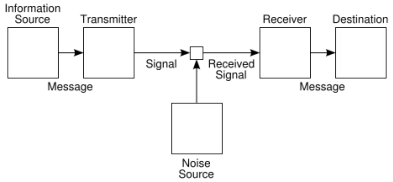
Элементы теории информации

Антон Алексеев
ПОМИ РАН
ЛШ, Дубна, 2018

Элементы теории информации: энтропия и её друзья

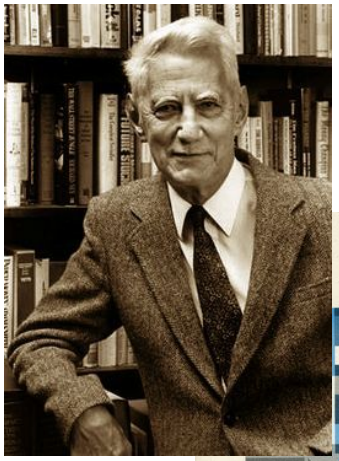


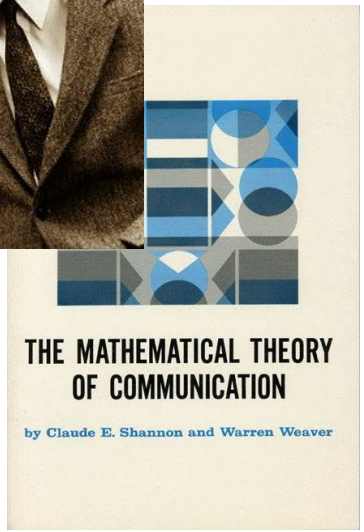
1948 - A Mathematical Theory of Communication,
Клод Шеннон; статья ввела основы теории информации

1949 - опубликована как книга с комментарием
Уоррена Уивера

Вводятся понятия информационной энтропии,
информационной избыточности, бита

Применения: алгоритмы сжатия, криптография,
обработка сигналов, ...





Собственная информация

- Сколько информации в себе несёт объект; чем менее вероятно (неожиданно) событие, тем больше информация

$$I(X) = -log_2 p(x)$$

(основание логарифма может быть другим)

- Предопределённый факт: $p(x) = 1$
Тогда

$$I(x) = 0$$

- Равномерное распределение: $p(x_i) = \frac{1}{N} \quad \forall x \in 1 : N$

$$I(x_i) = -log_2 N^{-1} = log_2 N,$$

то есть просто длина двоичного кода числа значений!

Собственная информация

- Если все слова в тексте встречаются одинаково часто и независимо друг от друга, то сжать его не получится, придётся кодировать номерами, а если нет —

$p(x_0) = \frac{1}{3}, p(x_1) = \frac{1}{3}, p(x_2) = \frac{1}{3}$

$$I_0 = log_2 3, I_1 = log_2 3, I_2 = log_2 3$$

$p(x_0) = \frac{2}{3}, p(x_1) = \frac{1}{6}, p(x_2) = \frac{1}{6}$

$$I_0 = log_2 3/2, I_1 = log_2 6, I_2 = log_2 6$$

- Редкие события наиболее «информативны»

$$=$$

мы можем позволить себе их сжимать длинными кодами

Собственная информация

- Кстати, если $p(x_0) = 0.5, p(x_1) = p_1, ..., p(x_n) = p_n$

$$I(x_0) = -log_2 0.5 = 1 \ bit$$

(название меры информации зависит от основания логарифма)

- NB! Мы не зависим от остального распределения!

Взаимная поточечная информация

PMI (Pointwise Mutual Information)

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

Интуиция:

- PMI говорит о том, сколько добавляется информации о слове word2, если видишь слово word1
- Может применяться не только к подряд идущим словам
- Дает большой вес редким словосочетаниям
- Разумно использовать как меру независимости или, наоборот, **неслучайности совместного упоминания слов** (это нам пригодится)

Взаимная поточечная информация: пример №1

Извлечение коллокаций: ***если слова встречаются вместе лишь немногим реже, чем по отдельности, они - коллокация***; вероятности оцениваем как частоты

Википедия, окт. 2015

word 1	*word 2*	*count word 1*	*count word 2*	count of co-occurrences	PMI
puerto	rico	1938	1311	1159	10.0349081703
hong	kong	2438	2694	2205	9.72831972408
los	angeles	3501	2808	2791	9.56067615065
carbon	dioxide	4265	1353	1032	9.09852946116
prize	laureate	5131	1676	1210	8.85870710982
san	francisco	5237	2477	1779	8.83305176711
...	...	...	...	...	...
to	and	1025659	1375396	1286	-3.08825363041
to	in	1025659	1187652	1066	-3.12911348956
of	and	1761436	1375396	1190	-3.70663100173

https://en.wikipedia.org/wiki/Pointwise_mutual_information

Взаимная поточечная информация: пример №2

Не SOTA (и вообще 2002), но остроумная идея подхода к анализу тональности с помощью поискового движка

1. Извлечь из текста по паттернам по частям речи определённые сочетания слов
2. Сделать запросы в AltaVista:
   “poor”, “<извл. фраза> NEAR poor”,
   “excellent”, “<извл. фраза> NEAR excellent”
3. Вычислить для всех фраз Semantic Orientation и усреднить; больше нуля - **positive**

Был такой оператор в альтависте

$$\mathrm{PMI}(word_1, word_2) = \log_2 \left[\frac{\mathrm{p}(word_1 \,\&\, word_2)}{\mathrm{p}(word_1)\,\mathrm{p}(word_2)} \right]$$

$$\mathrm{SO}(phrase) = \mathrm{PMI}(phrase, \text{“excellent”}) - \mathrm{PMI}(phrase, \text{“poor”})$$

$$\mathrm{SO}(phrase) = \log_2 \left[\frac{\mathrm{hits}(phrase \text{ NEAR “excellent”})\,\mathrm{hits}(\text{“poor”})}{\mathrm{hits}(phrase \text{ NEAR “poor”})\,\mathrm{hits}(\text{“excellent”})} \right]$$

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 417-424.

План на сегодня: отклонение от курса

1. ~~Элементы теории информации~~
    a. ~~Информация~~
        i. ~~Извлечение коллокаций~~
        ii. ~~Подход к анализу тональности~~
    b. Энтропия
        i. Связь энтропии и перплексии

Информационная энтропия

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x),$$

X — «предсказываемые значения»
Возможные трактовки формулы:

- матожидание собственной информации (как мера «содержательности»),
- оценка «непредсказуемости» системы $\mathbb{E}_{p_X} I(X)$,
- ...

Об истории термина “энтропия”

Кросс-энтропия

В теории информации перекрёстная энтропия — среднее число бит, необходимых для опознания события, если схема кодирования базируется на заданном распределении вероятностей q, вместо «истинного» p. «Википедия»

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log_2 q(x_i)$$

Истинным распределением взвешиваем информацию оценок.

Кросс-энтропия и её друзья

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log_2 q(x_i) + H(p) - H(p) =$$

$$= \sum_{i=1}^{n} p(x_i)(log_2 p(x_i) - log_2 q(x_i)) + H(p) = D_{KL}(p||q) + H(p)$$

- D_{KL} — расстояние («дивергенция») Кульбака-Лейблера
- **ОЧЕНЬ ВАЖНО**

$$H(p) \leq H(p,q) \ \forall p,q$$

именно поэтому к.-э. полезна: чем точнее оценка q, тем меньше разница + к.-э. никогда не «переоценит» истинную энтропию

Кстати*

Интересный взгляд на взаимную информацию

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y)).$$

Энтропия последовательности

- Часто нам важен текст как последовательность
- Нет проблем: для всякого языка L, задающего последовательности длины n

$$H(w_1, ..., w_n) = - \sum_{(w_1,...,w_n) \in L} p(w_1, ..., w_n) log_2 p(w_1, ..., w_n)$$

- Энтропия языка с последовательностями бесконечной длины

$$H(L) = \lim_{n \to \infty} \frac{1}{n} H(w_1, ...w_n) =$$

$$= - \lim_{n \to \infty} \frac{1}{n} \sum_{(w_1,...,w_n) \in L} p(w_1, ..., w_n) log_2 p(w_1, ..., w_n)$$

Энтропия последовательности

- Часто нам важен текст как последовательность
- Нет проблем: для всякого языка L, задающего последовательности длины n

$$H(w_1, ..., w_n) = - \sum_{(w_1,...,w_n) \in L} p(w_1, ..., w_n) log_2 p(w_1, ..., w_n)$$

- Энтропия языка с последовательностями бесконечной длины

$$H(L) = \lim_{n \to \infty} \frac{1}{n} H(w_1, ...w_n) =$$

$$= - \lim_{n \to \infty} \frac{1}{n} \sum_{(w_1,...,w_n) \in L} p(w_1, ..., w_n) log_2 p(w_1, ..., w_n)$$

УЖАС, как это считать?!

Стационарность стохастического процесса

Стохастический процесс называется **стационарным**, если вероятности последовательностей инвариантны относительно сдвигов позиций во времени

Википедия

- Случайный процесс называется *стационарным*, если все многомерные законы распределения зависят только от взаимного расположения моментов времени $t_1, t_2, \ldots, t_n$, но не от самих значений этих величин. Другими словами, случайный процесс называется стационарным, если его вероятностные закономерности неизменны во времени. В противном случае, он называется *нестационарным*.

Для естественного языка это, очевидно, не так, но иногда в рамках моделей мы можем себе позволить такое приближение

Эргодический стационарный стохастический процесс

В. Д. Колесник, Г. Ш. Полтырев
“Курс теории информации”

Пусть U_X — стационарный источник, выбирающий сообщения из множества X, и $\ldots x^{(-1)}, x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ — последовательность сообщений на его выходе. Пусть $\varphi(x_1, \ldots, x_k)$ — произвольная функция, определенная на множестве X^k и отображающая отрезки сообщений длины k в числовую ось. Пусть

$$z^{(i)} \triangleq \varphi(x^{(i+1)}, \ldots, x^{(i+k)}), \quad i = 1, 2, \ldots, \qquad (1.9.8)$$

— последовательность случайных величин, имеющих в силу стационарности одинаковые распределения вероятностей. Обозначим через m_z математическое ожидание случайных величин $z^{(i)}$.

Определение 1.9.1. Дискретный стационарный источник называется *эргодическим*, если для любого k, любой действительной функции $\varphi(x_1, \ldots, x_k)$, $\mathrm{M}\varphi(\cdot) < \infty$, определенной на X^k, любых положительных ε и δ найдется такое N, что для всех $n > N$

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} z^{(i)} - m_z\right| \geqslant \varepsilon\right) \leqslant \delta. \qquad (1.9.9)$$

Википедия

- Если при определении моментных функций стационарного случайного процесса операцию усреднения по статистическому ансамблю можно заменить усреднением по времени, то такой стационарный случайный процесс называется **эргодическим**.

Ответы Mail.RU

Попробую по-простому:
Помнишь что у случ процесс иногда записывают столбиком его реализации? Дак вот
можешь в любой момент времени провести сечение через неск-ко реализаций, найти
среднее значение, и оно окажется таким же, как если бы ты усреднял только одну
реализацию)))

Энтропия последовательности

- **Теорема Шэннона-МакМиллана-Бреймана** спешит на помощь: при стационарности и эргодичности последовательности верно, что

$$H(L) = -\lim_{n\to\infty} \frac{1}{n} log_2 p(w_1, ...w_n)$$

...то есть мы можем *просто* взять достаточно длинную последовательность для хорошей оценки

- То же верно при таких же допущениях и для перекрёстной энтропи

$$H(p, q) = -\lim_{n\to\infty} \frac{1}{n} log_2 q(w_1, ...w_n)$$

Зачем всё это: энтропия и перплексия

- Вспомним

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1})$$

- Выпишем формулу перплексии

$$PP(W) = \sqrt[n]{\frac{1}{p(x_1, ..., x_n)}} = 2^{-\frac{1}{n} log_2 p(x_1, ..., x_n)} =$$

$$= 2^{-\frac{1}{n} \sum_{i=1}^{n} \log P(x_i | x_1 ... x_{i-1})} \to 2^{H(W)}$$

- Перплексия — экспонента кросс-энтропии языка, которую мы оцениваем на достаточно длинном тексте

Где это пригодится?

- Уже знаем: оценка качества языковых моделей
- Оценка качества тематических моделей
- Ветвление в *деревьях принятия решений*