

第二章 数学基础

目录

2.1 概率论基础

2.2 信息论基础

2.3 应用实例

2.1 概率论基础

2.1 概率论基础

◆ 基本概念

- 概率(probability)
- 最大似然估计(maximum likelihood estimation)
- 条件概率(conditional probability)
- 全概率公式(full probability)
- 贝叶斯决策理论(Bayesian decision theory)
- 贝叶斯法则(Bayes' theorem)
- 二项式分布(binomial distribution)
- 期望(expectation)
- 方差(variance)

2.1 概率论基础

◆ 概率(probability):

概率是从随机实验中的事件到实数域的函数，用以表示事件发生的可能性。如果用 $P(A)$ 作为事件 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 $A_i \cap A_j = \emptyset$ ，则有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \quad (1)$$

2.1 概率论基础

◆ 最大似然估计:

(Maximization likelihood estimation, MLE)

如果一个实验的样本空间是 $\{S_1, S_2, \dots, S_n\}$, 在相同情况下重复实验 N 次, 观察到样本 $s_k (1 \leq k \leq n)$ 的次数为 $n_N(s_k)$, 则 s_k 的相对频率为:

$$q_N(s_k) = \frac{n_N(s_k)}{N} \quad (2)$$

由于 $\sum_{k=1}^n n_N(s_k) = N$, 因此, $\sum_{k=1}^n q_N(s_k) = 1$ 。

2.1 概率论基础

◆ 最大似然估计:

(Maximization likelihood estimation, MLE)

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 $P(s_k)$ 。事实上,

$$\lim_{N \rightarrow \infty} q_N(s_k) = P(s_k) \quad (3)$$

因此, 相对频率常被用作概率的估计值。这种概率值的估计方法称为**最大似然估计**。

2.1 概率论基础

◆ 条件概率(conditional probability):

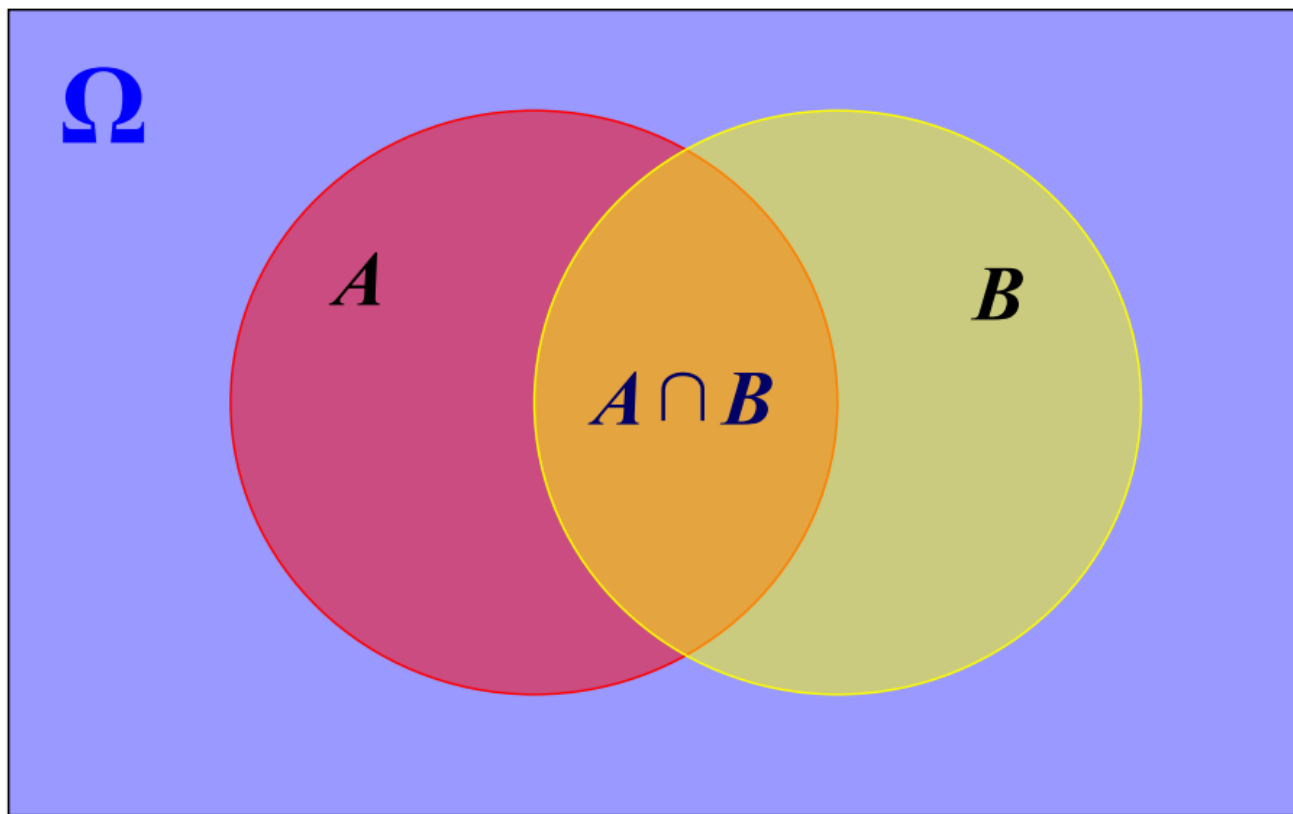
如果 A 和 B 是样本空间 Ω 上的两个事件, $P(A) > 0$, 那么在给定 B 时 A 的条件概率 $P(A|B)$ 为:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (4)$$

条件概率 $P(A|B)$ 给出了在已知事件 B 发生的情况下, 事件 A 发生的概率。

一般地, $P(A|B) \neq P(A)$ 。

2.1 概率论基础



2.1 条件概率示意图

2.1 概率论基础

◆ 全概率公式:

设 Ω 为实验 E 的样本空间, $B_1, B_2 \cdots B_n$ 为 Ω 的一组事件, 且他们两两互斥, 且每次实验中至少发生一个, 即

$$(1) B_i \cap B_j = \phi \quad (i \neq j; i, j = 1, 2 \cdots n) \quad (5)$$

$$(2) \bigcup_{i=1}^n B_i = \Omega \quad (6)$$

则称 $B_1, B_2 \cdots B_n$ 为样本空间 Ω 的一个划分。

2.1 概率论基础

◆ 全概率公式:

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

$$\begin{aligned} P(A) &= P(\cup_{i=1}^n AB_i) = \sum_{i=1}^n P(AB_i) \\ &= \sum_{i=1}^n P(B_i) P(A|B_i) \end{aligned} \quad (7)$$

2.1 概率论基础

◆ 贝叶斯法则(Bayes' theorem):

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

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)} \quad (8)$$

当 $n = 1$ 时,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (9)$$

2.1 概率论基础

◆ 贝叶斯决策理论(Bayesian decision theory):

假设研究的分类问题有 c 个类别, 各类别的状态用 w_i , $i = 1, 2 \cdots c$ 表示, w_i 出现的先验概率为 $P(w_i)$; 在特征空间已观察到某一向量 $\bar{x} = [x_1, x_2, \cdots, x_d]^T$ 是 d 维特征空间上的某一点, 且条件概率密度函数 $P(x|w_i)$ 是已知的。那么, 利用贝叶斯公式我们可以得到后验概率:

$$P(w_i|\bar{x}) = \frac{P(\bar{x}|w_i)P(w_i)}{\sum_{j=1}^n P(\bar{x}|w_j)P(w_j)} \quad (10)$$

2.1 概率论基础

➤ 基于最小错误率的贝叶斯决策规则为：

(1) 如果 $P(w_i|\bar{x}) = \max_{j=1,2\cdots c} P(w_j|\bar{x})$, 则 $\bar{x} \in w_i$;

(2) 或：如果 $P(\bar{x}|w_i)P(w_i) = \max_{j=1,2\cdots c} P(\bar{x}|w_j)P(w_j)$,
则 $\bar{x} \in w_i$;

(3) 或 ($c = 2$): 如果 $l(\bar{x}) = \frac{P(\bar{x}|w_1)}{P(\bar{x}|w_2)} > \frac{P(w_2)}{P(w_1)}$,
则 $\bar{x} \in w_1$, 否则 $\bar{x} \in w_2$.

贝叶斯决策理论在文本分类、词汇语义消歧(word sense disambiguation)等问题的研究中具有重要用途

2.1 概率论基础

例2-1:

给定语音信号 A ，找出对应的语句 S ，使得 $P(S|A)$ 最大，
即： $\hat{S} = \operatorname{argmax}_S P(S|A)$

根据贝叶斯公式：

$$\hat{S} = \operatorname{argmax}_S \frac{P(S)P(A|S)}{P(A)}$$

由于 $P(A)$ 在 A 给定时是归一化常数，因而

$$\hat{S} = \operatorname{argmax}_S P(A|S)P(S)$$

声学模型

语言模型

2.1 概率论基础

例2-2:

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？

2.1 概率论基础

解2-2:

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，我们有：

$$P(G) = \frac{1}{100000} = 0.00001 \quad P(T|G) = 0.95$$

$$\begin{aligned} P(G|T) &= \frac{P(G|T)P(G)}{P(T|G)P(G) + P(T|\bar{G})P(\bar{G})} \\ &= \frac{0.95 \times 0.00001}{0.95 \times 0.00001 + 0.005 \times 0.99999} \approx 0.002 \end{aligned}$$

2.1 概率论基础

◆ 二项式分布(binomial distribution) :

当重复一个只有两种输出(假定为 \bar{A} 或 A)的实验(伯努利实验), A 在一次实验中发生的概率为 p , 现把实验独立地重复 n 次。如果用 X 表示 A 在这 n 次实验中发生的次数, 那么, $X = 0, 1, \dots, n$ 。

考虑事件 $\{X = i\}$, 如果这个事件发生, 必须在这 n 次的原始记录中有 i 个 A , $n - i$ 个 \bar{A} 。

$$\underbrace{A\bar{A}AA\cdots\bar{A}}_{n\text{个}} \Longrightarrow p^i (1-p)^{n-i}$$

2.1 概率论基础

◆ 二项式分布(binomial distribution) :

A 可以出现在 n 个位置中的任何一个位置, 所以结果序列有 $\binom{n}{i}$ 种可能, 由此得出:

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, i = 1, 2 \cdots n \quad (11)$$

其中, $\binom{n}{i} = C_n^i = \frac{n!}{(n-i)!i!}, 0 \leq i \leq n$

X 所遵从的概率分布(11)称为二项式分布, 并记为:

$$X \sim B(n, p)$$

2.1 概率论基础

◆ 二项式分布(binomial distribution) :

在自然语言处理中，一般以句子为处理单位。假设一个句子独立于它前面的其它语句，句子的概率分布近似地认为符合二项式分布。

2.1 概率论基础

◆ 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 $P(X = x_k) = p_k, k = 1, 2, \dots$ 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：

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \quad (12)$$

2.1 概率论基础

◆ 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：

$$\begin{aligned} \text{Var}(X) &= E \left((X - E(X))^2 \right) \\ &= E(X^2) - E^2(X) \end{aligned} \tag{13}$$

2.2 信息论基础

2.2 信息论基础

◆ 熵(entropy):

香农(Claude Elwood Shannon)

于1940年获得MIT数学博士学位和电子工程硕士学位后，于1941年加入了贝尔实验室数学部，并在那里工作了15年。

1948年6月和10月，由贝尔实验室出版的《贝尔系统技术》杂志连载了香农博士的文章《通讯的数学原理》，该文奠定了香农信息论的基础。

熵是信息论中重要的基本概念。



2.2 信息论基础

◆ 熵(entropy):

如果 X 是一个离散型随机变量, 其概率分布为:

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

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (14)$$

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)。

2.2 信息论基础

◆ 熵(entropy):

熵又称为自信息(self-information), 表示信源 X 每发一个符号(不论发什么符号)所提供的平均信息量。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。那么, 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。

2.2 信息论基础

例2-3:

计算下列两种情况下英文(26个字母和1个空格, 共27个字符)信息源的熵: (1)假设27个字符等概率出现; (2)假设英文字母的概率分布如下:

字母	空格	E	T	O	A	N	I	R	S
概率	0.1956	0.105	0.072	0.0654	0.063	0.059	0.055	0.054	0.052

字母	H	D	L	C	F	U	M	P	Y
概率	0.047	0.035	0.029	0.023	0.0225	0.0225	0.021	0.0175	0.012

字母	W	G	B	V	K	X	J	Q	Z
概率	0.012	0.011	0.0105	0.008	0.003	0.002	0.001	0.001	0.001

2.2 信息论基础

解2-3:

(1) 等概率出现情况:

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x) \\ &= 27 \times \left\{ -\frac{1}{27} \log_2 \frac{1}{27} \right\} = 4.75(\text{bits/letter}) \end{aligned}$$

(2) 实际情况:

$$H(X) = -\sum_{i=1}^{27} p(x_i) \log_2 p(x_i) = 4.02(\text{bits/letter})$$

说明: 考虑了英文字母和空格实际出现的概率后, 英文信源的平均不确定性, 比把字母和空格看作等概率出现时英文信源的平均不确定性要小。

2.2 信息论基础

- 法语、意大利语、西班牙语、英语、俄语字母的熵[冯志伟, 1989]:

语言	熵 (bits)
法语	3.98
意大利语	4.00
西班牙语	4.01
英语	4.03
俄语	4.35

英语词的熵约为10bits

2.2 信息论基础

1970年代末期冯志伟首先开展了对汉字信息熵的研究，经过几年的文本收集和手工统计，在当时艰苦的条件下测定了汉字的信息熵为 **9.65** 比特(bit)。1980年代末期，刘源等测定了汉字的信息熵为 **9.71** 比特，而汉语词的熵为 **11.46** 比特。

汉语词汇平均长度约为 **2.5** 个汉字。

2.2 信息论基础

➤ 北京、香港、台北三地汉语词的熵[Tsou,2003]:

北京5年		台北5年		香港5年		京、港、台5年	
A1	A2	B1	B2	C1	C2	D1	D2
11.45	11.11	11.69	11.36	11.96	11.64	11.96	11.60

其中， A1, B1, C1 分别是从小LIVAC文本集中北京、台北、香港三地5年各约1000万字文本中所提取的数据；A2, B2, C2 为三地文本剔除专用名词之后的数据。D1, D2分别为三地文本合并剔除专用名词以后的数据。

专用名词主要指： 人名、地名、组织机构名。

2.2 信息论基础

◆ 联合熵(joint entropy):

如果 X, Y 是一对离散型随机变量 $X, Y \sim p(x, y)$, X, Y 的联合熵 $H(X, Y)$ 为:

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) \quad (15)$$

联合熵实际上就是描述一对随机变量平均所需要的信息量。

2.2 信息论基础

◆ 条件熵(conditional entropy):

给定随机变量 X 的情况下, 随机变量 Y 的条件熵定义为:

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) H(Y|X = x) \\ &= \sum_{x \in X} p(x) \left[- \sum_{y \in Y} p(y|x) \log_2 p(y|x) \right] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x) \end{aligned} \quad (16)$$

2.2 信息论基础

◆ 条件熵(conditional entropy):

将(15)式中的 $\log_2 p(x, y)$ 根据概率公式展开:

$$\begin{aligned} H(X, Y) &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x) p(y | x)] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) [\log p(x) + \log p(y | x)] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) \\ &= - \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x) \\ &= H(X) + H(Y | X) \end{aligned} \quad (17) \quad \textbf{(连锁规则)}$$

2.2 信息论基础

例2-4:

假设 (X, Y) 服从如下联合概率分布:

Y \ X	1	2	3	4
1	$1/8$	$1/16$	$1/32$	$1/32$
2	$1/16$	$1/8$	$1/32$	$1/32$
3	$1/16$	$1/16$	$1/16$	$1/16$
4	$1/4$	0	0	0

请计算 $H(X)$ 、 $H(Y)$ 、 $H(X|Y)$ 、 $H(Y|X)$ 和 $H(X, Y)$ 各是多少?

2.2 信息论基础

解2-4:

Y \ X	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0
P(X)	1/2	1/4	1/8	1/8

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x) \\ &= -\left(\frac{1}{2} \times \log_2 \left(\frac{1}{2} \right) + \frac{1}{4} \times \log_2 \left(\frac{1}{4} \right) + \frac{1}{8} \times \log_2 \left(\frac{1}{8} \right) + \frac{1}{8} \times \log_2 \left(\frac{1}{8} \right) \right) \\ &= \frac{7}{4} \end{aligned}$$

2.2 信息论基础

解2-4:

Y \ X	1	2	3	4	H(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) = 2 \text{ (bits)}$$

2.2 信息论基础

解2-4:

Y \ X	1	2	3	4	H(Y)
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
P(X)	1/2	1/4	1/8	1/8	

$$p(x_1 | y_1) = \frac{p(x_1, y_1)}{p(y_1)} = \frac{1}{8} \times \frac{4}{1} = \frac{1}{2} \quad p(x_2 | y_1) = \frac{p(x_2, y_1)}{p(y_1)} = \frac{1}{16} \times \frac{4}{1} = \frac{1}{4}$$

$$p(x_3 | y_1) = \frac{p(x_3, y_1)}{p(y_1)} = \frac{1}{32} \times \frac{4}{1} = \frac{1}{8} \quad p(x_4 | y_1) = \frac{p(x_4, y_1)}{p(y_1)} = \frac{1}{32} \times \frac{4}{1} = \frac{1}{8}$$

.....

2.2 信息论基础

解2-4:

$$-\sum_{i=0}^4 p(x_i|y_1) \log_2 p(x_i|y_1)$$

$$-\sum_{i=0}^4 p(x_i|y_2) \log_2 p(x_i|y_2)$$

$$H(X|Y) = \sum_{i=1}^4 p(y=i) H(X|Y=i)$$

$$= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right)$$

$$+ \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0)$$

$$= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 = \frac{11}{8}(\text{bit})$$

类似地, $H(Y|X) = 13/8$ (bits), $H(X, Y) = 27/8$ (bits)。

可见 $H(X|Y) \neq H(Y|X)$

2.2 信息论基础

例2-5:

简单的波利尼西亚语(Polynesian)是一些随机的字符序列，其中部分字符出现的概率为：

p: 1/8, t: 1/4, k: 1/8, a: 1/4, i: 1/8, u: 1/8

那么，每个字符的熵为：

$$\begin{aligned} H(P) &= \sum_{i \in \{p, t, k, a, i, u\}} P(i) \log P(i) \\ &= - \left[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4} \right] \\ &= 2.5 \text{ (bits)} \end{aligned}$$

2.2 信息论基础

- 这个结果表明，我们可以设计一种编码，传输一个字符平均只需要2.5个比特：

p	t	k	a	i	u
100	00	101	01	110	111

这种语言的字符分布并不是随机变量，但可以将其近似地看作随机变量。

2.2 信息论基础

- 如果将字符按元音和辅音分成两类，元音随机变量 $V = \{a, i, u\}$ ，辅音随机变量 $C = \{p, t, k\}$ 。
- 假定所有的单词都由CV(consonant-vowel)音节序列组成，其联合概率分布 $P(C, V)$ 、边缘分布 $P(C, \cdot)$ 和 $P(\cdot, V)$ 如下表所示：

V \ C	p	t	k	$P(\cdot, V)$
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
$P(C, \cdot)$	1/8	3/4	1/8	

2.2 信息论基础

注意，这里的边缘概率是基于每个音节的，其值是基于每个字符的概率的两倍，因此，每个字符的概率值应该为相应边缘概率的 $1/2$ ，即：

p: $1/16$ t: $3/8$ k: $1/16$ a: $1/4$ i: $1/8$ u: $1/8$

现在我们来求联合熵为多少？

2.2 信息论基础

求联合熵可以有几种方法， 以下我们采用连锁规则方法可以得到：

$$\begin{aligned} H(C) &= - \sum_{c=p,t,k} p(c) \log p(c) = -2 \times \frac{1}{8} \times \log \frac{1}{8} - \frac{3}{4} \times \log \frac{3}{4} \\ &= \frac{9}{4} - \frac{3}{4} \log 3 \approx 1.061 \quad (\text{bits}) \end{aligned}$$

$$\begin{aligned} H(V | C) &= \sum_{c=p,t,k} p(C=c) H(V | C=c) \\ &= \frac{1}{8} H\left(\frac{1}{2}, \frac{1}{2}, 0\right) + \frac{3}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{8} H\left(\frac{1}{2}, 0, \frac{1}{2}\right) = \frac{11}{8} = 1.375 \quad (\text{bits}) \end{aligned}$$

2.2 信息论基础

因此,

$$\begin{aligned} H(C, V) &= H(C) + H(V | C) \\ &= \frac{9}{4} - \frac{3}{4} \log 3 + \frac{11}{8} \approx 2.44 \quad (\text{bits}) \end{aligned}$$

2.2 信息论基础

◆ 熵率(entropy rate):

- 一般地, 对于一条长度为 n 的信息, 每一个字符或字的熵为:

$$H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n}) \quad (18)$$

这个数值我们也称为**熵率**。其中, 变量 X_{1n} 表示随机变量序列 $(X_1, \cdots X_n)$, $x_{1n} = x_1, \cdots, x_n$, 有时写成: $x_1^n = (x_1, \cdots x_n)$ 。

2.2 信息论基础

例如：

为传播科学知识、弘扬科学精神、宣传科学思想和科学方法，增进公众对科学的理解，5月20日中国科学院举办了“公众科学日”科普开放日活动。

- $n = 66$ (每个数字、标点均按一个汉字计算)
- $x_{1n} = (\text{为, 传, 播,, 活, 动, 。})$
- $H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{66} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$

2.2 信息论基础

◆ 相对熵(relative entropy) (或称Kullback-Leibler divergence, KL 距离)

两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为:

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (19)$$

该定义中约定 $0 \log(0/q) = 0$, $p \log(p/0) = \infty$

2.2 信息论基础

相对熵常被用以衡量两个随机分布的差距。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。

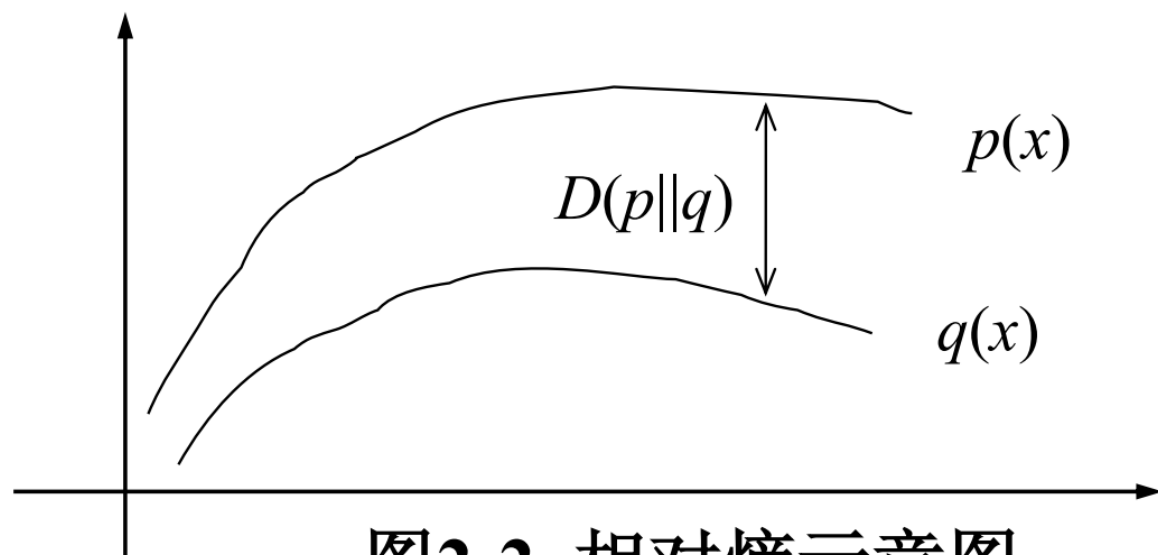


图2-2. 相对熵示意图

2.2 信息论基础

◆ 交叉熵(cross entropy):

如果一个随机变量 $X \sim p(x)$, $q(x)$ 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:

$$H(X, q) = H(X) + D(p||q) = -\sum_x p(x) \log q(x) \quad (20)$$

交叉熵的概念用以衡量估计模型与真实概率分布之间的差异。

2.2 信息论基础

◆ 交叉熵(cross entropy):

对于语言 $L = (X_i) \sim p(x)$ 与其模型 q 的交叉熵定义为:

$$H(L, q) = - \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x_1^n} p(x_1^n) \log q(x_1^n) \quad (21)$$

其中, $x_1^n = x_1, \dots, x_n$ 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

2.2 信息论基础

◆ 交叉熵(cross entropy):

假设这种语言是“理想”的，即 n 趋于无穷大时，其全部“单词”的概率之和为1。据信息论的定理：假定语言 L 是稳态(stationary ergodic)随机过程， x_1^n 为 L 的样本， L 与其模型 q 的交叉熵计算公式为：

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


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

2.2 信息论基础

◆ 交叉熵(cross entropy):

由此，我们可以根据模型 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，我们的目的是使交叉熵最小，从而使模型最接近真实的概率分布 $p(x)$ 。

2.2 信息论基础

◆ 困惑度(perplexity):

在设计语言模型时，我们通常用困惑度来代替交叉熵衡量语言模型的好坏。给定语言 L 的样本 $l_1^n = l_1 \cdots l_n$ ， L 的困惑度 PP_q 定义为：

$$PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n} \log q(l_1^n)} = [q(l_1^n)]^{-\frac{1}{n}} \quad (23)$$

语言模型设计的任务就是寻找困惑度最小的模型，使其最接近真实的语言。

2.2 信息论基础

◆ 互信息(mutual information):

如果 $(X, Y) \sim p(x, y)$, X, Y 之间的互信息 $I(X; Y)$ 定义为:

$$I(X; Y) = H(X) - H(X|Y) \quad (24)$$

根据 $H(X)$ 和 $H(X|Y)$ 的定义:

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

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

2.2 信息论基础

◆ 互信息(mutual information):

$$\begin{aligned} I(X; Y) &= H(X) - H(X | Y) \\ &= -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x | y) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x, y) (\log_2 p(x | y) - \log_2 p(x)) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \left(\log_2 \frac{p(x | y)}{p(x)} \right) \\ I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (25) \end{aligned}$$

互信息 $I(X; Y)$ 是在知道了 Y 的值以后 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。

2.2 信息论基础

◆ 互信息(mutual information):

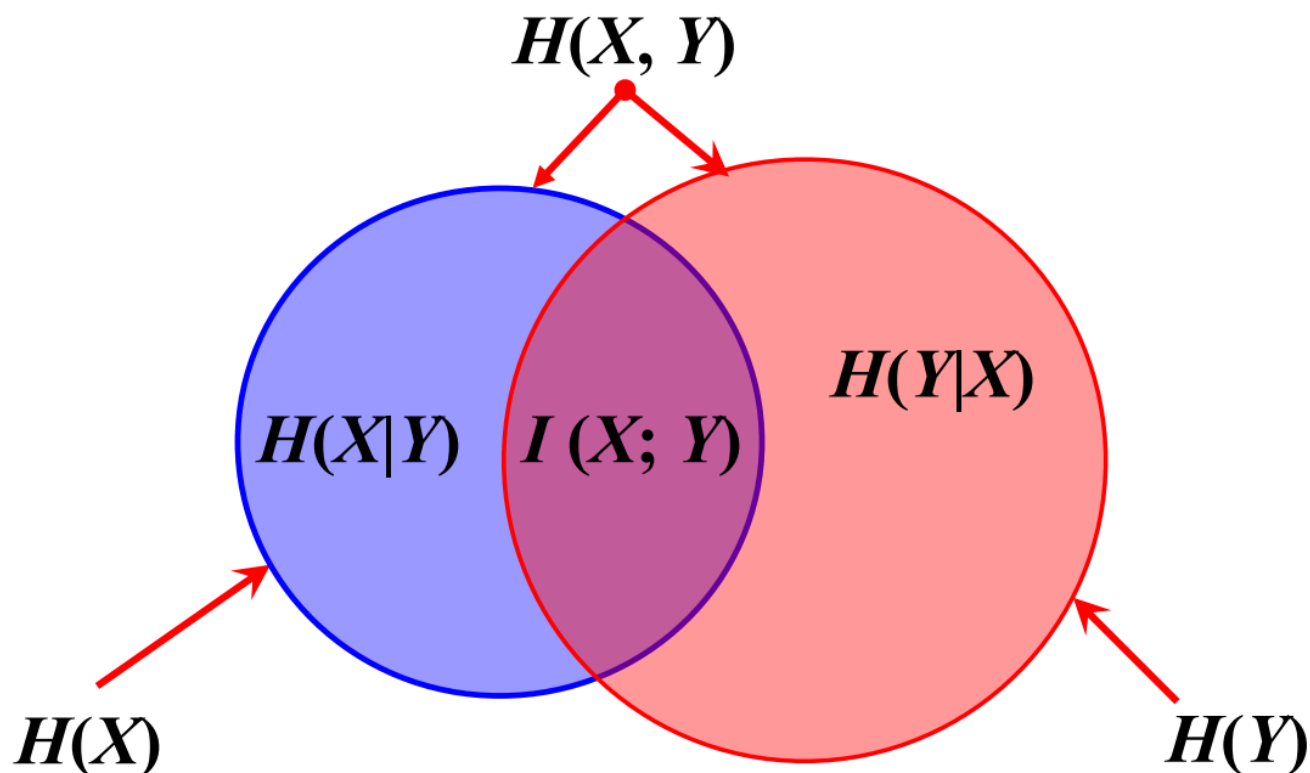


图 2-3. 互信息、条件熵与联合熵

2.2 信息论基础

◆ 互信息(mutual information):

由于 $H(X|X) = 0$, 所以

$$H(X) = H(X) - H(X|X) = I(X; X) \quad (26)$$

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一个常量，而是取决于它们的熵。

2.2 信息论基础

◆ 互信息(mutual information):

➤ 例如：汉语分词问题

为 人 民 服 务。
?

利用互信息值估计两个汉字结合的程度:

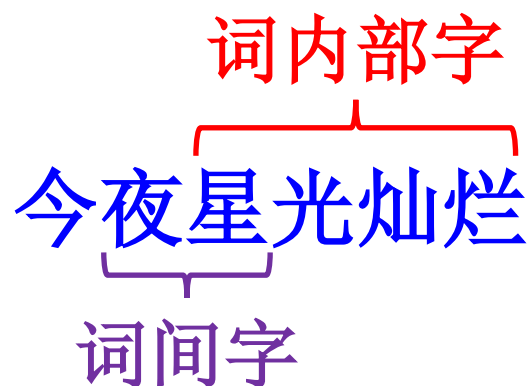
$$I(x; y) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{p(y|x)}{p(y)}$$

互信息值越大，表示两个汉字之间的结合越紧密，越可能成词。反之，断开的可能性越大。

2.2 信息论基础

◆ 互信息(mutual information):

- 也就是说，词与词之间两个临近字的互信息应该小于词内部相邻字之间的互信息。



- 当两个汉字 x 和 y 关联度较强时，其互信息值 $I(x, y) > 0$ ； x 与 y 关系弱时， $I(x, y) \approx 0$ ；而当 $I(x, y) < 0$ 时， x 与 y 称为“互补分布”。

2.2 信息论基础

◆ 互信息(mutual information):

- 在汉语分词研究中，有学者用**双字耦合度**的概念代替互信息：

设 c_i, c_{i+1} 是两个连续出现的汉字，统计样本中 c_i, c_{i+1} 连续出现在一个词中的次数和连续出现的总次数，二者之比就是 c_i, c_{i+1} 的双字耦合度：

$$Couple(c_i, c_{i+1}) = \frac{N(c_i, c_{i+1})}{N(c_i, c_{i+1}) + N(\cdots c_i | c_{i+1} \cdots)}$$

注意：此处 “|” 不表示条件概率！

2.2 信息论基础

◆ 互信息(mutual information):

其中 c_i, c_{i+1} 是一个有序字对, 表示两个连续汉字, 且 $c_i c_{i+1}$ 不等于 $c_{i+1} c_i$ 。 $N(c_i c_{i+1})$ 表示字符串 $c_i c_{i+1}$ 构成的词出现的频率, $N(\cdots c_i | c_{i+1} \cdots)$ 表示 c_i 作为上一个词的词尾且 c_{i+1} 作为相邻下一个词的词头出现的频率。

例如:

“为人” 出现5次, “为人民” 出现20次, 那么,
 $Couple(\text{为}, \text{人}) = 0.2$ 。

2.2 信息论基础

◆ 互信息(mutual information):

理由： 互信息是计算两个汉字连续出现在一个词中的概率，而两个汉字在实际应用中出现的概率情况共有三种：

- (1) 两个汉字连续出现，并且在一个词中；
- (2) 两个汉字连续出现，但分属于两个不同的词；
- (3) 非连续出现。

2.2 信息论基础

◆ 互信息(mutual information):

有些汉字在实际应用中出现虽然比较频繁，但是连续在一起出现的情况比较少，一旦连在一起出现，就很可能是一个词。这种情况下计算出来的互信息会比较小，而实际上两者的结合度应该是比较高的。而双字耦合度恰恰计算的是两个连续汉字出现在一个词中的概率，并不考虑两个汉字非连续出现的情况。

2.2 信息论基础

◆ 互信息(mutual information):

例如：“教务”以连续字符串形式在统计样本中共出现了16次，而“教”字和“务”字分别出现了14945次、6015次。(教, 务)的互信息只有-0.5119。如果用互信息来判断该字对之间位置的切分，是要断开的。但实际上，字对(教, 务)在文本集中出现的16次全部都是“教务”、“教务长”、“教务处”这几个词。连续字对(教, 务)的双字耦合度是1。因此，在判断两个连续汉字之间的结合强度方面，双字耦合度要比互信息更合适一些。

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

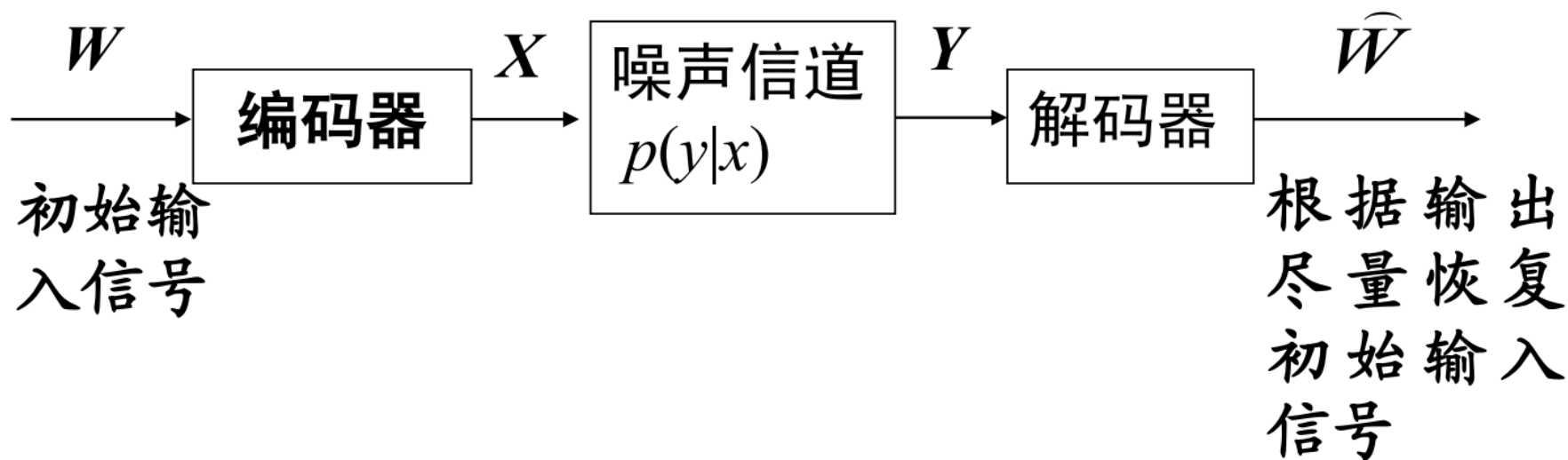
在信号传输的过程中都要进行双重性处理：一方面要通过压缩消除所有的冗余，另一方面又要通过增加一定的可控冗余以保障输入信号经过噪声信道后可以很好地恢复原状。信息编码时要尽量占用少量的空间，但又必须保持足够的冗余以便能够检测和校验错误。接收到的信号需要被解码使其尽量恢复到原始的输入信号。

噪声信道模型的目标就是优化噪声信道中信号传输的吞吐量和准确率，其基本假设是一个信道的输出以一定的概率依赖于输入。

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

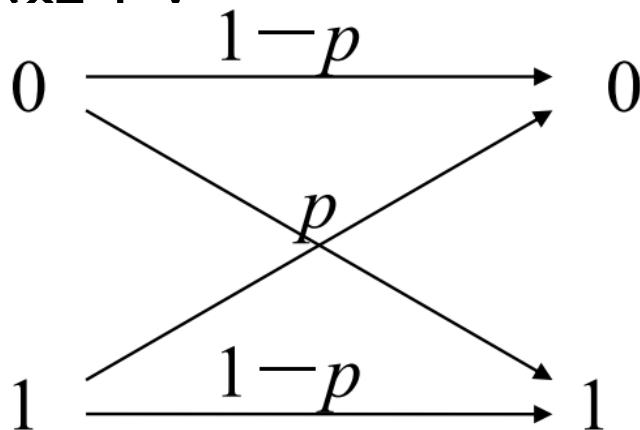
过程示意图:



2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

一个二进制的对称信道 (binary symmetric channel, BSC) 的输入符号集 $X: \{0, 1\}$, 输出符号集 $Y: \{0, 1\}$ 。在传输过程中如果输入符号被误传的概率为 p , 那么, 被正确传输的概率就是 $1 - p$ 。这个过程我们可以用一个对称的图型表示如下:



2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

信息论中很重要的一个概念就是信道容量(capacity), 其基本思想是用降低传输速率来换取高保真通讯的可能性。其定义可以根据互信息给出:

$$C = \max_{p(x)} I(X; Y) \quad (27)$$

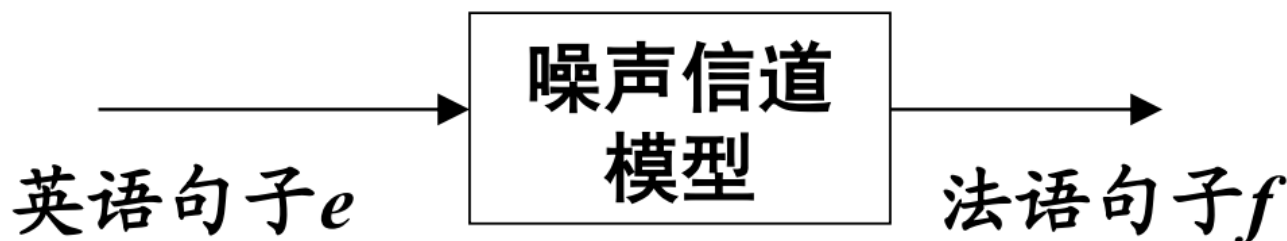
根据这个定义, 如果我们能够设计一个输入编码 X , 其概率分布为 $p(X)$, 使其输入与输出之间的互信息达到最大值, 那么, 我们的设计就达到了信道的最大传输容量。

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

在自然语言处理中，我们不需要进行编码，只需要进行解码，使系统的输出更接近于输入。

例如，法语翻译成英语：



根据贝叶斯公式：

$$p(e|f) = \frac{p(e) \times p(f|e)}{p(f)}$$

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

求该式的最大值相当于寻找一个使得右边分子的两项乘积 $p(e) \times p(f|e)$ 最大, 即:

$$\hat{e} = \underset{e}{\operatorname{argmax}} p(e) \times p(f|e) \quad (28)$$

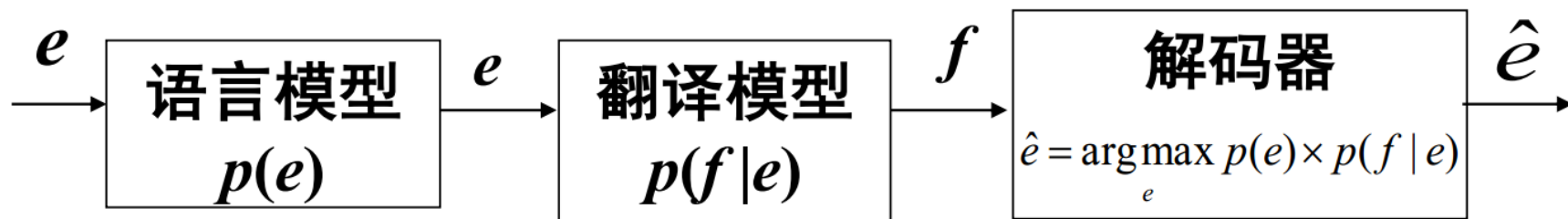
语言模型

翻译模型

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

统计翻译系统框架:



法语句子 f \Longrightarrow 英语句子 \hat{e}

2.2 信息论基础

◆ 噪声信道模型(noisy channel model):

如果我们要建立一个源语言 f 到目标语言 e 的统计翻译系统，我们必须解决三个关键的问题：

- (1) 估计语言模型概率 $p(e)$;
- (2) 估计翻译概率 $p(f|e)$;
- (3) 设计快速有效的搜索算法求解 \hat{e} ，使得 $p(e) \times p(f|e)$ 最大。

2.3 应用实例

2.3 应用实例

◆ 词汇歧义消解（例2-6）

➤ 问题的提出

任何一种自然语言中，一词多义（歧义）现象是普遍存在的。如何区分不同上下文中的词汇语义，就是词汇歧义消解问题，或称词义消歧 (Word Sense Disambiguation, WSD)。

词义消歧是自然语言处理中的基本问题之一。

2.3 应用实例

例如：

- | | |
|-------------------------|------------------------|
| (1) 他 打 鼓很在行。 | (9) 她会用毛线 打 毛衣。 |
| (2) 他会 打 家具。 | (10) 他用尺子 打 个格。 |
| (3) 他把碗 打 碎了。 | (11) 他 打 开了箱子盖。 |
| (4) 他在学校 打 架了。 | (12) 她 打 着伞走了。 |
| (5) 他很会与人 打 交道。 | (13) 他 打 来了电话。 |
| (6) 他用土 打 了一堵墙。 | (14) 他 打 了两瓶水。 |
| (7) 用面 打 浆糊贴对联。 | (15) 他想 打 车票回家。 |
| (8) 他 打 铺盖卷儿走人了。 | (16) 他以 打 鱼为生。 |

2.3 应用实例

(1) 基于上下文分类的消歧方法

—基于贝叶斯分类器 (Gale et al., 1992)基本思路

➤ 数学描述


假设某个多义词 w 所处的上下文语境为 C ，如果 w 的多个语义记作 $s_i (i \geq 2)$ ，那么，可通过计算 $\underset{s_i}{\operatorname{argmax}} p(s_i|C)$ 确定 w 的词义。

$$\text{根据贝叶斯公式: } p(s_i|C) = \frac{p(s_i) \times p(C|s_i)}{p(C)}$$

2.3 应用实例

(1) 基于上下文分类的消歧方法

考虑分母的不变性，并运用如下独立性假设：

$$p(C|s_i) = \prod_{v_k \in C} p(v_k|s_i)$$


出现在上下文中的词

因此，

$$\hat{s}_i = \operatorname{argmax}_{s_i} [p(s_i) \prod_{v_k \in C} p(v_k|s_i)] \quad (29)$$

概率 $p(v_k|s_i)$ 和 $p(s_i)$ 都可用最大似然估计求得：

2.3 应用实例

(1) 基于上下文分类的消歧方法

$$p(v_k | s_i) = \frac{N(v_k, s_i)}{N(s_i)} \quad (30)$$

其中, $N(s_i)$ 是在训练数据中词 w 用于语义 s_i 时的次数, 而 $N(v_k, s_i)$ 为 w 用于语义 s_i 时词 v_k 出现在 w 的上下文中的次数。

$$p(s_i) = \frac{N(s_i)}{N(w)} \quad (31)$$

$N(w)$ 为多义词 w 在训练数据中出现的总次数。

2.3 应用实例

(1) 基于上下文分类的消歧方法

举例说明: $p(s_i) = \frac{N(s_i)}{N(w)}$

对于“打”字而言，假设做实词用的25个语义分别标记为： $s_1 \cdots s_{25}$ ，两个虚词语义分别标记为： s_{26} 、 s_{27} 。假设 s_1 的语义为“敲击(beat)”。那么， $N(s_1)$ 表示“打”字的意思为“敲击(beat)”时在所有统计样本中出现的次数； $N(v_k, s_1)$ 表示某个词 v_k 出现在 s_1 的上下文中时出现的次数。例如，句子：

他 对 打 鼓 很 在 行 。 (取上下文: ± 2)
-2 -1 \uparrow +1 +2

2.3 应用实例

(1) 基于上下文分类的消歧方法

他 对 打 鼓 很 在 行 。 (取上下文: ± 2)
-2 -1 \uparrow +1 +2

那么, 上下文 $C = (\text{他}, \text{对}, \text{鼓}, \text{很})$ 。如果 $v_k = \text{他}$,
 $N(\text{他}, s_1) = 5$, $N(s_1) = 100$, 则

$$p(v_k | s_i) = p(\text{他} | s_1) = \frac{N(\text{他}, s_1)}{N(s_1)} = \frac{5}{100} = 0.05$$

假若 “打” 在所有样本中总共出现了800次, 则

$$p(s_i) = \frac{N(s_i)}{N(w)} = \frac{N(s_1)}{N(\text{打})} = \frac{100}{800} = 0.125$$

2.3 应用实例

(1) 基于上下文分类的消歧方法

➤ 消歧算法描述

对于多义词 w 的每个语义 s_i 执行如下循环：

(1) 对于词典中所有的词 v_k 利用训练语料

计算

$$p(v_k | s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$

(2) 对于 w 的每个语义 s_i 计算：

$$p(s_i) = \frac{N(s_i)}{N(w)}$$

利用已标注的大规模数据——训练过程

2.3 应用实例

(1) 基于上下文分类的消歧方法

➤ 消歧算法描述

(3) 对于 w 的每个语义 s_i 计算 $p(s_i)$, 并根据上下文中的每个词 v_k 计算 $p(w|s_i)$, 选择:

$$\hat{s}_i = \operatorname{argmax}_{s_i} \left[p(s_i) \prod_{v_k \in C} p(v_k | s_i) \right]$$

标注过程或
测试过程

说明: 在实际算法实现中, 通常将概率 $p(w|s_i)$ 和 $p(s_i)$ 的乘积运算转换为对数加法运算:

$$\hat{s}_i = \operatorname{argmax}_{s_i} \left[\log p(s_i) + \sum_{v_k \in C} \log p(v_k | s_i) \right]$$

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 基本思想:

在只掌握关于未知分布的部分知识的情况下，符合已知知识的概率分布可能有多多个，但使熵值最大的概率分布最真实地反映了事件的分布情况，因为熵定义了随机变量的不确定性，当熵最大时，随机变量最不确定，最难准确地预测其行为。也就是说，在已知部分知识的前提下，关于未知分布最合理的推断应该是符合已知知识最不确定或最大随机的推断。

2.3 应用实例

(2) 基于最大熵的消歧方法

对于词义消歧问题来说，确定一个多义词的某个义项可以看成是一个事件 a ，多义词周围（上下文）出现的词及其词性看成是这个事件发生的条件 b 。利用条件熵 $H(a|b)$ 最大时的概率 $p(a|b)$ 推断多义词使用某一义项的可能性。

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 问题描述:

用 A 表示某一多义词所有义项的集合, B 为所有上下文的集合。可定义 $\{0, 1\}$ 域上的二值函数 $f(a, b)$ 来表示上下文条件与义项之间的关系:

$$f(a, b) = \begin{cases} 1 & \text{若}(a, b) \in (A, B), \text{且满足某种条件} \\ 0 & \text{否则} \end{cases}$$

2.3 应用实例

(2) 基于最大熵的消歧方法

如果有特征函数 $f_j(a, b)$, 它在已知样本中的经验概率分布 $\tilde{p}(a, b)$ 可由下式计算得出:

$$\tilde{p}(a, b) \approx \frac{Count(a, b)}{\sum_{A, B} Count(a, b)}$$

其中, $Count(a, b)$ 为 (a, b) 在训练语料中出现的次数。
 f_j 在训练样本中关于经验概率分布的数学期望为:

$$E_{\tilde{p}}(f_j) = \sum_{a, b} \tilde{p}(a, b) f_j(a, b) \quad (32)$$

2.3 应用实例

(2) 基于最大熵的消歧方法

假设所建模型的概率分布为 $p(a, b)$, 则特征 f_i 关于 $p(a, b)$ 的数学期望为

$$E_p(f_i) = \sum_{a,b} \underline{p(a, b)} f_i(a, b) \quad (33)$$

而 $p(a, b) = p(b)p(a|b)$ 。由于所建立模型应该符合已知中的概率分布, 如果用 $\tilde{p}(b)$ 表示 b 在已知样本中的概率分布, 那么可令 $p(b) = \tilde{p}(b)$, 因此(33)式可以变为

$$E_p(f_i) = \sum_{a,b} \underline{\tilde{p}(b)p(a|b)} f_i(a, b) \quad (34)$$

2.3 应用实例

(2) 基于最大熵的消歧方法

如果特征 f_i 对所建模型是有用的，那么(34)式所表示的特征 f_i 的数学期望与它在已知样本的数学期望值是相同的，即

$$E_p(f_i) = E_{\tilde{p}}(f_i) \quad (35)$$

(35)式称为该问题建模的**约束方程**，简称**约束**。

2.3 应用实例

(2) 基于最大熵的消歧方法

假设有 $k(k > 0)$ 个特征 $f_i(i = 1, 2, \dots, k)$ ，它们都在建模过程中对输出有影响，我们所建立的模型应满足所有这些特征，即所建立的模型 p 应该属于这 k 个特征约束下所产生的所有模型的集合 C ：

$$C = \{p \in \Gamma | E_p(f_j) = E_{\tilde{p}}(f_j), j \in \{1, 2 \dots k\}\} \quad (36)$$

其中， Γ 表示所有无条件或无约束的概率模型空间，则 C 是在加入特征约束条件后得到的一个概率模型子集。

2.3 应用实例

(2) 基于最大熵的消歧方法

根据条件熵的定义

$$\begin{aligned} H(p) &= H(A|B) \\ &= \sum_{b \in B} p(b) H(A|B = b) \\ &= - \sum_{a,b} p(b) p(a|b) \log p(a|b) \end{aligned}$$

由于所建模型的概率分布 $p(b)$ 应符合已知样本中的概率分布 $\tilde{p}(b)$, 即: $p(b) = \tilde{p}(b)$, 因此,

$$H(p) = - \sum_{a,b} \tilde{p}(b) p(a|b) \log p(a|b)$$

2.3 应用实例

(2) 基于最大熵的消歧方法

根据(37)式指出的求解本问题的基本思想, 在 k 个约束条件的前提下, 那么, 具有使 $H(p)$ 最大的条件概率模型用于推断使用某一义项的可能性, 即

$$\begin{aligned} p^*(a|b) &= \operatorname{argmax}_{p \in P} H(p) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{a,b} \tilde{p}(b) p(a|b) \log p(a|b) \right) \end{aligned}$$

或是使熵值最大, 并符合约束条件的概率分布。

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 总结

根据最大熵方法的基本思路，估计概率 $p(a|b)$ 时应满足如下两个基本约束：

① $p^* = \operatorname{argmax}_{p \in P} H(p)$

② P : 所建模型中的概率分布 p 应与已知样本中的概率分布相吻合。

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 总结

基于约束条件，求解使 $H(p)$ 值最大的条件概率 $p^*(a|b)$:

$$\begin{aligned} p^*(a|b) &= \operatorname{argmax}_{p \in P} H(p) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{a,b} \tilde{p}(b) p(a|b) \log p(a|b) \right) \end{aligned}$$

目标函数

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 关于最大熵方法在NLP中的应用，请参阅：

1. A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity Resolution, PhD Dissertation, UPenn., 1998
2. A. Ratnaparkhi. A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical Report IRCS-97-08, Dept. of Computer Science, UPenn., 1997
3. J. N. Darroch, D. Ratcliff. Generalized Iterative Scaling for Log-linear Models. Annals of Math. Statistics, 1972, 43: 1470-1480
4. Rosenfeld R. A maximum entropy to adaptive statistical language learning[J]. Computer Speech and Language, 1996, 10(3): 187-228
5. 张仰森：面向语言资源建设的汉语词义消歧与标注方法研究，北京大学博士后出站报告， 2006年12月

2.3 应用实例

(2) 基于最大熵的消歧方法

➤ 相关开源工具：

[1] OpenNLP: <http://incubator.apache.org/opennlp/>

[2] 张乐: <http://homepages.inf.ed.ac.uk/lzhang10/maxent.html>

[3] Malouf: <http://tadm.sourceforge.net/>

[4] Tsujii: <http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/>

[5] 林德康: <http://webdocs.cs.ualberta.ca/~lindek/downloads.htm>

本章小结

◆ 概率论基础

概率 vs. 条件概率

二项式分布

贝叶斯决策理论

最大似然估计

贝叶斯公式

期望 vs. 方差

◆ 信息论基本概念

熵

互信息

交叉熵

噪声信道模型

联合熵

相对熵

困惑度

习题

- 2-1. 任意摘录一段文字，统计这段文字中所有字符的相对频率。假设这些相对频率就是这些字符的概率，请计算其分布的熵。
- 2-2. 任意取另外一段文字（与上题中文字的用字一样），按上述同样的方法计算字符分布的概率，然后计算两段文字中字符分布的 KL 距离。
- 2-3. 举例说明（任意找两个分布 p 和 q ），KL 距离是不对称的，即 $D(p \parallel q) \neq D(q \parallel p)$ 。

习题

北京大学计算语言学研究所在(<http://icl.pku.edu.cn/>)
提供部分标注语料，可供学习和研究参考。

谢谢!