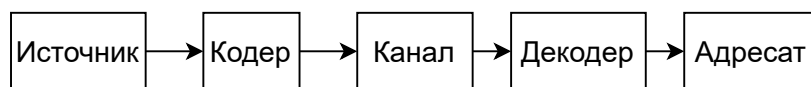


Теория информации

На этой неделе мы рассмотрим основы теории информации, которая напрямую адресует вопрос о том, что такое информация, как её измерять и какое отношение информация имеет к вероятности и статистике.

2.1 Передача информации

Пока неформально определим, что информация – это факт некоторого события, которое описывается случайной величиной или набором случайных величин. Об этом событии (реализации случайных величин) можно кому-то рассказать, тем самым совершив передачу информации. Если мы начинаем задумываться о том, как именно *рассказывается*, то мы приходим к огромному числу сценариев, которые, тем не менее, все объединяются схемой ниже.



Попробуем на примере понять, за что отвечает каждое звено этой схемы.

Пример 2.1. Лектор на паре пытается дать определение предела последовательности. Источником является он сам, он сам же кодирует информацию путём слов, жестов и выражений лица. Среда передачи – пространство аудитории, в котором есть своя искажающая акустика, из-за которой, например, без усиления лектора будет очень плохо слышно на верхних рядах. Слушатели могут быть заняты своими делами в ноутбуке и минимальное внимание уделять лектору. Так сами слушатели декодируют искажённый сигнал, возможно, иногда не распознавая его в точности так, как планировалось.



Есть и более сложные примеры.

Кодер – это некоторый механизм, преобразующий информацию в виде символов (представляющих события в широком смысле) в кодовые последовательности, а *декодер* – наоборот. Мы будем рассматривать случай, когда символов конечное число, то есть мы хотели бы уметь кодировать и декодировать конечное число m различных событий, представленных обычно цифрами $I = \{e_0, \dots, e_{m-1}\}$ и называемых *алфавитом источника*.

Кодовые последовательности состоят из q различных кодовых знаков из специального *алфавита кодера* $J = \{0, \dots, q-1\}$. Очень часто $J = \{0, 1\}$, такой алфавит называется двоичным. *Код* – это отображение f , которое переводит символ u из алфавита источника в *кодовое слово* $x_1x_2\dots x_t$. Набор всех кодовых слов называют ещё *кодовой книгой*.

Вообще, мы можем как угодно строить код, коль скоро его можно однозначно декодировать. Если ни одно кодовое слово не является префиксом другого, то код называется *беспрефиксным*; такой допускает однозначное декодирование, но в обратную сторону это утверждение неверно (можно рассмотреть кодовую книгу из 0, 01, 011).

Оказывается, что для таких кодов верен по-настоящему сильный факт, который связывает длины кодовых слов и размер алфавита.

Теорема 2.1. (*Неравенство Крафта*) Для данных натуральных чисел s_1, \dots, s_m декодируемый код f с кодовыми словами длин s_1, \dots, s_m существует тогда и только тогда, когда

$$\sum_{i=1}^m q^{-s_i} \leq 1.$$

Дополнительно, если это неравенство выполнено, то существует непрефиксный код с этими длинами кодовых слов.

Важно, что теорема говорит только о существовании кода с заданными длинами, но не об его единственности; кроме того, если один конкретный код подчиняется неравенству, то он не обязательно декодируемый, хотя нам гарантируется, что какой-то (другой) непрефиксный код можно построить. Мы не будем доказывать это неравенство для экономии времени, но если интересно, то это несложно и детали можно посмотреть в [15, Гл. 1.1].

Вопрос состоит в том, насколько короткие кодовые слова можно подобрать, если мы обладаем дополнительными сведениями об источнике? Для этого предположим, что источник таков, что вероятность получить символ e_j составляет p_j . Такие предположения вполне реалистичны: например, широко известны таблицы частот слов и символов для текстов различных типов на различных языках. Зная, что какие-то символы чаще, мы можем попробовать в среднем сэкономить на длине кодовых слов.

2.3 Энтропия и оптимальность кода

Если имеем код с известными длинами s_j кодовых слов, символ можно напрямую связать с кодовым словом, сказав, что вероятность символа p_j – это вероятность j -го кодового слова; так мы можем определить среднюю длину случайного кодового слова S как

$$\mathbb{E}[S] = \sum_{j=1}^m p_j s_j.$$

Мы хотели бы декодируемый код, обладающий минимальной средней длиной кодового слова. С помощью неравенства Крафта задачу поиска длин оптимального кода можно формализовать:

$$\begin{aligned} &\text{минимизировать } \mathbb{E}[S] \text{ по длинам } s_1, \dots, s_m \\ &\text{при условии } \sum_i q^{-s_i} \leq 1. \end{aligned}$$

Кроме того, в оптимальном решении можно оценить среднюю длину кодовых слов, которое задаёт ключевое ограничение в теории информации.

Теорема 2.2. *Для оптимального решения верно*

$$\min_{s_j} \mathbb{E}[S] \geq H_q(p_1, \dots, p_m),$$

где

$$H_q(p_1, \dots, p_m) = - \sum_i p_i \log_q p_i$$

называется энтропией дискретного распределения (p_i) .

▷ Ограничения в дискретной задаче можно ослабить, позволив нецелые s_i , а затем применить метод множителей Лагранжа. Функция Лагранжа после введения фиктивной переменной $t^2 \geq 0$ (чтобы избавиться от неравенства) будет равна

$$L(s; \lambda, t) = \sum_i s_i p_i + \lambda \left(1 - \sum_i q^{-s_i} - t^2 \right).$$

Взяв производные по s_i

$$\partial_{s_i} L = p_i + \lambda q^{-s_i} \ln q$$

и приравняв к нулю, получим, что

$$s_i = -\log_q p_i + \log_q(-\lambda \ln q).$$

Продифференцировав по другим параметрам, увидим, что $\lambda < 0, t = 0$. Подставим решение в ограничение:

$$\sum_i \frac{p_i}{-\lambda \ln q} = 1,$$

что даёт $-\lambda \ln q = 1$ и для релаксированной задачи

$$s_i = -\log_q p_i$$

является оптимальным решением. \square

Таким образом, энтропия выступает естественным порогом для оптимальности кодирования на самом базовом уровне. Важно отметить, что q -ичная энтропия H_q – это не безразмерная величина: она измеряется в q -ичных битах, а если $q = e$, то в натах, – и связана напрямую с длиной кодовых слов в оптимальном коде. Оказывается, что есть и близкая верхняя оценка.

Теорема 2.3. (Шеннон) *Для минимальной средней длины кодового слова верно*

$$H_q(p) \leq \min_{s_j} \mathbb{E}[S] \leq H_q(p) + 1.$$

▷ Для верхней оценки нам нужно построить набор длин кодовых слов. Например, можем всегда подобрать s_i так, чтобы

$$q^{-s_i} \leq p_i \leq q^{-s_i+1}.$$

Просуммировав неравенство слева, получим неравенство Крафта

$$\sum_i q^{-s_i} \leq 1,$$

то есть, есть декодируемый код с такими длинами. Правое неравенство утверждает, что

$$s_i \leq -\log_q p_i + 1$$

и

$$\mathbb{E}[S] \leq H_q(p) + 1.$$

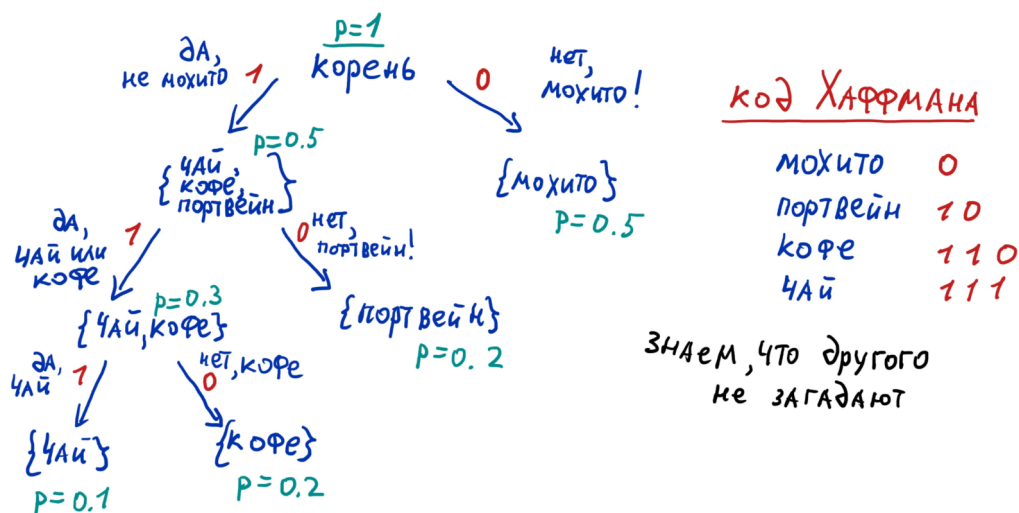
\square

Отсюда идёт частое сопоставление энтропии со средней длиной оптимального кода. Длина кодового слова для e_j при этом предлагается $\log_q p_j$ с точностью до округления, это отлично иллюстрируется старой известной игрой.

Пример 2.3. Допустим мы с вами играем в данетки, я загадываю напиток. Для простоты предположим, что мы настолько хорошо знакомы, что вы можете однозначно сказать, что этот напиток один из пяти: кофе, чай, мохито, портвейн и лимонад, – и более того известно, что кофе я загадываю примерно в 2 из 10 случаев, чай в 1 из 10 случаев, мохито в 5 из 10 случаев, а портвейн в 2 из 10 случаев.

Если играть с таким знанием, то оптимальной стратегией будет отсекаать как можно больше вариантов на каждом шаге игры. Тогда сначала вы бы спросили про то, загадал ли я мохито (если да, то в 50% случаев вы сразу выигрываете). Если нет, то вы далее спрашиваете про мохито, если нет, то вы далее спрашиваете про то, вдруг это портвейн (если да, то вы тоже сразу победили) и так далее.

На основе такого бинарного дерева отгадываний можно построить оптимальный код, который называется кодом Хаффмана:



Высота дерева (самая большая длина кодового слова) это и есть примерно $-\log_2 0.1 \approx 3.3$, то есть, 3 с точностью до округления. Ещё можно заметить, что если мы с вами так будем играть, то в среднем вам будет нужно

$$1 \cdot 0.5 + 2 \cdot 0.2 + 3 \cdot 0.2 + 3 \cdot 0.1 = 1.8$$

вопросов, а энтропия равна примерно 1.76.

В реальных играх игроки тоже сначала спрашивают очень общие вопросы, например, горячий ли напиток или алкогольный ли он, потому что такие вопросы сразу исключают много гипотез как невозможные. Крайне редко сразу пытаются назвать ответ. Сложнее всего играть с людьми, которые равновероятно загадывают что угодно, в этом случае энтропия распределения наибольшая, а количество задаваемых вопросов в среднем выше при оптимальной игре.

2.4 Кросс-энтропия и метод максимального правдоподобия

Если мы рассмотрим два кодера с разными кодами, то можно ли понять, как дорого происходит между ними трансфер информации? В контексте языка это означает буквально

следующее: если есть сообщение на одном языке в своей среде, то насколько в этой же среде оно вырастет при переводе на другой язык при условии оптимального перевода? Такой перевод можно сделать, переведя сообщение на коде 1 на естественный язык символов, а потом снова закодировав последовательность символов кодом 2.

2.4.1 Кросс-энтропия и KL-дивергенция

На этот вопрос отвечает *кросс-энтропия*, которая определяется для двух вероятностных распределений p, w , имеющих одинаковый носитель, как

$$CE(w|p) = - \sum_i w_i \log_q p_i,$$

при этом она несимметрична и читается как *кросс-энтропия из p в w* . Представим себе, что язык p изначально построен оптимально, то есть, реальные символы действительно обладают вероятностями p . Язык w строился в другом предположении, что сообщения обладают вероятностями w . И вот мы, уроженцы страны P и говорящие на p , оказываемся в стране W , где мир вероятностно устроен по-другому. Язык p станет избыточным – в мире W средняя длина кодовых слов с использованием этого кода будет выше.

Разница при переходе в мир, где код неоптимальный, занимает особое место и называется дивергенцией *Кульбака-Лейблера* или *KL-дивергенцией*

$$D_{KL}(w|p) = CE(w|p) - H(w).$$

Можно также переписать это определение в форме

$$D_{KL}(w|p) = - \sum_i w_i \log_q \frac{p_i}{w_i}.$$

По своему смыслу KL-дивергенция неотрицательна, это следует из нашей интуиции с оптимальностью выше. Это можно показать формально, используя, к примеру, в силу вогнутости $f(x) = \ln x$ неравенство Йенсена:

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]).$$

Теорема 2.4. (*Неравенство Гиббса*) Пусть p, w – два вероятностных распределения на одном и том же конечном множестве. Тогда для любого q верно

$$\sum_i p_i \log_q \frac{w_i}{p_i} \leq 0,$$

при этом полагается $0 \log_q 0 = 0$ и $p \log_q 0 = -\infty$.

▷ Не теряя общности рассмотрим случай $q = e$. Положим, что все $p_i > 0$; иначе они дадут $0 \ln \infty = 0$ в силу асимптотики при $p \rightarrow 0$. Неравенство Йенсена даст сразу

$$\sum_{i=0}^{m-1} p_i \ln \frac{w_i}{p_i} \leq \ln \sum_{i=0}^{m-1} w_i = 0.$$

□

Да, KL-дивергенция может быть бесконечной; это происходит когда мы пытаемся закодировать оптимальным кодом мира W некоторое событие e_j , которое в W было невозможным, а в P бывает.

2.4.2 Непрерывные аналоги энтропии

Мы рассматривали энтропию для дискретных распределений, можно определить её по аналогии в более общем случае. Для наглядности мы попробуем ввести определения для случая p, w – абсолютно непрерывных распределений, то есть, таких, которые обладают плотностью.

Дифференциальная энтропия – аналог энтропии в непрерывном случае – задаётся как

$$H_q(p) = - \int_{\mathbb{R}} p(x) \log_q p(x) dx.$$

Далее, кроссэнтропия и KL-дивергенция вводятся точно так же, требуется только то, чтобы вводимые интегралы были хорошо определены:

$$CE(p|w) = - \int_{\mathbb{R}} p(x) \log_q w(x) dx, \quad D_{KL}(p|w) = - \int_{\mathbb{R}} p(x) \log_q \frac{w(x)}{p(x)} dx.$$

Можно достаточно быстро заметить, что дифференциальная энтропия и кроссэнтропия уже не имеют физического смысла, связанного с информацией, потому что они могут быть отрицательными. Несмотря на это, KL-дивергенция всё ещё остаётся неотрицательной и является одним из важных способов сравнивать распределения, которым пользуются во многих областях современной статистики и в байесовских методах.

В отличие от дискретного случая в самом общем случае могут возникнуть вопросы, чему равна KL-дивергенция из дискретного (например, Пуассоновского) в непрерывное (например, гауссовское). Для того, чтобы такая KL-дивергенция $D_{KL}(p|w)$ была определена, нужно, чтобы вероятностная мера p была абсолютно непрерывна относительно w , то есть:

$$\forall A \in \mathcal{F} \quad w(A) = 0 \implies p(A) = 0.$$

В этом случае существует измеримая функция, обозначаемая dp/dw , которая называется *производной Радона-Никодима* меры p по мере w ; и KL-дивергенция определяется более общо как

$$D_{KL}(p|w) = \int_{\mathbb{R}} \log_q \frac{dp}{dw}(x) dp(x).$$

В противном случае говорят, что KL-дивергенция бесконечна.

2.4.3 Метод максимального правдоподобия

Давайте вернёмся на неделю назад и вспомним метод максимального правдоподобия. Нам дана выборка X_1, \dots, X_n , которую кратко мы обозначим за X ; мы предполагаем, что она пришла из распределения p_θ . Для оценки параметров θ предлагалось максимизировать лог-правдоподобие

$$l(\theta, X) = \ln p_\theta(X).$$

Положим, настоящий параметр равен θ_0 (но мы его не знаем). Тогда получается, что выборка пришла из распределения p_{θ_0} , а если мы могли бы взять матожидание по X_1 , то внезапно видим, что

$$\mathbb{E}_{\theta_0} [l(\theta, X_1)] = -CE(p_{\theta_0} | p_\theta).$$

В этом смысле метод максимального правдоподобия имеет чёткую информационную интерпретацию: он подгоняет такое распределение, которое бы давало примерно оптимальное кодирование информации. Но в силу того, что θ_0 нам неизвестно, мы пользуемся тем, что есть, и пытаемся в случае выборки независимых наблюдений аппроксимировать эту высокую цель более достижимой

$$\mathbb{E}_{\theta_0} [l(\theta, X)] \approx \frac{1}{n} \sum_{i=1}^n l(\theta, x_i),$$

где справа просто применена оценка Монте-Карло. Мы можем посмотреть на это ещё с другой стороны:

$$\mathbb{E}_{\theta_0} [l(\theta, X)] \approx \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta_0} [l(\theta, X_i) | X_i = x_i] = \frac{1}{n} \sum_{i=1}^n l(\theta, x_i).$$

В случае более общей вероятностной модели тоже можно заметить, что мы будто просто убрали матожидание:

$$\mathbb{E}_{\theta_0} [l(\theta, X)] \approx? \approx l(\theta, x) = \mathbb{E}_{\theta_0} [l(\theta, X) | X = x].$$

В этом случае $l(\theta, X)$ всё ещё является несмещённой оценкой $\mathbb{E}_{\theta_0} [l(\theta, X)]$. Такой взгляд тоже достаточно естественен: мы берём условное матожидание при условии того, что известно, то есть, при условии данной выборки. К этой идее мы ещё вернёмся далее.

Литература

- [1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.
- [2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.
- [3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.
- [7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.
- [8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.
- [9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.
- [10] Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.
- [11] O. A. Stepanov. Kalman filtering: Past and present. an outlook from russia. (on the occasion of the 80th birthday of rudolf emil kalman). *Gyroscopy and Navigation*, 2(2):99–110, Apr 2011.
- [12] R. L. Stratonovich. Conditional markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.
- [13] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

- [14] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.
- [15] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.
- [16] Р.Л. Стратонович. *Условные марковские процессы и их применение к теории оптимального управления*. Московский государственный университет, 1966.
- [17] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.