Откуда получаем что $H(\mathcal{B} | \mathcal{A}) \leq H(\mathcal{A})$. Рассмотрим случай равенства. Так как доказательство верхней оценки применяет неравенство Гиббса к n_a суммам, случай равенства означает что для любого $i_a = 1, \dots, n_a$ выполнено равенство:

$$\sum_{i_b=1}^{n_b} p_{i_b|i_a} \log p_{i_b|i_a} = \sum_{i_b=1}^{n_b} p_{i_b|i_a} \log p_{i_b}.$$

Из этого как известно следует что $p_{i_b|i_a} = p_{i_b}$, что и означает что для любых $A \in \mathcal{A}, B \in \mathcal{B}$, события A и B независимы; \square

Пример (Случайная монета). Рассмотрим опять модель броска монетки. Пусть наугад выбирается монета: с вероятностью $\frac{1}{2}$ она честная, с вероятностью $\frac{1}{2}$ она фальшивая и выпадает всегда одной стороной — для определенности скажем что решкой. Можем рассмотреть два опыта: первый состоит в наблюдении первого броска монеты и имеет

два результата, $\mathcal{A}_1 = \{O_1, P_1\}$. Второй состоит в наблюдении второго броска монеты и также имеет два результата, $\mathcal{A}_2 = \{O_2, P_2\}$. Нетрудно вычислить вероятности событий: $\mathbb{P}(O_1) = \mathbb{P}(O_2) = \frac{1}{4}$, $\mathbb{P}(P_1) = \mathbb{P}(P_2) = \frac{3}{4}$ и значит:

$$H(\mathcal{A}_1) = H(\mathcal{A}_2) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}$$

С другой стороны в среднем наблюдение первого броска снижает нашу неопределенность в вопросе фальшивости монеты — как минимум, если при первом броске выпал орёл то монетка заведомо честная. Поэтому:

$$H(\mathcal{A}_2 | \mathcal{A}_1) = -\frac{3}{4} \left(\frac{3}{4} \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4} \right) - \frac{1}{4} \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right)$$

Определение 1.3.4 (Независимость опытов). Два опыта \mathcal{A} и \mathcal{B} называются независимыми, если любая пара результатов $A \in \mathcal{A}$ и $B \in \mathcal{B}$ суть независимые события.

Замечание. Таким образом $H(\mathcal{B} | \mathcal{A}) = H(\mathcal{B})$ тогда и только тогда, когда опыты \mathcal{A} и \mathcal{B} независимы.

Определение 1.3.5. Будем говорить что опыт \mathcal{B} определяется опытом \mathcal{A} если для любых событий $A \in \mathcal{A}, B \in \mathcal{B}$, выполнено что $\mathbb{P}(B | A) \in \{0, 1\}$.

Замечание. Таким образом $H(\mathcal{B} | \mathcal{A}) = 0$ тогда и только тогда, когда опыт \mathcal{A} определяет опыт \mathcal{B} .

Замечание. Почему такое определение естественно? Оказывается что если опыт \mathcal{B} определяется опытом \mathcal{A} , то для любого A_{i_a} результата опыта \mathcal{A} существует ровно один результат $B_{r(i_a)}$ опыта \mathcal{B} , такой что $\mathbb{P}(B_{r(i_a)} | A_{i_a}) = 1$; Это следует из того что:

$$\sum_{i_b=1}^{n_b} \mathbb{P}(B_{i_b} | A_{i_a}) = 1$$

Поэтому зная результат опыта \mathcal{A} можно достоверно предсказать результат опыта \mathcal{B} .

Пример (Два броска монеты). Пусть $\mathcal{A}_1 = \{O_1, P_1\}$ и $\mathcal{A}_2 = \{O_2, P_2\}$ — два опыта состоящие в броске монеты. Если броски независимы и монета честная, то:

$$H(\mathcal{A}_1) = H(\mathcal{A}_1 | \mathcal{A}_2) = 1 \text{ бит}, \quad H(\mathcal{A}_2) = H(\mathcal{A}_2 | \mathcal{A}_1) = 1 \text{ бит}.$$

Если второй бросок полностью зависит от первого (например, всегда противоположен первому), то

$$H(\mathcal{B} | \mathcal{A}) = H(\mathcal{A} | \mathcal{B}) = 0 \text{ бит}.$$

Аналогично тому как вводится понятие независимости в совокупности для событий⁴, можно ввести понятие независимых в совокупности опытов.

Определение 1.3.6 (Независимость опытов в совокупности). Опыты $\mathcal{A}_1, \dots, \mathcal{A}_r$ называются независимыми в совокупности, если для любого поднабора $\mathcal{A}_{j_1}, \dots, \mathcal{A}_{j_k}$ любые события $A_1 \in \mathcal{A}_{j_1}, \dots, A_k \in \mathcal{A}_{j_k}$ будут независимыми в совокупности.

Замечание. Определение можно упростить: $\mathcal{A}_1, \dots, \mathcal{A}_r$ называются независимыми в совокупности, если для любого поднабора $\mathcal{A}_{j_1}, \dots, \mathcal{A}_{j_k}$ и любых событий $A_1 \in \mathcal{A}_{j_1}, \dots, A_k \in \mathcal{A}_{j_k}$:

$$\mathbb{P}(A_1 \cap \dots \cap A_k) = \mathbb{P}(A_1) \cdot \dots \cdot \mathbb{P}(A_k)$$

Так как равенство выполнено для любых наборов событий (A_1, \dots, A_k) из разных опытов, то в частности выполнено для любого поднабора.

Пример. Пусть честная монетка независимо бросается r раз. Обозначим за $O_i(P_i)$, $i = 1, \dots, r$ событие «при i -ом броске выпал(а) орёл (решка)». Рассмотрим следующие опыты:

$$\begin{aligned} \mathcal{A}_i &= \{O_i, P_i\}, \quad i = 1, \dots, r \\ \mathcal{A}_0 &= \{T_0, T_1\}, \quad T_0 — \text{число выпавших орлов чётно}, T_1 — \text{нечётно} \end{aligned}$$

Тогда опыты $\mathcal{A}_1, \dots, \mathcal{A}_r$ независимы в совокупности, а если добавить к ним \mathcal{A}_0 , то независимость в совокупности пропадёт и останется только попарная независимость⁵.

⁴И для случайных величин

⁵Проверьте

1.3.3. Совместная энтропия

Определение 1.3.7 (Совместная энтропия). Пусть $p_{i_a i_b} = \mathbb{P}(A_{i_a} \cap B_{i_b})$. Совместная энтропия опытов \mathcal{A} и \mathcal{B} определяется как:

$$H_m(\mathcal{A}, \mathcal{B}) = - \sum_{i_a=1}^{n_a} \sum_{i_b=1}^{n_b} \mathbb{P}(A_{i_a}, B_{i_b}) \cdot \mathcal{I}_m(A_{i_a}, B_{i_b}) = - \sum_{i_a=1}^{n_a} \sum_{i_b=1}^{n_b} p_{i_a i_b} \log_m p_{i_a i_b}.$$

Мы можем думать о совместной энтропии, как об энтропии опыта, состоящего в одновременном проведении опытов \mathcal{A} и \mathcal{B} . Формально это можно выразить следующим образом.

Определение 1.3.8 (Произведение двух опытов). Пусть имеются два опыта:

$$\mathcal{A} = \{A_{i_a}\}_{i_a=1}^{n_a}, \quad \mathcal{B} = \{B_{i_b}\}_{i_b=1}^{n_b},$$

Их произведением называется опыт:

$$\mathcal{A} \wedge \mathcal{B} = \{A_{i_a} \cap B_{i_b} \mid \mathbb{P}(A_{i_a} \cap B_{i_b}) > 0\}_{i_a=1, \dots, n_a, i_b=1, \dots, n_b}.$$

Тогда получим что $H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A} \wedge \mathcal{B})$. Естественным образом можно теперь ввести совместную энтропию нескольких опытов:

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1 \wedge \dots \wedge \mathcal{A}_r).$$

Пример (Случайная монета, продолжение - 1). Вернемся к модели, где выбор между настоящей и фальшивой монетой делается наугад. Два опыта: первый состоит в наблюдении того, какая была вытащена монета: $\mathcal{B} = \{H, \Phi\}$. Второй состоит в наблюдении первого броска монеты и также имеет два результата, $\mathcal{A}_1 = \{O_1, P_1\}$. Их произведение при этом будет состоять из трёх событий:

$$\mathcal{A}_1 \wedge \mathcal{B} = \{H \cap O_1, \Phi \cap P_1, H \cap P_1\}$$

Поскольку при вытаскивании фальшивой монеты, орёл выпасть не может. Совместная энтропия при этом равна:

$$H(\mathcal{A}_1, \mathcal{B}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{2} \log_2 \frac{1}{2} = 1.5 \text{ бит}$$

Предложение 1.3.9 (Правило цепочки). Для двух случайных опытов \mathcal{A} и \mathcal{B} выполняется соотношение

$$H_m(\mathcal{A}, \mathcal{B}) = H_m(\mathcal{A}) + H(\mathcal{B} \mid \mathcal{A}) = H_m(\mathcal{B}) + H_m(\mathcal{A} \mid \mathcal{B}).$$

Доказательство. Не умаляя общности $m = 2$. По определению совместной энтропии:

$$H(\mathcal{A}, \mathcal{B}) = - \sum_{i_a, i_b} p_{i_a i_b} \log p_{i_a i_b} = - \sum_{i_a, i_b} p_{i_a i_b} \log(p_{i_a} p_{i_b \mid i_a}) = - \sum_{i_a} p_{i_a} \log p_{i_a} - \sum_{i_a} p_{i_a} \sum_{i_b} p_{i_b \mid i_a} \log p_{i_b \mid i_a} = H(\mathcal{A}) + H(\mathcal{B} \mid \mathcal{A}).$$

Симметрично, меняя местами \mathcal{A} и \mathcal{B} , получаем второе равенство. \square

Пример (Случайная монета, продолжение - 2). Можем убедиться, рассматривая модель, где выбор между настоящей и фальшивой монетой делается наугад. Как и ранее, рассматриваем два опыта: первый состоит в наблюдении того, какая была вытащена монета: $\mathcal{B} = \{H, \Phi\}$. Второй состоит в наблюдении первого броска монеты и также имеет два результата, $\mathcal{A}_1 = \{O_1, P_1\}$.

По примеру ранее $H(\mathcal{A}_1 \mid \mathcal{B}) = 1.5$ бит. Энтропия $H(\mathcal{B}) = 1$ бит. Посчитаем:

$$H(\mathcal{A}_1 \mid \mathcal{B}) = -\mathbb{P}(H) \cdot \underbrace{[\mathbb{P}(O_1 \mid H) \log \mathbb{P}(O_1 \mid H) + \mathbb{P}(P_1 \mid H) \log \mathbb{P}(P_1 \mid H)]}_{1 \text{ бит}} - \mathbb{P}(\Phi) \cdot \underbrace{[\mathbb{P}(P_1 \mid \Phi) \log \mathbb{P}(P_1 \mid \Phi)]}_{0 \text{ бит}} = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = 0.5 \text{ бит}$$

Все вычисления сошлись.

Следствие 1.3.10. Для любых двух опытов:

$$\max\{H(\mathcal{A}), H(\mathcal{B})\} \leq H(\mathcal{A}, \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$$

Причем

- Равенство $H(\mathcal{A}) = H(\mathcal{A}, \mathcal{B})$ достигается тогда и только тогда, когда опыт \mathcal{B} определяется опытом \mathcal{A} .
- Равенство $H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$ достигается тогда и только тогда, когда опыты \mathcal{A} и \mathcal{B} независимы.

Доказательство. Запишем правило цепочки (предложение 1.3.9) и применим к нему с двух сторон оценки на условную энтропию (предложение 1.3.3)

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B} \mid \mathcal{A}) \geq H(\mathcal{A})$$

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B} \mid \mathcal{A}) \leq H(\mathcal{A}) + H(\mathcal{B})$$

Случаи равенства разбираются исходя из случаев равенства оценок в предложении 1.3.3. \square

О значениях совместной энтропии

Теорема 1.3.11 (О значениях совместной энтропии). Для любых опытов $\mathcal{A}_1, \dots, \mathcal{A}_r$:

$$\max_{j=1, \dots, r} H_m(\mathcal{A}_j) \leq H_m(\mathcal{A}_1, \dots, \mathcal{A}_r) \leq \sum_{j=1}^r H_m(\mathcal{A}_j)$$

Причем

- Равенство $H_m(\mathcal{A}_1, \dots, \mathcal{A}_r) = \max H_m(\mathcal{A}_j)$ достигается тогда и только тогда, когда один из опытов \mathcal{A}_k определяет любой другой опыт $\mathcal{A}_j, j \neq k$.
- Равенство $H_m(\mathcal{A}_1, \dots, \mathcal{A}_r) = \sum H_m(\mathcal{A}_j)$ достигается тогда и только тогда, когда опыты $\mathcal{A}_1, \dots, \mathcal{A}_r$ независимы в совокупности.

Доказательство. Не умаляя общности $m = 2$. Индукцией по r . Базовый случай $r = 2$ уже доказан в следствии 1.3.10. Начнем с самих неравенств при $r \geq 3$. Имеем:

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1, \mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) \geq \max\{H(\mathcal{A}_1), H(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)\} = \max\{H(\mathcal{A}_1), H(\mathcal{A}_2, \dots, \mathcal{A}_r)\} \geq \max_{* j=1, \dots, r} H(\mathcal{A}_j)$$

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1, \mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) \leq H(\mathcal{A}_1) + H(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) = H(\mathcal{A}_1) + H(\mathcal{A}_2, \dots, \mathcal{A}_r) \leq \sum_{* j=1}^r H(\mathcal{A}_j),$$

где переходы * сделаны по предположению индукции. Разберем случай равенства.

Равенство максимуму. Не умаляя общности пусть

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1)$$

Тогда, так как $H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1, \mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)$, получаем что опыт \mathcal{A}_1 полностью определяет опыт $\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r$. Формально это означает, что для любого набора событий $A_1 \in \mathcal{A}_1, \dots, A_r \in \mathcal{A}_r$:

$$\mathbb{P}(A_2 \cap \dots \cap A_r \mid A_1) \in \{0, 1\}$$

Покажем что $\mathbb{P}(A_2 \mid A_1) \in \{0, 1\}$: распишем:

$$\mathbb{P}(A_2 \mid A_1) = \sum_{A_3 \in \mathcal{A}_3} \dots \sum_{A_r \in \mathcal{A}_r} \mathbb{P}(A_2 \cap \dots \cap A_r \mid A_1)$$

Каждое слагаемое в правой части это либо 0, либо 1. С другой стороны, поскольку условная вероятность лежит в отрезке $[0; 1]$, получаем что вся сумма принимает значение или 0, или 1, что и требовалось доказать. Аналогично получаем что $\mathbb{P}(A_j \mid A_1) \in \{0, 1\}$ при $j \neq 1$. В другую сторону случай равенства следует из того что, если опыт \mathcal{A}_1 определяет опыты \mathcal{A}_2 и \mathcal{A}_3 , то определяет и опыт $\mathcal{A}_2 \wedge \mathcal{A}_3$, поскольку пересечение событий вероятностей 0 или 1, само есть событие вероятности 0 или 1.

Равенство сумме. Имеем:

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = \sum_{j=1}^r H(\mathcal{A}_j)$$

Тогда, так как $H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1, \mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)$ имеем:

1. Опыты \mathcal{A}_1 и $\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r$ независимы;
2. $H(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) = H(\mathcal{A}_2) + \dots + H(\mathcal{A}_r)$

Хотим проверить что для любого поднабора $\mathcal{A}_{j_1}, \dots, \mathcal{A}_{j_k}$ и любых событий $A_1 \in \mathcal{A}_{j_1}, \dots, A_{j_k} \in \mathcal{A}_{j_k}$ выполнено равенство

$$\mathbb{P}(A_1 \cap \dots \cap A_k) = \mathbb{P}(A_1) \cdot \dots \cdot \mathbb{P}(A_k) \quad (1.4)$$

Применяя индукционное предположение к пункту 2, получаем что опыты $\mathcal{A}_2, \dots, \mathcal{A}_r$ независимы в совокупности. Таким образом, равенство (1.4) выполнено, когда среди набора $\mathcal{A}_{j_1}, \dots, \mathcal{A}_{j_k}$ нет опыта \mathcal{A}_1 . Тогда, чтобы доказать независимость в совокупности опытов $\mathcal{A}_1, \dots, \mathcal{A}_r$ нужно показать равенство (1.4) для поднаборов $\mathcal{A}_{j_1}, \dots, \mathcal{A}_{j_k}$ с $j_1 = 1$ и произвольных $A_1 \in \mathcal{A}_{j_1}, \dots, A_k \in \mathcal{A}_{j_k}$.

Обозначим за t_1, \dots, t_{r-k} индексы не вошедшие в набор j_1, \dots, j_k , т.е. $\{t_1, \dots, t_{r-k}\} = \{1, \dots, r\} \setminus \{j_1, \dots, j_k\}$; Имеем по пункту 1:

$$\begin{aligned} \mathbb{P}(A_1 \cap \dots \cap A_k) &= \sum_{A_{k+1} \in \mathcal{A}_{t_1}} \dots \sum_{A_r \in \mathcal{A}_{t_{r-k}}} \underbrace{\mathbb{P}(A_1 \cap \dots \cap A_k \cap A_{k+1} \dots A_r)}_{\in \mathcal{A}_1 \wedge \dots \wedge \mathcal{A}_r} = \\ &= \sum_{A_{k+1} \in \mathcal{A}_{t_1}} \dots \sum_{A_r \in \mathcal{A}_{t_{r-k}}} \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \cap \dots \cap A_k \cap A_{k+1} \dots A_r) = \mathbb{P}(A_1) \cdot \sum_{A_{k+1} \in \mathcal{A}_{t_1}} \dots \sum_{A_r \in \mathcal{A}_{t_{r-k}}} \mathbb{P}(A_2 \cap \dots \cap A_k \cap A_{k+1} \dots A_r) = \\ &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \cap \dots \cap A_k) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \dots \cdot \mathbb{P}(A_k) \end{aligned}$$

Что и требовалось доказать. В обратную сторону случай равенства напрямую следует из 1.3.10. \square

Приложение. О независимых опытах, σ -алгебрах и энтропии

Доказательство теоремы выглядит довольно хитрым и возникает соблазн как-то его упростить. Можно; но для этого придется сделать язык чуть более абстрактным и доказать пару предварительных результатов.

Замечание. Это приложение — скорее черновик того, как можно развивать мысль. Многие переходы намерено расписаны не очень подробно или даже туманно, дабы дать читателю некоторую творческую свободу. Можно расценивать это приложение как руководство по тому, как не стоит записывать доказательства.

Для начала, дадим теоретико-множественную трактовку тому, что один опыт определяет другой.

Определение (Эквивалентность опытов mod 0). *Опыты \mathcal{A} и \mathcal{B} называются равными mod 0, если для любого результата $A \in \mathcal{A}$ существует результат $B \in \mathcal{B}$ такой, что*

$$\mathbb{P}(A \triangle B) = 0,$$

где $A \triangle B$ — симметрическая разность множеств. Записывается как $\mathcal{A} = \mathcal{B} \pmod{0}$.

Замечание. $\cdot = \cdot \pmod{0}$ это отношение эквивалентности.

Предложение. *Опыт \mathcal{B} определяется опытом \mathcal{A} тогда и только тогда, когда:*

$$\mathcal{A} = \mathcal{A} \wedge \mathcal{B} \pmod{0}.$$

Доказательство. Пусть опыт \mathcal{B} определяется опытом \mathcal{A} . Так как \mathcal{B} определяется \mathcal{A} , для любого $A_{i_a} \in \mathcal{A}$ существует единственное $B_{i_b} \in \mathcal{B}$ такое что

$$\mathbb{P}(B_{i_b} | A_{i_a}) = 1$$

Иначе говоря $\mathbb{P}(B_{i_b} \cap A_{i_a}) = \mathbb{P}(A_{i_a})$. Значит, по аддитивности вероятности, $\mathbb{P}((A_{i_a} \cap B_{i_b}) \triangle A_{i_a}) = 0$. Отсюда прямым образом следует $\mathcal{A} = \mathcal{A} \wedge \mathcal{B} \pmod{0}$.

Теперь обратно. Если $\mathcal{A} = \mathcal{A} \wedge \mathcal{B} \pmod{0}$, то любое событие $B \in \mathcal{B}$ представимо, с точностью до множества меры 0, как объединение нескольких множеств $A_1, \dots, A_k \in \mathcal{A}$, а значит все условные вероятности вырождены. \square

Замечание. Интуиция тут может быть следующая. Каждый опыт задает некоторое разбиение пространства элементарных исходов Ω . Опыт \mathcal{A} определяет \mathcal{B} , если опыт \mathcal{A} как разбиение Ω является измельчением разбиения \mathcal{B} (с точностью до множеств меры 0).

Однако как подсказывает заголовок при как подсказывает заголовок приложения, речь в какой-то момент должна зайти о σ -алгебрах. Действительно, формулировки о том что две (или больше) системы событий независимы или определяют друг друга в некотором смысле выглядят кустарно. Оказывается, если вместо полной системы событий \mathcal{A} рассматривать порождённую ей σ -алгебру $\sigma(\mathcal{A})$ мы придем к более естественным формулировкам.

Определение. *Для двух σ -алгебр $\mathcal{F} \subset \mathcal{H}$ и $\mathcal{G} \subset \mathcal{H}$ на одном вероятностном пространстве $(\Omega, \mathcal{H}, \mathbb{P})$, их совместная σ -алгебра есть*

$$\mathcal{F} \vee \mathcal{G} =: \sigma(\mathcal{F} \cup \mathcal{G}),$$

то есть наименьшая σ -алгебра, содержащая \mathcal{F} и \mathcal{G} .

Предложение. *Пусть \mathcal{A}, \mathcal{B} — полные системы событий. Тогда*

$$\sigma(\mathcal{A}) \vee \sigma(\mathcal{B}) = \sigma(\mathcal{A} \wedge \mathcal{B}),$$

Доказательство. С одной стороны, каждый элемент $A \cap B$ принадлежит $\sigma(\mathcal{A})$ и $\sigma(\mathcal{B})$, следовательно,

$$\sigma(\mathcal{A} \wedge \mathcal{B}) \subseteq \sigma(\mathcal{A}) \vee \sigma(\mathcal{B}).$$

С другой стороны, любые события из \mathcal{A} и \mathcal{B} могут быть получены как объединения пересечений $A \cap B$. Поэтому есть включение $\sigma(\mathcal{A}) \cup \sigma(\mathcal{B}) \subseteq \sigma(\mathcal{A} \wedge \mathcal{B})$, а значит и их порождённая σ -алгебра $\sigma(\mathcal{A}) \vee \sigma(\mathcal{B})$ содержится в $\sigma(\mathcal{A} \wedge \mathcal{B})$. Получаем равенство. \square

Для σ -алгебр аналогично полным системам событий можно ввести понятие равенство mod 0 и включения mod 0 (последнее означает, что каждый элемент из более грубой σ -алгебры с точностью до множества вероятности 0 содержится в более тонкой σ -алгебре).

Оказывается что верно следующее:

Предложение. *Опыт \mathcal{B} определяется опытом \mathcal{A} тогда и только тогда, когда $\sigma(\mathcal{B}) \subseteq \sigma(\mathcal{A}) \pmod{0}$.*

Доказательство. Следует из ранее доказанного. \square

Теперь надо разобраться с независимостью опытов. Для σ -алгебр понятие независимости вводится следующим образом:

Определение (Независимость сигма-алгебр). Сигма-алгебры $\mathcal{F}_1 \subset \mathcal{H}, \dots, \mathcal{F}_r \subset \mathcal{H}$ на одном вероятностном пространстве $(\Omega, \mathcal{H}, \mathbb{P})$ называются независимыми в совокупности, если для всех $F_1 \in \mathcal{F}_1, \dots, F_r \in \mathcal{F}_r$:

$$\mathbb{P}(F_1 \cap \dots \cap F_r) = \mathbb{P}(F_1) \dots \mathbb{P}(F_r).$$

Замечание. Полагая некоторые из F_i равными Ω , получаем что события F_1, \dots, F_r будут независимы в совокупности.

Лемма. Пусть E — событие, \mathcal{A} — некоторый опыт, и E независимо от любого $A \in \mathcal{A}$. Тогда E независимо от любого события $C \in \sigma(\mathcal{A})$.

Доказательство. Каждое $C \in \sigma(\mathcal{A})$ представимо как объединение некоторых множеств $A_{j_1}, \dots, A_{j_k} \in \mathcal{A}$. Распишем

$$\mathbb{P}(E \cap C) = \mathbb{P}(E \cap [A_{j_1} \cup \dots \cup A_{j_k}]) = \sum_{i=1}^k \mathbb{P}(E \cap A_{j_i}) = \mathbb{P}(E) \sum_{i=1}^k \mathbb{P}(A_{j_i}) = \mathbb{P}(E) \mathbb{P}(C)$$

□

Следствие. Опыты \mathcal{A} и \mathcal{B} независимы тогда и только тогда, когда независимы $\sigma(\mathcal{A})$ и $\sigma(\mathcal{B})$.

Доказательство. Из независимости σ -алгебр независимость опытов следует по определению. В другую сторону. Пусть теперь $A \in \sigma(\mathcal{A}), B \in \sigma(\mathcal{B})$. Поскольку опыты \mathcal{A} и \mathcal{B} независимы, событие $B \in \sigma(\mathcal{B})$ и событие $A' \in \mathcal{A}$ независимы для любого события A' в силу леммы выше. Тогда применяя лемму теперь уже к событию B получаем что B независимо с любым $A \in \sigma(\mathcal{A})$. □

Замечание. Обобщение на случай нескольких независимых в совокупности опытов доказывается mutatis mutandis.

Предложение. Пусть $\mathcal{F}_1, \mathcal{F}_2$ — две независимые σ -алгебры на $(\Omega, \mathcal{H}, \mathbb{P})$, $\mathcal{G} \subset \mathcal{F}_1$. Тогда \mathcal{G} и \mathcal{F}_2 независимы.

Доказательство. Следует напрямую из определения. □

Упрощённое доказательство теоремы о совместной энтропии.

1. **Равенство с максимумом.** Предположим

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1).$$

По определению совместной энтропии

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = H(\mathcal{A}_1 \wedge (\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)).$$

Следовательно,

$$H(\mathcal{A}_1) = H(\mathcal{A}_1 \wedge (\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)),$$

а значит $\sigma(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) \subseteq \sigma(\mathcal{A}_1) \pmod{0}$. Поскольку $\sigma(\mathcal{A}_j) \subseteq \sigma(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r)$ для любого $j \geq 2$, получаем

$$\sigma(\mathcal{A}_j) \subseteq \sigma(\mathcal{A}_1) \pmod{0}, \quad j = 2, \dots, r.$$

Это эквивалентно тому, что каждый опыт \mathcal{A}_j определяется \mathcal{A}_1 , что и требовалось.

2. **Равенство с суммой.** Предположим

$$H(\mathcal{A}_1, \dots, \mathcal{A}_r) = \sum_{j=1}^r H(\mathcal{A}_j).$$

По уже доказанному случаю $r = 2$ получаем

$$\sigma(\mathcal{A}_1) \text{ независима от } \sigma(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r),$$

а также

$$H(\mathcal{A}_2 \wedge \dots \wedge \mathcal{A}_r) = \sum_{j=2}^r H(\mathcal{A}_j),$$

что по индукционному предположению означает независимость сигма-алгебр $\sigma(\mathcal{A}_2), \dots, \sigma(\mathcal{A}_r)$ в совокупности. Теперь, используя независимость $\sigma(\mathcal{A}_1)$ от их объединения $\sigma(\mathcal{A}_2 \vee \dots \vee \mathcal{A}_r)$ и факт $\sigma(\mathcal{A}_2 \vee \dots \vee \mathcal{A}_r) = \sigma(\mathcal{A}_2) \wedge \dots \wedge \sigma(\mathcal{A}_r)$, получаем независимость всех $\sigma(\mathcal{A}_j)$, $j = 1, \dots, r$ (поскольку это подалгебры $\sigma(\mathcal{A}_2) \wedge \dots \wedge \sigma(\mathcal{A}_r)$). Это эквивалентно независимости самих опытов $\mathcal{A}_1, \dots, \mathcal{A}_r$.