

# 第2章 数学基础

---

北京市海淀区中关村东路95号  
邮编：100190



电话：+86-10-8254 4688  
邮件：cqzong@nlpr.ia.ac.cn



# 本章内容

---

 **2.1 概率论基础**

**2.2 信息论基础**

**2.3 应用举例**

**2.4 附录**



## 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)

在自然语言处理中，以句子为处理单位时一般假设句子独立于它前面的其它语句，句子的概率分布近似地符合二项式分布。



# 本章内容

---

2.1 概率论基础

➡ 2.2 信息论基础

2.3 应用举例

2.4 附录



## 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

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：

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (1)$$

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2-1:** 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：（1）等概率出现情况：

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

（2）实际情况：

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

**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

英语单词  
的熵约为  
10 bits。



## 2.2 信息论基础

---

1970年代末期冯志伟教授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，北京航空学院刘源等测定了汉字的信息熵为9.71 比特，汉语词的熵为11.46比特。

规范文本中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之后剔除**专用名词**前后的数据。

专用名词又称命名实体(named entity)，主要指：人名、地名、组织机构名、时间、数字及货币等。



## 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:

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) \quad (2)$$

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:

$$\begin{aligned} H(Y | X) &= \sum_{x \in X} p(x) H(Y | X = x) \\ &= \sum_{x \in X} p(x) \left[ - \sum_{y \in Y} p(y | x) \log_2 p(y | x) \right] \\ &= - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y | x) \end{aligned} \quad (3)$$

$p(x) \cdot p(y|x) = p(x, y)$

## 2.2 信息论基础

将 (2) 式:  $H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$  中的  $\log_2 p(x, y)$  根据概率公式展开:

$$\begin{aligned} H(X, Y) &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x)p(y|x)] \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) [\log p(x) + \log p(y|x)] \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) \\ &= -\sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned} \quad (4) \quad (\text{连锁规则})$$

## 2.2 信息论基础

例2-2：假设 $(X, Y)$ 服从如下联合概率分布：

$Y \backslash X$	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

$Y \backslash X$	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

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

$$= - \left( \frac{1}{2} \times \log_2 \left( \frac{1}{2} \right) + \frac{1}{4} \times \log_2 \left( \frac{1}{4} \right) + \frac{1}{8} \times \log_2 \left( \frac{1}{8} \right) + \frac{1}{8} \times \log_2 \left( \frac{1}{8} \right) \right)$$

$$= \frac{7}{4}$$



## 2.2 信息论基础

类似地，可以计算 $H(Y)$ 。

$Y \backslash X$	1	2	3	4	$p(Y)$
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

$Y \backslash X$	1	2	3	4	$p(Y)$
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

$$\begin{aligned} H(X | Y) &= \sum_{i=1}^4 p(y=i) H(X | Y=i) \\ &= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\ &\quad + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\ &= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 = \frac{11}{8} \quad (\text{bits}) \end{aligned}$$

$-\sum_{i=1}^4 p(x_i | y_1) \log p(x_i | y_1)$

## 2.2 信息论基础

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

$$-\sum_{i=1}^4 p(x_i | y_2) \log p(x_i | y_2)$$

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

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

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

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

可见,  $H(Y|X) \neq H(X|Y)$ 。



## 2.2 信息论基础

**例2-3:** 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：

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



## 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，但是，我们可以近似地将其看作随机变量。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  $C=\{p, t, k\}$ 。



## 2.2 信息论基础

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  $P(C, \bullet)$  和  $P(\bullet, V)$ 如下表所示：

$V \backslash C$	p	t	k	$P(\bullet, V)$
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
$P(C, \bullet)$	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：

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

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



## 2.2 信息论基础

---

因此,

$$H(C, V) = H(C) + H(V | C)$$

$$= \frac{9}{4} - \frac{3}{4} \log 3 + \frac{11}{8} \approx 2.44 \quad (\text{bits})$$



## 2.2 信息论基础

一般地，对于一条长度为  $n$  的信息，每一个字符或字的熵为：

$$H_{\text{rate}} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n}) \quad (5)$$

这个数值我们也称为 **熵率(entropy rate)**。其中，变量  $X_{1n}$  表示随机变量序列  $(X_1, \dots, X_n)$ ， $x_{1n} = (x_1, \dots, x_n)$  表示随机变量的具体取值。有时将  $x_{1n}$  写成： $x_1^n$ 。



## 2.2 信息论基础

例如，有如下文字：

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)
- $x_{1n}=(\text{为}, \text{传}, \text{播}, \dots, \text{活}, \text{动}, \text{。})$
- $H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{66} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$



## 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:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (6)$$

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。

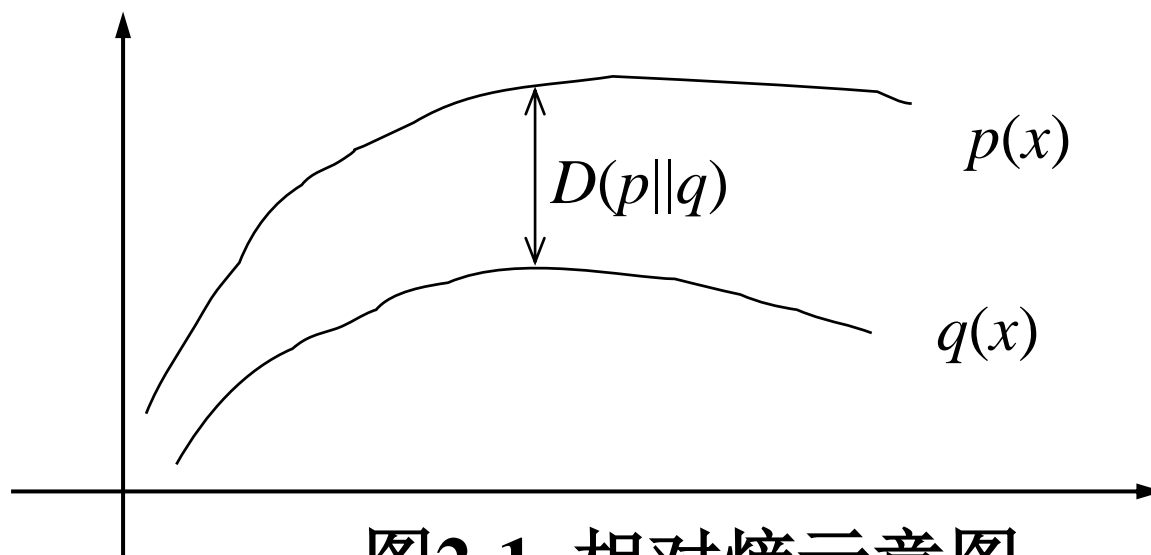


图2-1. 相对熵示意图



## 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:

$$\begin{aligned} H(X, q) &= H(X) + D(p \parallel q) \\ &= -\sum_x p(x) \log q(x) \end{aligned} \quad (7)$$

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

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

$$H(L, q) = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x_1^n} p(x_1^n) \log q(x_1^n) \quad (8)$$

其中， $x_1^n = x_1, \dots, x_n$  为语言  $L$  的词序列（样本）；

$p(x_1^n)$  为  $x_1^n$  的概率（理论值）；

$q(x_1^n)$  为模型  $q$  对  $x_1^n$  的概率估计值。



## 2.2 信息论基础

信息论中有如下**定理**：

假定语言  $L$  是稳态(stationary)遍历性(ergodic)随机过程， $x_1^n$  为  $L$  的样本，那么，有：

$$H(L, q) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log q(x_1^n) \quad (9)$$

证明见本章讲义**附录1**。

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：

$$PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n} \log q(l_1^n)} = [q(l_1^n)]^{-\frac{1}{n}} \quad (10)$$

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:

$$I(X; Y) = H(X) - H(X | Y) \quad (11)$$

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

$$\begin{aligned} I(X; Y) &= H(X) - H(X | Y) \\ &= -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x | y) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x, y) (\log_2 p(x | y) - \log_2 p(x)) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \left( \log_2 \frac{p(x | y)}{p(x)} \right) \\ I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x) p(y)} \end{aligned} \quad (12)$$

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

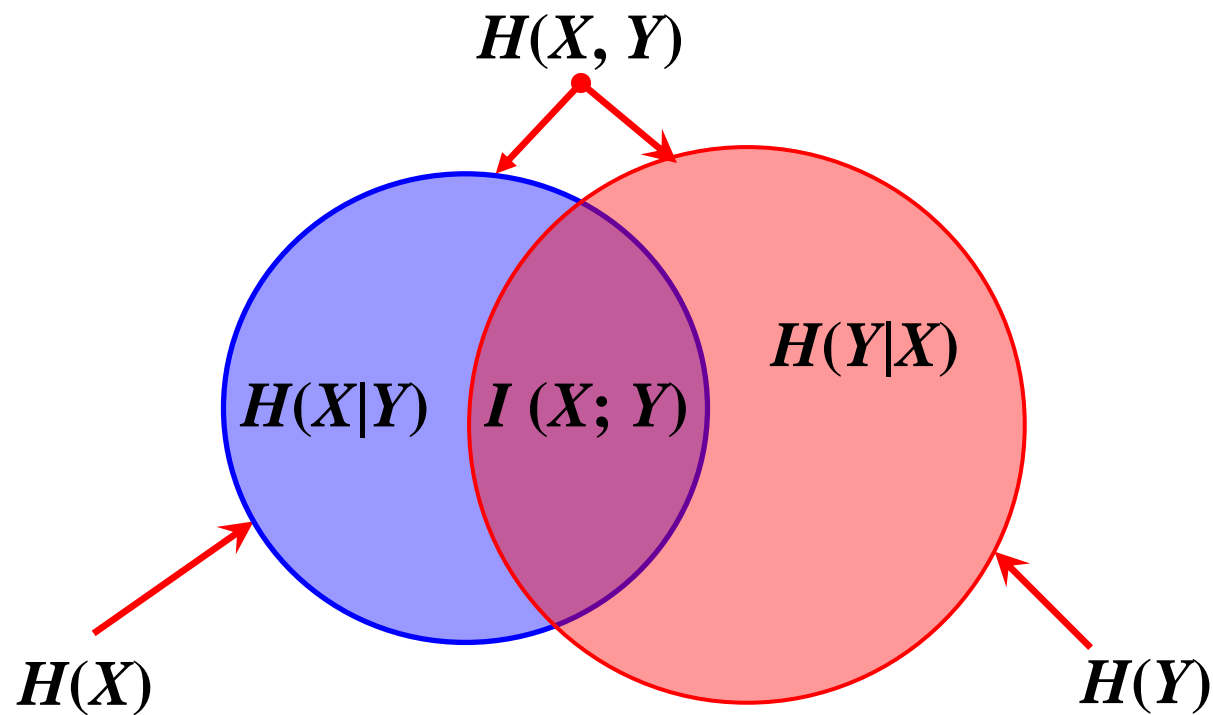


图 2-2. 互信息、条件熵与联合熵



## 2.2 信息论基础

---

由于  $H(X|X) = 0$ , 所以,

$$H(X) = H(X) - H(X|X) = I(X; X) \quad (13)$$

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一个常量，而是取决于它们的熵。

## 2.2 信息论基础

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

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”。

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：

## 2.2 信息论基础

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

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， $N(\cdots c_i \mid c_{i+1} \cdots)$  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， $\text{Couple}(\text{为}, \text{人}) = 0.2$ 。

**注意：此处“|”不表示条件概率！**



## 2.2 信息论基础

**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。

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

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”字出现了14 945次，“务”字出现了6 015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

**说明：**两个单个离散事件 $(x_i, y_j)$ 之间的互信息 $I(x_i, y_j)$ 通常称为点式互信息(point-wise mutual information)，点式互信息可能为负值。两个随机变量 $(X, Y)$ 之间的互信息 $I(X, Y)$ 称为平均互信息，平均互信息不可能为负值。

关于两个随机变量之间平均互信息为非负值的证明见本课件**附录2**。



## 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

过程示意图：

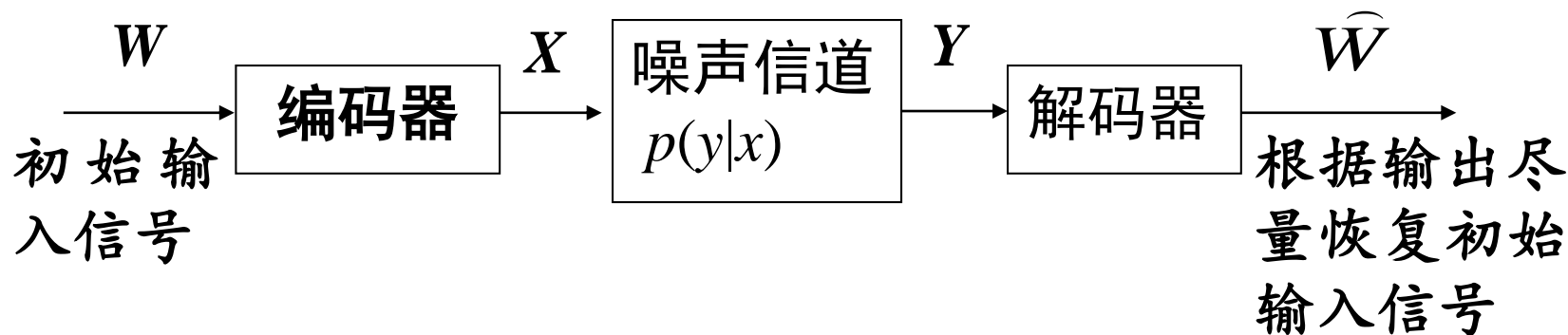


图 2-3. 噪声信道模型示意图

## 2.2 信息论基础

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:

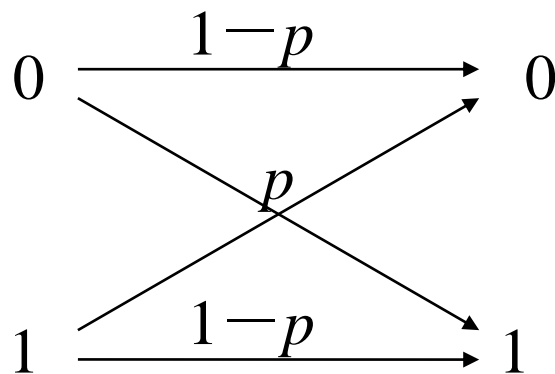


图 2-4. 对称信道





## 2.2 信息论基础

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:

$$C = \max_{p(X)} I(X; Y) \quad (14)$$

据此定义, 如果我们能够设计一个输入编码  $X$ , 其概率分布为  $p(X)$ , 使其输入和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。

在语言处理中, 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, 如机器翻译。



# 本章内容

---

2.1 概率论基础

2.2 信息论基础

 **2.3 应用举例**

**2.4 附录**



## 2.3 应用举例

### 例2-4: 词汇歧义消解

#### ❖ 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 应用举例

例如：

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

## 2.3 应用举例

## ❖ 基本思路

每个词表达不同的含意时其上下文（语境）往往不同，也就是说，不同的词义对应不同的上下文，因此，如果能够将多义词的上下文区别开，其词义自然就明确了。

他/P 很/D 会/V 与/C 人/N 打/V 交道/N 。 /PU  
 -2      -1      0      +1      +2

## 基本的上下文信息：词、词性、位置



## 2.3 应用举例

### ❖ 实现方法

#### (1) 基于贝叶斯分类器(Gale *et al.*, 1992)

##### ● 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, 可以通过计算  $\arg \max_{s_i} p(s_i | C)$  确定  $w$  的词义。

## 2.3 应用举例

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

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，

$$\hat{s}_i = \arg \max_{s_i} \left[ p(s_i) \prod_{v_k \in C} p(v_k | s_i) \right] \quad (15)$$

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



## 2.3 应用举例

---

$$p(v_k | s_i) = \frac{N(v_k, s_i)}{N(s_i)} \quad (16)$$

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。

$$p(s_i) = \frac{N(s_i)}{N(w)} \quad (17)$$

$N(w)$  为多义词  $w$  在训练数据中出现的总次数。



## 2.3 应用举例

举例说明：
$$p(v_k | s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$

对于“打”字而言，假设做实词用的25个语义分别标记为： $s_1 \sim s_{25}$ ，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：  $\pm 2$ )  
-2 -1  $\uparrow$  +1 +2



## 2.3 应用举例

他对打鼓很在行。(取上下文:  $\pm 2$ )  
-2 -1  $\uparrow$  +1 +2

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

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

假若“打”在所有样本中总共出现了800次, 那么,

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

## 2.3 应用举例

- 算法描述：

- ①对于多义词  $w$  的每个语义  $s_i$  执行如下循环：  
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：

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

模  
数  
据  
—  
训  
练  
过  
程  
利  
用  
已  
标  
注  
的  
大  
规

## 2.3 应用举例

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：

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

标注过程或  
称测试过程

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

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



## 2.3 应用举例

### (2) 基于最大熵的消歧方法

- 数学描述：

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个，但使熵值最大的概率分布能够最真实地反映事件的分布情况，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。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 应用举例

对于求解的问题，就是估计在条件  $b \in B$  下(已知知识)，发生某个事件  $a \in A$ (未知分布) 的概率  $p(a|b)$ ，该概率使熵  $H(p(A|B))$  最大。

经推导(见本章附录3)，有：

$$p^*(a|b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (18)$$

特征函数

特征权重

$$\text{其中, } Z(b) = \sum_a \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (19)$$

$Z(b)$  为保证对所有  $b$ ，使得  $\sum_a p(a|b) = 1$  的归一常量。



## 2.3 应用举例

- 确定特征函数

对于词义消歧而言，设  $A$  为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：

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

如：“打”字的义项集合： $A = \{s_1, s_2, s_3, \dots, s_{27}\}$   
 $B = \{\text{“打”字出现的上下文}\}$

## 2.3 应用举例

上下文条件(*b*)表示有:

(1) 词形信息:      他 很 会 与 人 打 交道。

(2) 词性信息: 他/PN 很/D 会/V 与/C 人/N 打 交道/N。/PU

(3) 词形+词性信息:

他/PN 很/D 会/V 与/C 人/N 打 交道/N。/PU



## 2.3 应用举例

存在两种表示方法：

- ①位置无关：目标词周围的词形、词性或其组合构成的集合，如取 $\pm 2$ 窗口范围内的词形：

{与，人，交道，。}  
={交道，与，。，人}

词袋模型  
(通常用向量表示)



他/PN 很/D 会/V 与/C 人/N **打** 交道/N 。/PU

- ②位置有关：词形( $\pm 2$ )： $\langle \text{与}_{-2}, \text{人}_{-1}, \text{交道}_{+1}, \text{。}_{+2} \rangle$

模板表示

$\neq \langle \text{与}_{-2}, \text{。}_{+2}, \text{交道}_{+1}, \text{人}_{-1} \rangle$

## 2.3 应用举例

他/P 很/D 会/V 与/C 人/N 打/V !5\$ 交道/N 。 /PU  
↑

假设以字形向量表示条件，那么，特征函数为：

$$f_1(a,b) = \begin{cases} 1 & \text{If } a = s_5 \text{ and } b = \langle (\text{与}, \text{人}), (\text{交道}, \text{。}) \rangle \\ 0 & \text{Otherwise} \end{cases}$$

$$f_2(a,b) = \begin{cases} 1 & \text{If } a = s_5 \text{ and } b = \langle (\text{C}, \text{N}), (\text{N}, \text{PU}) \rangle \\ 0 & \text{Otherwise} \end{cases}$$

... ..



## 2.3 应用举例

如果上下文条件由如下三类信息表示：

- (1)特征的类型：词形、词性、词形+词性，3种情况；
- (2)上下文窗口大小：当前词的左右2个词，1种情况；
- (3)是否考虑位置：是或否，2种情况。

上述3种情况组合，可得到如下  $n$  种特征模板：

$$n = 3 \times 1 \times 2 = 6$$

考虑到词形、词性又有很多种可能，因此，可得到若干上下文特征构成的条件模板，这就需要筛选。



## 2.3 应用举例

特征选择一般有三种方法：

- ① 从候选特征集中选择那些在训练数据中出现频次超过一定阈值的特征；
- ② 利用互信息作为评价尺度从候选特征集中选择满足一定互信息要求的特征；
- ③ 利用增量式特征选择方法(Della Pietra *et al.*)从候选特征集中选择特征。(比较复杂)

最终选定  $k$  ( $k > 0$ ) 个特征，对应  $k$  个特征函数  $f$ 。在以下叙述中不再区分特征和特征函数。



## 2.3 应用举例

---

回顾前面的公式(18)、(19):

$$p^*(a | b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (18)$$

$$Z(b) = \sum_a \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (19)$$

此处  $l = k+1$ 。



## 2.3 应用举例

- 获取 $\lambda$ 参数

— 利用GIS(generalized interactive scaling) 算法  
GIS 迭代过程要求对于训练集中每个实例的任意  $(a, b) \in A \times B$ ,  $k$  个特征函数之和为一常量  $C$ , 即:

$$\sum_{j=1}^k f_j(a, b) = C$$

若该条件不满足, 则根据训练集取:  $C = \max_{a \in A, b \in B} \sum_{j=1}^k f_j(a, b)$

并增加一个修正特征  $f_l$ :  $f_l(a, b) = C - \sum_{j=1}^k f_j(a, b)$

与其它特征不一样,  $f_l(a, b)$  的取值范围为:  $0 \sim C$ 。

## 2.3 应用举例

- GIS算法描述：

(a) 初始化：  $\lambda[1..l]=0$ ;

(b) 计算每一个特征函数  $f_j$  的训练样本期望值  $E_{\tilde{p}}(f_j)$ ;

(c) 迭代计算特征函数的模型期望值  $E_p(f_j)$ :

①利用公式(18)、(19)

计算概率  $p^*$ ;

$$Z(b) = \sum_a \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right)$$
$$p^*(a|b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right)$$

②若满足终止条件，则结束迭代;

否则，修正  $\lambda$ ，继续下轮迭代。

(d) 算法结束，确定  $\lambda$ ，算出每个  $p^*$ 。



## 2.3 应用举例

迭代终止条件:

- i. 限定迭代次数;
- ii. 对数似然( $L(p)$ )的变化小到可以忽略:

$$|L_{i+1} - L_i| < \varepsilon$$

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

$\tilde{p}(a,b)$  为  $(a, b)$  在训练样本中出现的概率。



## 2.3 应用举例

$\lambda$  修正方法:

$$\lambda^{(n+1)} = \lambda^{(n)} + \frac{1}{C} \ln \left( \frac{E_{\tilde{p}}(f_j)}{E_{p^{(n)}}(f_j)} \right)$$

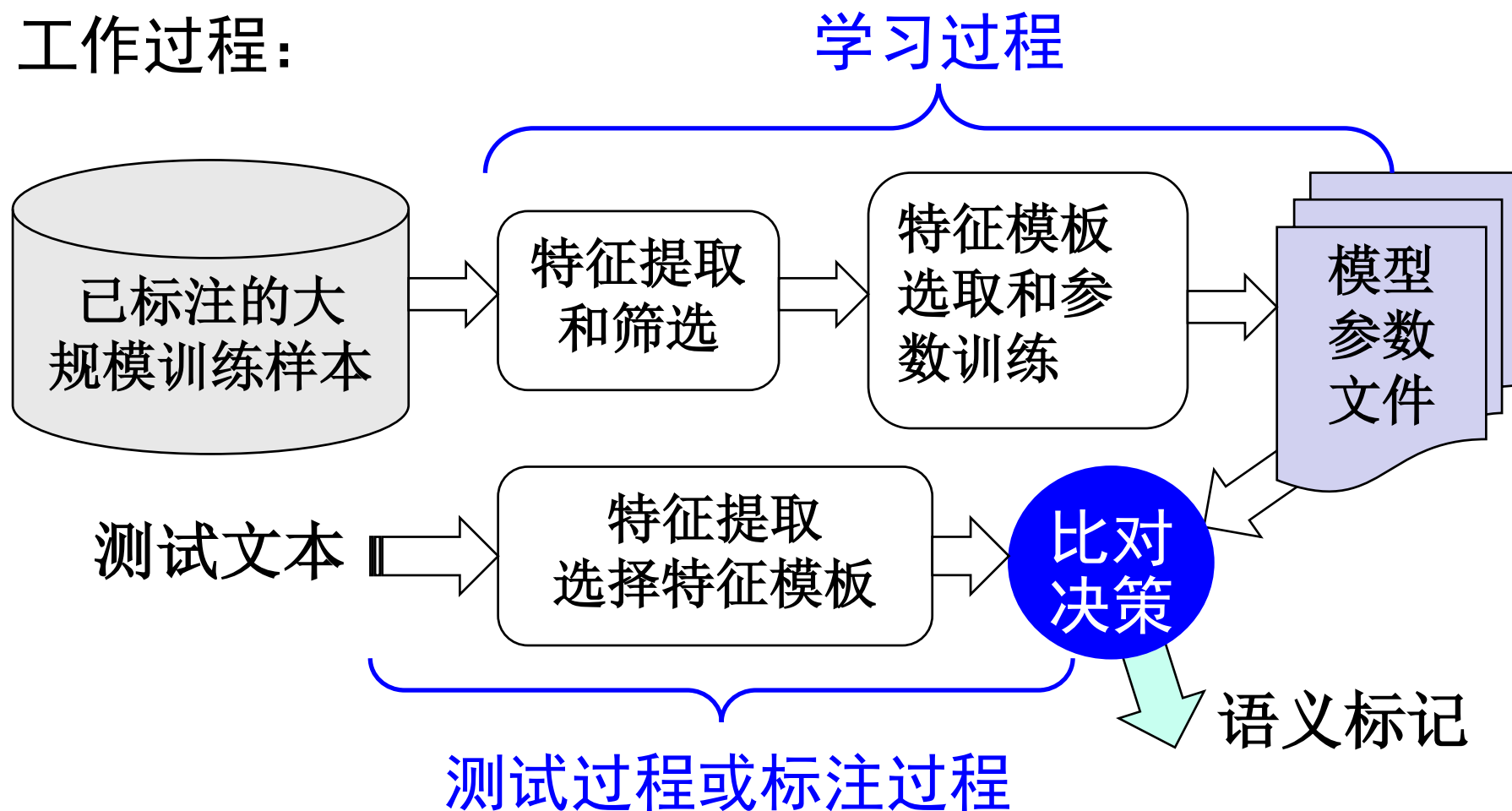
由于  $\lambda$  的收敛速度受  $C$  的取值的影响，因此，人们对 GIS 算法做出了改进：

A. L. Berger. The improved iterative scaling algorithm: A gentle introduction, *Technical report*, Carnegie Mellon University, 1997

D. Pietra et al. Inducing Features of Random Fields, *IEEE Trans. on PAMI*, 1997, 19(4): 380-393

## 2.3 应用举例

工作过程：





## 2.3 应用举例

### ● 实验结果：

- 训练数据：用2000年1月1～28日28天的《人民日报》标注文本作为训练数据（全部进行了词义标注）；
- 测试数据：2000年1月29～31日三天的文本作为测试数据，利用所建立的最大熵模型算法对其进行义项标注实验，多义词有4931个；
- 特征模板：特征类型=词形，窗口大小=全句，不考虑位置特征；
- 标注结果：正确率为 94.34%。



## 2.3 应用举例

关于最大熵方法在NLP中的应用及GIS，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 应用举例

---

### ◆ 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

## ◆ 信息论基础

➤ 熵

➤ 互信息

➤ 交叉熵

➤ 噪声信道模型

➤ 联合熵

➤ 相对熵

➤ 困惑度

## ◆ 应用举例



# 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

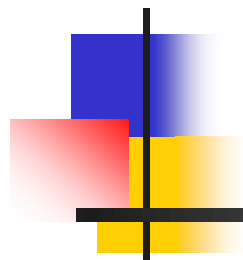
2-4. 设  $X \sim p(x)$ ,  $q(x)$  为用于近似  $p(x)$  的一个概率分布, 则  $p(x)$  与  $q(x)$  的交叉熵定义为:

$H(p, q) = H(p) + D(p \parallel q)$ 。请证明:

$$H(p, q) = -\sum_x p(x) \log q(x)$$

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





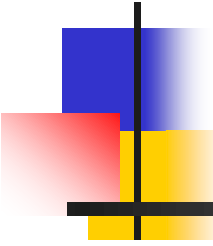
# ***Thanks***

谢谢!



## 2.4 附录

1. 证明前面P34公式(9):  $H(L, q) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log q(x_1^n)$   
(又见《统计自然语言处理》P28, 公式(2-39))
2. 证明前面P45上的结论: 两个随机变量之间的平均互信息为非负值。
3. 前面P62上概率  $p^*(a|b)$  的推导说明



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

布莱曼渐近均分性（Breiman's AEP）定理：如果 $X$ 是稳态的遍历性随机过程，那么

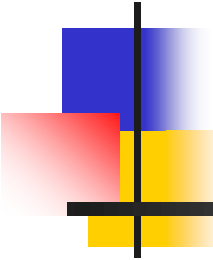
$$H_{\text{rate}}(X) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log p(x_1, x_2, \dots, x_n) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log p(x_1^n)$$

### 该定理的证明：

假设  $(x_1, x_2, \dots, x_n)$  符合独立同分布。

根据熵率的定义, 左边：

$$\begin{aligned} H_{\text{rate}}(X) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(x_1^n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} H(x_1, x_2, \dots, x_n) \\ &= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n) \right\} \end{aligned}$$



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

$$\begin{aligned} &= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \sum_{x_1 x_2 \dots x_n} p(x_1, x_2, \dots, x_n) \left( \sum_{x_i} \log p(x_i) \right) \right\} \\ &= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \left[ \sum_{x_1 x_2 \dots x_n} p(x_1, x_2, \dots, x_n) \log p(x_1) + \sum_{x_1 x_2 \dots x_n} p(x_1, x_2, \dots, x_n) \log p(x_2) + \dots \right. \right. \\ &\quad \left. \left. + \sum_{x_1 x_2 \dots x_n} p(x_1, x_2, \dots, x_n) \log p(x_n) \right] \right\} \\ &= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \left[ \sum_{x_1 x_2 \dots x_n} p(x_1) p(x_2) \dots p(x_n) \log p(x_1) + \sum_{x_1 x_2 \dots x_n} p(x_1) p(x_2) \dots p(x_n) \log p(x_2) \right. \right. \\ &\quad \left. \left. + \dots + \sum_{x_1 x_2 \dots x_n} p(x_1) p(x_2) \dots p(x_n) \log p(x_n) \right] \right\} \end{aligned}$$

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

$$\begin{aligned}
 &= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \left[ \sum_{x_1} p(x_1) \log p(x_1) \left( \sum_{x_2 x_3 \dots x_n} p(x_2) p(x_3) \dots p(x_n) \right) \right. \right. \\
 &\quad + \sum_{x_2} p(x_2) \log p(x_2) \left( \sum_{x_1, x_3 x_4 \dots x_n} p(x_1) p(x_3) p(x_4) \dots p(x_n) \right) \\
 &\quad \left. \left. + \dots + \sum_{x_n} p(x_n) \log p(x_n) \left( \sum_{x_1 x_2 \dots x_{n-1}} p(x_1) p(x_2) \dots p(x_{n-1}) \right) \right] \right\}
 \end{aligned}$$

由于  $(x_1, x_2, \dots, x_n)$  符合概率同分布，所以红线部分可以被看作是联合概率分布对所有的可能取值的求和，其值为1。

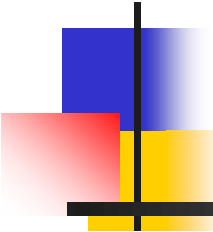
$$= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \left[ \sum_{x_1} p(x_1) \log p(x_1) + \sum_{x_2} p(x_2) \log p(x_2) + \dots + \sum_{x_n} p(x_n) \log p(x_n) \right] \right\}$$

$$= \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \left[ \sum_{x_1} p(x_1) \log p(x_1) + \sum_{x_2} p(x_2) \log p(x_2) + \dots + \sum_{x_n} p(x_n) \log p(x_n) \right] \right\}$$

$$= \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} [H(x_1) + H(x_2) + \dots + H(x_n)] \right\}$$

$$= \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \times n \times H(x_i) \right\}$$

$$= H(x_i)$$



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

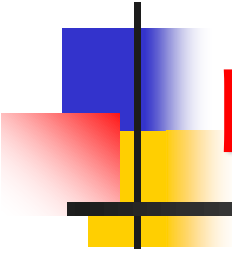
而定理右边：

$$\begin{aligned} -\lim_{n \rightarrow \infty} \frac{1}{n} \log p(x_1^n) &= -\lim_{n \rightarrow \infty} \frac{1}{n} \log p(x_1, x_2, \dots, x_n) \\ &= -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x_i} \log p(x_i) \quad (\text{基于概率同分布}) \end{aligned}$$

该式中， $\frac{1}{n} \sum_{x_i} \log p(x_i)$  可以看作是  $\log p(x_i)$  的均值，而  $E(\log p(x_i))$

为其期望值（相当于下面式子中的  $\mu$ ）。根据**辛钦大数定律** (Wiener-Khinchin Law of Large Numbers)：**样本均值依概率收敛于期望值  $\mu$** ，即

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$



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

因此，

$$\lim_{n \rightarrow \infty} P \left( \left| \frac{1}{n} \sum_{x_i} \log p(x_i) - E(\log p(x_i)) \right| < \varepsilon \right) = 1$$

即

$$-\frac{1}{n} \sum_{x_i} \log p(x_i) \rightarrow -E(\log p(x_i)) = H(x_i)$$

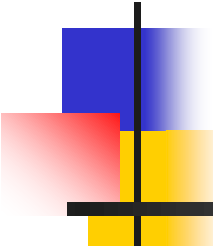
(依概率)

(参见下面的[1]，英文版P58，中文版P33)

所以，左边等于等式右边。

---

[1] Thomas M.Cover and Joy A.Thomas. Elements of Information Theory, 2nd edition, John Wiley & Sons. 2006



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

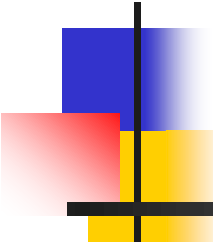
类似地，可以证明在  $(x_1, x_2, \dots, x_n)$  不满足独立同分布的条件下，该定理同样成立。请参阅下面的文献[2]。

根据布莱曼渐近均分性定理(Breiman's AEP)，可以推广到：

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(\log p(x_1^n)) = \lim_{n \rightarrow \infty} \frac{1}{n} \log p(x_1^n)$$

- 
- [2] Paul H. Algoet and Thomas M. Cover. A Sandwich Proof of The Shannon-McMillan-Breiman Theorem. In The Annals of Probability 1998, Vol. 16, No.2 899-909





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

本章讲义前面P34，公式(9)，《统计自然语言处理》P28，公式(2-39)：

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

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

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

(利用布莱曼渐进均分性定理的推广)

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

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

证毕。



## 2.4 附录

1. 证明前面P34公式(9):  $H(L, q) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log q(x_1^n)$

(又见《统计自然语言处理》P28, 公式(2-39))

2. 证明前面P45上的结论: 两个随机变量之间的平均互信息为非负值。

3. 前面P62上概率  $p^*(a|b)$  的推导说明

## 附录2: 证明: 平均互信息为非负值

方法一:

证明: 根据琴生不等式 (Jensen inequality) 的积分形式:

$$\frac{\int_a^b f(g(x)) p(x) dx}{\int_a^b p(x) dx} \geq f\left(\frac{\int_a^b g(x) p(x) dx}{\int_a^b p(x) dx}\right)$$

其中,  $f(x)$  是凸函数,  $g(x)$  为任意函数。那么,

$$\begin{aligned} I(X, Y) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) = \int \int p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) dx dy \\ &\geq -\log \left( \int \int p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) dx dy \right) \\ &= 0 \end{aligned}$$

证毕。

## 附录2: 证明: 平均互信息为非负值

### 方法二:

**证明:** 根据互信息的定义有:  $I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$

那么,

$$-I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log_2 \frac{p(x_i)p(y_j)}{p(x_i y_j)}$$

利用不等式:  $\ln z \leq z-1$ , 且  $\log_2 z = \ln z \cdot \log_2 e$

所以,  $\log_2 z \leq (z-1) \cdot \log_2 e$ ,  $\log_2 \frac{p(x_i)p(y_j)}{p(x_i y_j)} \leq \left[ \frac{p(x_i)p(y_j)}{p(x_i y_j)} - 1 \right] \cdot \log_2 e$

## 附录2: 证明: 平均互信息为非负值

$$\begin{aligned} -I(X; Y) &= \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log_2 \frac{p(x_i)p(y_j)}{p(x_i y_j)} \\ &\leq \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \left[ \frac{p(x_i)p(y_j)}{p(x_i, y_j)} - 1 \right] \log_2 e \\ &= \left[ \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i)p(y_j) - \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i y_j) \right] \log_2 e \\ &= \left[ \sum_{x_i \in X} p(x_i) \sum_{y_j \in Y} p(y_j) - \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i y_j) \right] \log_2 e = 0 \end{aligned}$$

$$I(X; Y) \geq 0$$

证毕。

根据自然对数的性质:  $\ln z \leq z-1, z > 0$ , 当且仅当  $z=1$  时取等号, 因此, 当且仅当  $\frac{p(x_i)p(y_j)}{p(x_i y_j)} = 1$  时, 即  $p(x_i, y_j) = p(x_i)p(y_j)$  时,  $I(X; Y) = 0$ 。



## 2.4 附录

1. 证明前面P34公式(9):  $H(L, q) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log q(x_1^n)$

(又见《统计自然语言处理》P28, 公式(2-39))

2. 证明前面P45上的结论: 两个随机变量之间的平均互信息为非负值。

3. 前面P62上概率  $p^*(a|b)$  的推导说明



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

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  $p^* = \arg \max_{p \in P} H(p)$  (20)

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。



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

根据条件熵的定义(理论值):

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

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,

$$H(p) = - \sum_{a,b} \tilde{p}(b) p(a | b) \log p(a | b) \quad (21)$$





## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

即求解使  $H(p)$  值最大的条件概率  $p^*(a|b)$ :

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

目标函数



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

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:

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

其中,  $\text{Count}(a, b)$  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:

$$E_{\tilde{p}}(f_j) = \sum_{a, b} \tilde{p}(a, b) f_j(a, b) \quad (22)$$



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

假设所建模型的概率分布为  $p(a, b)$ , 则特征  $f_j$  关于  $p(a, b)$  的数学期望(理论值)为:

$$E_p(f_j) = \sum_{a,b} p(a, b) f_j(a, b) \quad (23)$$

由于  $p(a, b) = p(b)p(a|b)$ , 且所建模型应符合已知样本中的概率分布, 即:  $p(b) = \tilde{p}(b)$ , 由此, (23)式变为:

$$E_p(f_j) = \sum_{a,b} \tilde{p}(b) p(a|b) f_j(a, b) \quad (24)$$



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

如果特征  $f_j$  对所建的模型是有用的，那么，所建模型中特征  $f_j$  的数学期望与它在已知样本中的数学期望应该是相同的，即：

$$E_p(f_j) = E_{\tilde{p}}(f_j) \quad (25)$$

该式称为该问题建模的**约束方程**，简称**约束**。



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

假设存在  $k$  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  $P$ ：

$$P = \{ p \mid E_p(f_j) = E_{\tilde{p}}(f_j), j \in \{1, 2, \dots, k\} \} \quad (26)$$



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

根据以上阐述，归纳如下：

$$p^* = \arg \max_{p \in P} H(p) \quad (20)$$

$$H(p) = -\sum_{a,b} \tilde{p}(b) p(a|b) \log p(a|b) \quad (21)$$

$$P = \{ p \mid E_p(f_j) = E_{\tilde{p}}(f_j), j \in \{1, 2, \dots, k\} \} \quad (26)$$

$$E_{\tilde{p}}(f_j) = \sum_{a,b} \tilde{p}(a,b) f_j(a,b) \quad (22)$$

$$E_p(f_j) = \sum_{a,b} \tilde{p}(b) p(a|b) f_j(a,b) \quad (24)$$



## 附录3: 关于概率 $p^*(a|b)$ 的推导说明

这样，问题就变成了在满足一组约束的条件下求最优解的问题，可用拉格朗日乘子法解决此问题。可以证明，满足(26)式约束条件的解具有如下形式：

$$p^*(a|b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (27)$$

$$\text{其中, } Z(b) = \sum_a \exp\left(\sum_{j=1}^l \lambda_j \cdot f_j(a, b)\right) \quad (28)$$

$Z(b)$ 为保证对所有 $b$ ，使得  $\sum_a p(a|b) = 1$  的归一常量。