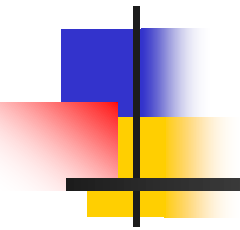
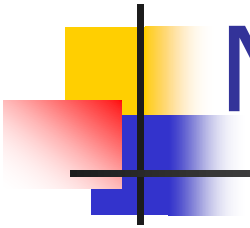


NLP数学基础

刘秉权

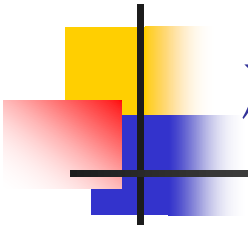
哈工大智能技术与自然语言处理研究室

2017年4月



NLP数学基础

- 概率论与数学统计
- 信息论
- 建模方法
- 最优化方法



概率论

- 概率
- 最大似然估计
- 条件概率
- 贝叶斯公式
- 二项式分布
- 期望
- 方差

概率(Probability)

$P(A)$ 为事件 A 的概率， Ω 是实验的样本空间，则概率函数必须满足如下公理：

公理 1： $P(A) \geq 0$ ；

公理 2： $P(\Omega) = 1$ ；

公理 3：如果对任意的 i 和 j ($i \neq j$),事件 A_i 和 A_j 不相交 ($A_i \cap A_j = \Phi$),则： $P(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$ 。

最大似然估计(Maximization likelihood estimation, MLE)

一个试验的样本空间是$\{s_1, s_2, \ldots, s_n\}$,在相同情况下重复试验 N 次，观察到样本$s_k (1 \le k \le n)$的次数为： $n_N(s_k)$， 则s_k 的相对频率为：

$$q_N(s_k) = \frac{n_N(s_k)}{N},$$

$$\because \sum_{k=1}^{n} n_N(s_k) = N, \quad \therefore \sum_{k=1}^{n} q_N(s_k) = 1$$

当 N 越来越大时， 相对频率 $q_N(s_k)$ 就越来越接近 s_k 的概率 $P(s_k)$：

$\lim_{N\to\infty} q_N(s_k) = P(s_k)$，因此相对频率常被用作概率的估计值，这种估计方法称为最大似然估计。

现代汉语字频统计结果：前20个最高频汉字及其频率

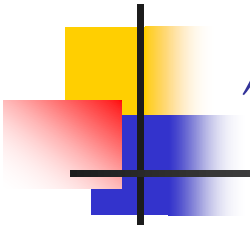
汉字	频率	汉字	频率	汉字	频率	汉字	频率
的	0.040855	了	0.008470	中	0.006012	国	0.005406
一	0.013994	有	0.008356	大	0.005857	我	0.005172
是	0.011758	和	0.007297	为	0.005720	以	0.005117
在	0.010175	人	0.006821	上	0.005705	要	0.004824
不	0.009034	这	0.006557	个	0.005488	他	0.004685

条件概率(conditional probability)

如果 A 和 B 是样本空间 Ω 上的两个事件，$P(B)>0$，那么在给定 B 时 A 的条件概率 $P(A|B)$ 为：

$$P(A|B)=\frac{P(A\cap B)}{P(B)},$$

一般地， $P(A|B)\neq P(A)$。



例

- 当预测“大学”一词出现的概率时，如果已经知道出现在它前面的两个词是“哈尔滨”和“工业”，“大学”一词出现的概率会大大增加

N-gram模型中的条件概率估计

- 极大似然估计

$$p(w_i \mid w_{i-N+1} \cdots w_{i-1}) = \frac{c(w_{i-N+1} \cdots w_i)}{\sum_{w_i} c(w_{i-N+1} \cdots w_i)}$$

$$= \frac{c(w_{i-N+1} \cdots w_i)}{c(w_{i-N+1} \cdots w_{i-1})}$$

$P(朋友|漂亮)=c(漂亮，朋友)/c(漂亮)$

$P(大学|哈尔滨,工业)=c(哈尔滨,工业,大学)/c(哈尔滨,工业)$

全概率公式

设 Ω 为实验的样本空间， $B_1, B_2, \ldots, B_n$ 为 Ω 的一组两两互斥的事件，且每次试验中至少发生一个，则称 $B_1, B_2, \ldots, B_n$ 为样本空间 Ω 的一个划分。

如果 A 为样本空间 Ω 的事件， $B_1, B_2, \ldots, B_n$ 为样本空间 Ω 的一个划分，且 $P(B_i) > 0$（ $i = 1,2,\cdots,n$ ），则全概率公式为：

$$P(A) = P(\bigcup_{i=1}^{n} AB_i) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

贝叶斯定理(Bayes' Theorem)

如果 A 为样本空间 Ω 的事件，$B_1, B_2, \ldots, B_n$ 为样本空间 Ω 的一个划分，且 $P(A) > 0$，$P(B_i) > 0$（$i = 1,2,\cdots,n$），则：

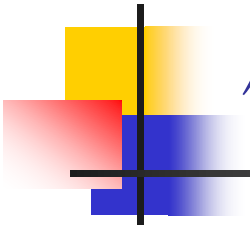
$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)},$$

当 $n = 1$ 时，$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$

例

例：输入语音信号 A，找到对应的语句 S，使得 $P(S \mid A)$ 最大，则 $\hat{S} = \arg\max_S P(S \mid A)$，根据贝叶斯公式，$\hat{S} = \arg\max_s \frac{P(S)P(A \mid S)}{P(A)}$，由于 $P(A)$ 在 A 给定时是归一化常数，因而，$\hat{S} = \arg\max_s P(S)P(A \mid S)$。

其中 $P(A \mid S)$ 为语音识别中的声学模型，$P(S)$ 为语言模型。



例

- 假设某一种特殊的句法结构很少出现，平均大约每100,000个句子中才可能出现一次。我们开发了一个程序来判断某个句子中是否存在这种特殊的句法结构。如果句子中确实含有该特殊句法结构时，程序判断结果为“存在”的概率为0.95。如果句子中实际上不存在该句法结构时，程序错误地判断为“存在”的概率为0.005。那么，这个程序测得句子含有该特殊句法结构的结论是正确的概率有多大？

解

假设G表示事件“句子确实存在该特殊句法结构”，T 表示事件“程序判断的结论是存在该特殊句法结构”。那么：

$$P(G)=\frac{1}{100000}=0.00001,\quad P(\overline{G})=\frac{100000-1}{100000}=0.99999,$$

$$P(T\mid G)=0.95,\quad P(T\mid \overline{G})=0.005$$

求解：$P(G\mid T)=?$

$$P(G\mid T)=\frac{P(T\mid G)P(G)}{P(T\mid G)P(G)+P(T\mid \overline{G})P(\overline{G})}$$

$$=\frac{0.95\times 0.00001}{0.95\times 0.00001+0.005\times 0.99999}\approx 0.002$$

二项式分布 (binomial distribution)

当重复一个只有两种输出（假定为 $\overline{A}$ 和 A ）的试验（伯努利试验），A 在一次实验中发生的概率为 p ，现将实验独立地重复 n 次，如果用 X 表示 A 在这 n 次实验中发生的次数，那么， $X = 0,1,\cdots,n$ 。则 n 次独立实验中成功的次数为 r 的概率为： $p_r = C_n^r p^r (1-p)^{n-r}$ ，其中， $C_n^r = \frac{n!}{(n-r)!r!}$ ，$0 \le r \le n$ 。此时 X 所遵从的概率分布称为二项式分布，并记为： $X \sim B(n, p)$ 。

自然语言处理中常以句子为处理单位，一般假设一个语句独立于它前面的其他语句，句子的概率分布近似地认为符合二项式分布。

期望(Expectation)

期望值是一个随机变量所取值的概率平均。设 X 为一随机变量，其分布为 $P(X = x_k) = p_k$ ， $k = 1,2,\cdots,$

若级数 $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛，那么随机变量 X 的数学期望

或概率平均值为： $E(X) = \sum_{k=1}^{\infty} x_k p_k$ 。

方差(Variance)

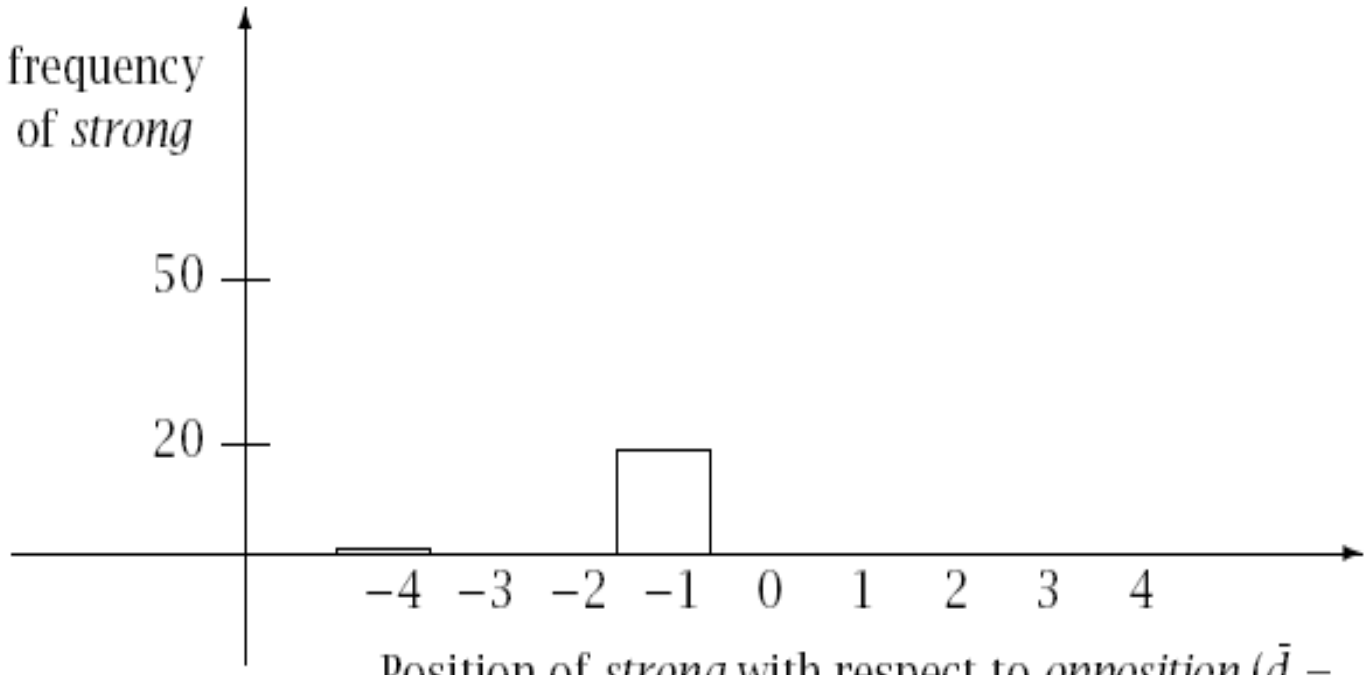
一个随机变量的方差描述的是该随机变量的值偏离其期望值的程度。设 X 为一随机变量，其方差为：

$$Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

计算两个词之间偏移量的均值和方差

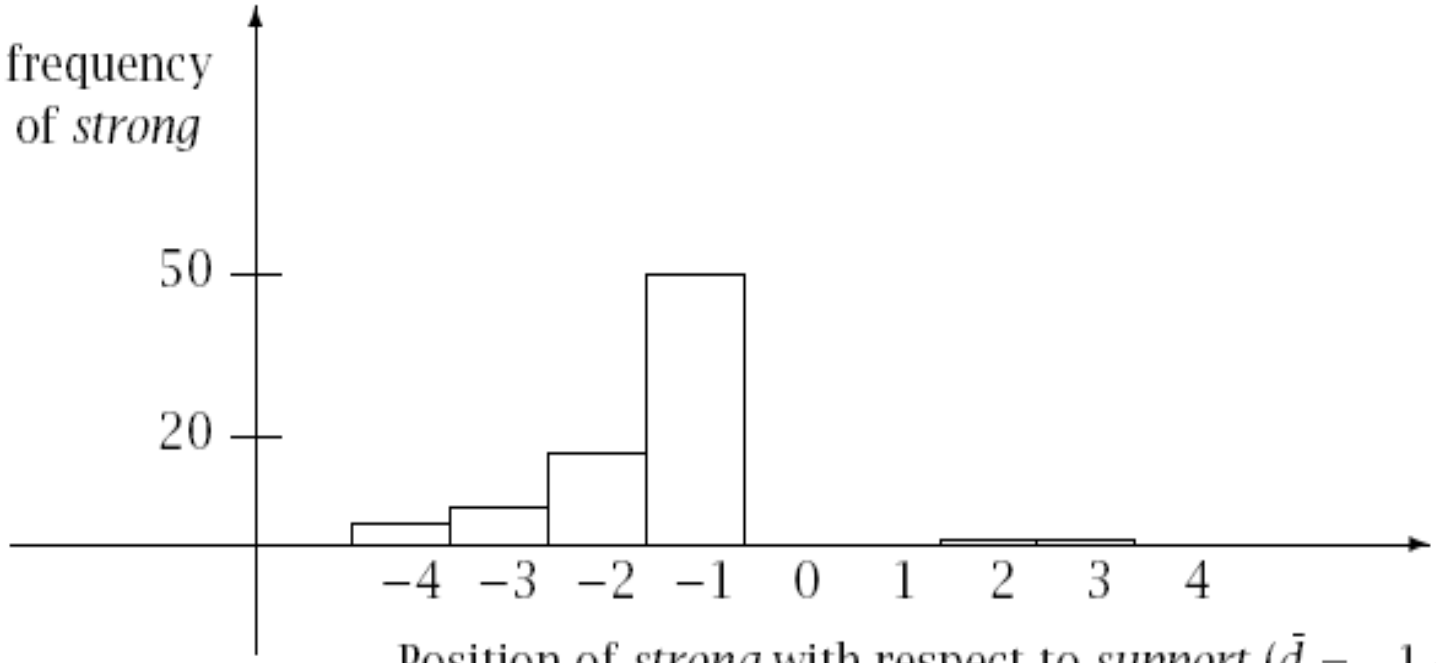
- 均值是简单的平均偏移量
- 方差衡量单独的偏移量偏离均值的距离
  - 低的偏差值意味两个词通常会以大致相同的距离出现
- 方差是关于一个词相对于其他词分布峰值情况的量度

strong相对于opposition的位置



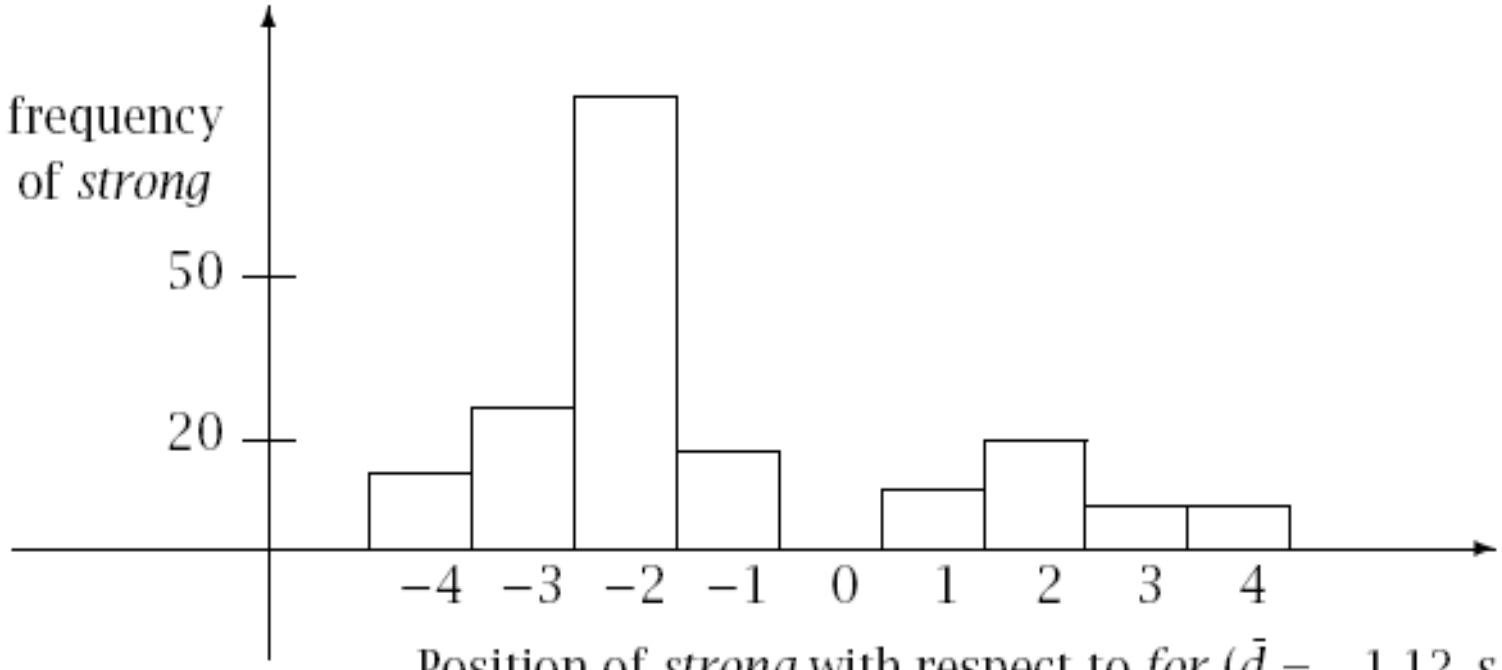
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

strong相对于support的位置



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

strong相对于for的位置



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

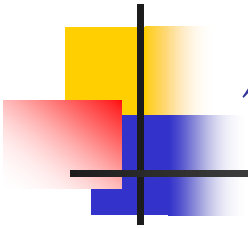
基于均值和方差的搭配发现

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

汉语搭配实例

表 2　语义搭配超常度计算实验示例

超常搭配				常规搭配			
序号	中心词 w_h	受支配词 w_d	$Uncolloca(w_h, w_d)$	序号	中心词 w_h	受支配词 w_d	$Uncolloca(w_h, w_d)$
1	编织	梦想	32.00	1	编织	花篮	1.82
2	阅读	人生	3.64	2	阅读	书籍	1.00
3	捕捉	歌声	14.54	3	治愈	创伤	1.00
4	摧毁	健康	20.00	4	展示	风景画	1.36
5	撕破	天	3.64	5	流露	气质	1.82
6	酿造	黑暗	32.00	6	建立	帝国	1.36
7	雕镌	人生	14.54	7	失去	目标	1.82
8	敞开	心扉	3.64	8	需要	养料	1.00
9	拥抱	大地	14.54	9	感到	失望	1.00
10	劈开	海浪	1.09 (判断错误)	10	挖掘	宝石	3.63 (判断错误)



信息论

- 1948年美国Shannan香农“通信的数学理论”，用概率测度和数理统计的方法，系统地讨论了通信的基本问题，奠定了信息论的基础
- 什么是信息？
- 信息的度量有三个基本方向：结构的、语义的和统计的
- 香农所说的信息是狭义的信息，是统计信息，依据是概率的不确定性度量

Claude Shannon

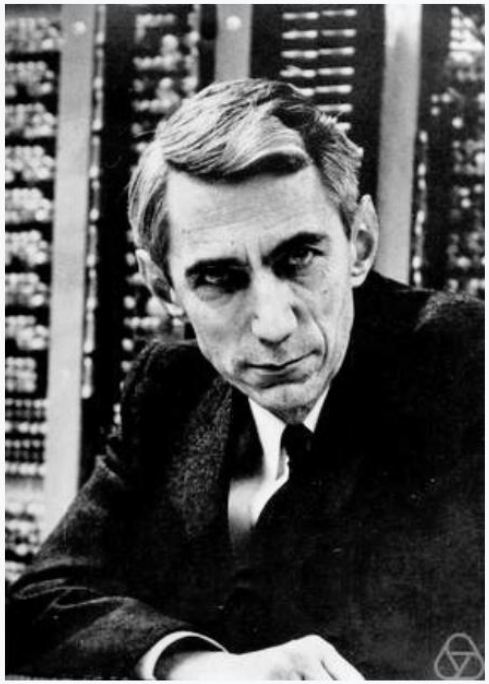
From Wikipedia, the free encyclopedia

Claude Elwood Shannon (April 30, 1916 – February 24, 2001) was an American mathematician, electrical engineer, and cryptographer known as "the father of information theory".[1][2]

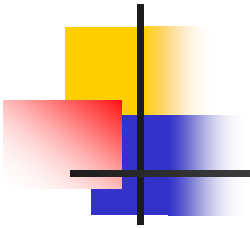
Shannon is noted for having founded information theory with a landmark paper, *A Mathematical Theory of Communication*, that he published in 1948. He is, perhaps, equally well known for founding digital circuit design theory in 1937, when—as a 21-year-old master's degree student at the Massachusetts Institute of Technology (MIT)—he wrote his thesis demonstrating that electrical applications of Boolean algebra could construct any logical, numerical relationship.[3] Shannon contributed to the field of cryptanalysis for national defense during World War II, including his fundamental work on codebreaking and secure telecommunications.

Contents [hide]

Claude Shannon



Born	April 30, 1916 Petoskey, Michigan, United States
Died	February 24, 2001 (aged 84) Medford, Massachusetts, United States
Nationality	American



内容

- 熵
- 联合熵
- 互信息
- 相对熵
- 交叉熵
- 迷惑度
- 噪声信道模型

熵(Entropy)

熵表示信息源 X 每发一个符号所需要的平均信息量。

熵也可以被视为描述一个随机变量的不确定性的数量。

一个随机变量的熵越大，它的不确定性越大，正确估计其值的可能性就越小。越不确定的随机变量越需要大的信息量用以确定其值。

熵的定义

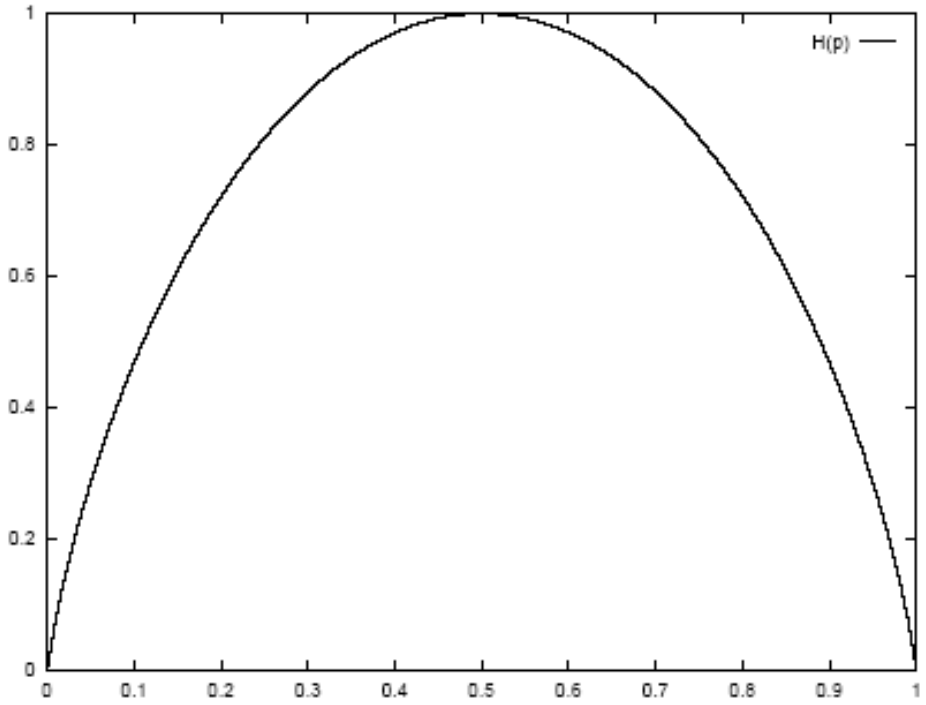
如果 X 是一个离散型随机变量，其概率分布为：

$p(x) = P(X = x)$ ， $x \in X$ ，则 X 的熵 $H(X)$ 为：

$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$，其中约定 $0 \log 0 = 0$ 。

$H(X)$ 也可以写为 $H(p)$，通常熵的单位为二进制位比特（bit）。

抛非均匀硬币事件的熵值



横轴表示硬币正面朝上的概率，纵轴表示抛一次硬币试验的熵值

英语字母的熵-等概率情况

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

$$= 26 \times (-\frac{1}{26} \log_2 \frac{1}{26}) = \log_2 26 = 4.70$$

英语字母的熵-实际出现情况

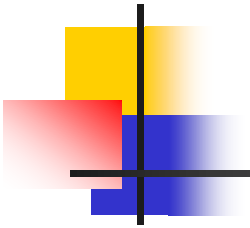
- 例：考察英语中特定字母出现的频率。当观察字母的个数较少时，频率有较大幅度的随机波动，但当观察数目增大时，频率即呈现出稳定性，有人统计了438023个字母得到如下表所示的数据。

特定英语字母的出现频率

字母	频率	字母	频率	字母	频率
E	0.1268	L	0.0394	P	0.0186
T	0.0978	D	0.0389	B	0.0156
A	0.0788	U	0.0280	V	0.0102
O	0.0776	C	0.0268	K	0.0060
I	0.0707	F	0.0256	X	0.0016
N	0.0706	M	0.0244	J	0.0010
S	0.0634	W	0.0214	Q	0.0009
R	0.0594	Y	0.0202	Z	0.0006
H	0.0573	G	0.0187		

英语字母的熵-实际结果

- 根据熵的定义计算，每收到一个英文字母信号的不确定程度是4.1606比特。



比较

- 考虑了英文字母实际出现的概率后，英文信源的平均不确定性，比把字母看作等概率出现时英文信源的平均不确定性要小
- 均衡分布的熵最大

汉字的熵

- 中文有6000多个常用字，经中国冯志伟等人测算，汉字的信息熵随着汉字个数的增加而增加，当汉字的个数达到12366个汉字时，汉字的信息熵值为9.65比特。因此，汉字机内码必须用两个字节才能表示一个汉字。

联合熵(Joint Entropy)

如果 X 、 Y 是一对离散型随机变量，其联合概率分布密度函数为 $p(x,y)$，X 、Y 的联合熵定义为：

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

联合熵就是描述一对随机变量平均所需要的信息量。

条件熵(Conditional Entropy)

如果离散型随机变量(X , Y)的联合概率分布密度函数为 $p(x,y)$，已知随机变量 X 的情况下随机变量 Y 的条件熵定义为：

$$
\begin{aligned}
H(Y \mid X) &= \sum_{x \in X} p(x) H(Y \mid X = x) \\
&= \sum_{x \in X} p(x)[-\sum_{y \in Y} p(y \mid x) \log p(y \mid x)] \\
&= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y \mid x)
\end{aligned}
$$

条件熵表示的是在已知 X 的情况下，传输 Y 额外所需的平均信息量。

熵的连锁规则

$$H(X,Y) = H(X) + H(Y \mid X)$$

$$H(X_1, \cdots X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \cdots X_{n-1})$$

熵的连锁规则

证明：

$$
\begin{aligned}
H(X,Y) &= -E_{p(x,y)}(\log p(x,y)) \\
&= -E_{p(x,y)}(\log(p(x)p(y\mid x))) \\
&= -E_{p(x,y)}(\log p(x)+\log p(y\mid x)) \\
&= -E_{p(x)}(\log p(x)) - E_{p(x,y)}(\log p(y\mid x)) \\
&= H(X)+H(Y\mid X)
\end{aligned}
$$

互信息(Mutual Information)

如果离散型随机变量(X , Y)的联合概率分布密度函数为 $p(x,y)$， X , Y 之间的互信息定义为：

$I(X;Y)=H(X)-H(X\mid Y)$，展开得到：

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

互信息 $I(X;Y)$ 是在知道了 Y 的值后 X 的不确定性的减少量。即 Y 的值透露了多少关于 X 的信息量。

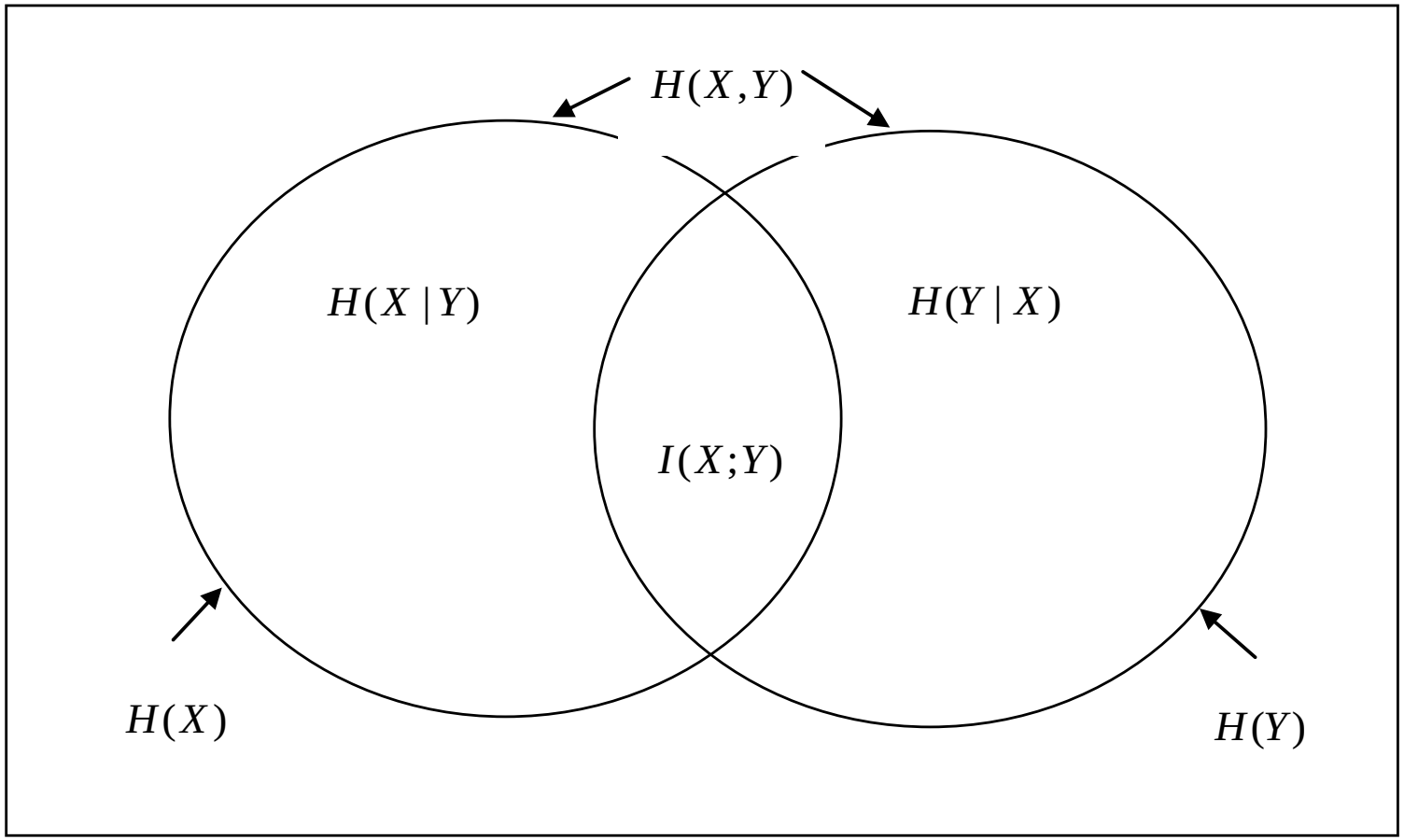
互信息(Mutual Information)

$$\because H(X \mid X) = 0,$$

$$\therefore H(X) = H(X) - H(X \mid X) = I(X;X)$$

这一方面说明了为什么熵又称自信息，另一方面说明了两个完全相互依赖的变量之间的互信息并不是常量，而是取决于他们的熵。

互信息与熵之间的关系



基于互信息的搭配发现：按互信息大小排列的出现20次的10个二元组

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

相对熵(Relative Entropy or Kullback-Leibler Divergence)

两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为：

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} ,$$

其中约定 $0 \log(0/q) = 0, p \log(p/0) = \infty$ 。

相对熵常被用以衡量两个随机分布的差距。 $D(p \| q) \geq 0$ ， 当且仅当两个随机分布相同时， 其相对熵为 0， 当两个随机分布的差别增加时， 其相对熵也增加。 $D(p \| q) \neq D(q \| p)$ 。

交叉熵(Cross Entropy)

如果一个随机变量 X 的概率分布为 $p(x)$，

$q(x)$ 为用于近似 $p(x)$ 的概率分布，那么随机变量

X 和模型 q 之间的交叉熵定义为：

$$H(X,q) = H(X) + D(p \| q)$$
$$= -\sum_{x} p(x) \log q(x)$$

交叉熵的概念是用来衡量估计模型与正式概率分布之间差异的。

语言与其模型的交叉熵

概率分布为 $p(x)$ 的语言 $L=(X_i)$ 与其模型 q 的交叉熵定义为：

$$H(L,q)=-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n}p(x_1^n)\log q(x_1^n)$$

其中， $x_1^n=x_1,\cdots,x_n$ 为语言 L 的语句；

$p(x_1^n)$ 为 L 中语句 x_1^n 的概率；

$q(x_1^n)$ 为模型 q 对 x_1^n 的概率估计。

语言与其模型的交叉熵

假设在理想情况下，即 n 趋于无穷大时，其全部“单词”的概率和为 1。即根据信息论的定理：假定语言 L 是稳态（stationary）ergodic 随机过程， L 与其模型 q 的交叉熵计算公式就变为：

$$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n)。$$

因此可以根据模型 q 和一个含有大量数据的 L 的样本来计算交叉熵。在设计模型 q 时，目的是使交叉熵最小，从而使模型最接近真实的概率分布 $p(x)$。

迷惑度(复杂度,困惑度,Perplexity)

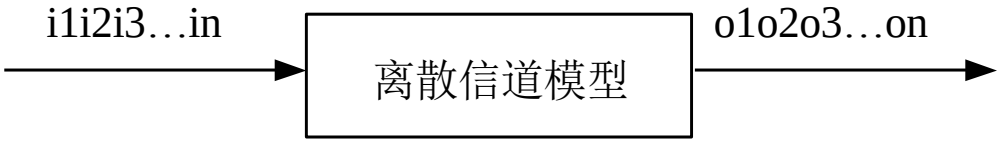
在设计语言模型时，通常用迷惑度来代替交叉熵衡量语言模型的优劣。给定语言 L 的样本 $l_1^n = l_1 \cdots l_n$，L 的迷惑度定义为：

$$PP_q = 2^{H(L,q)} \approx 2^{\frac{1}{n}\log(l_1^n)} = [q(l_1^n)]^{\frac{1}{n}}。$$

语言模型设计的任务就是寻找迷惑度最小的模型，使其最接近真实的语言。

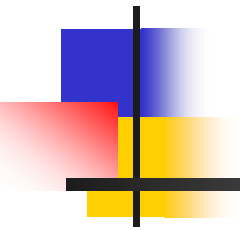
信源信道模型

- 信源—信道模型：



$$\hat{I} = \arg\max_{I}(p(I|O)) = \arg\max_{I}\frac{p(I)p(O|I)}{p(O)} = \arg\max_{I} p(I)p(O|I)$$

- I:语言文本；O:声音信号、字符图像信号、拼音输入等
- 语言模型： $p(I)$



谢谢！