自然语言处理

赵云蒙
华东理工大学 信息科学与工程学院
能源化工过程智能制造教育部重点实验室
2022-2023 第一学期

第二章 数学基础

2.1 概率论基础

2.1 概率论基础

✦ 基本概念

- 概率
- 最大似然估计
- 条件概率
- 全概率公式
- 贝叶斯法则
- 贝叶斯决策理论
- 二项式分布
- 期望
- 方差

2.1 概率论基础

✦ 概率：

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。

第一公理（非负性）：$P(A) \geq 0$

第二公理（规范性）：$P(\Omega) = 1$

第三公理（可加性）：如果对任意的i和$j\,(i \neq j)$，事件A_i和A_j不相交$(\mathrm{A}_i \cap \mathrm{A}_j = \emptyset)$，则有：

$$P\left(\bigcup_{i=1}^{\infty} \mathrm{A}_i\right) = \sum_{i=0}^{\infty} P(\mathrm{A}_i)$$

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

如果一个实验的样本空间是$\{s_1, s_2, \cdots, s_n\}$，在相同情况下重复实验$N$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：

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

由于 $\sum_{i=1}^{n} n_k(s_k) = N$，因此，$\sum_{i=1}^{n} q_N(s_k) = 1$

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，通常用相对频率作为概率的估计值。

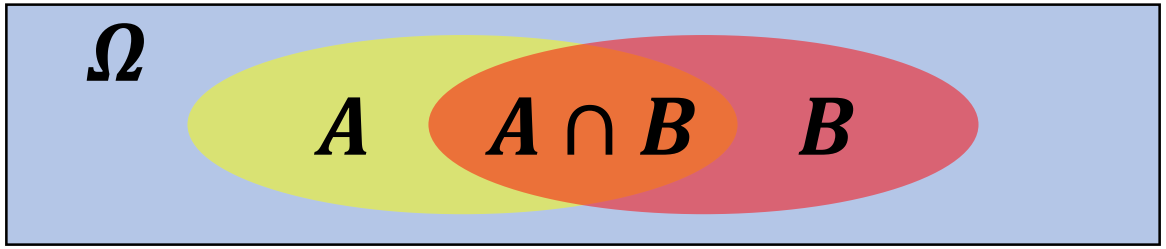
✦ 条件概率(conditional probability)

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：

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

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的概率。

一般地， $P(A|B)\neq P(A)$



$$P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$$

✦ 全概率公式(full probability)

假设B是样本空间Ω的一个划分，即事件B_i和B_j不相交($B_i \cap B_j = \emptyset \ (i \neq j; i,j = 1,2,\cdots,n)$，$\bigcup_{i=1}^{n} B_i = \Omega$)。如果$A \subseteq \bigcup_{i}^{n} B_i$，那么$A = \sum_{i}^{n} B_i$，

于是：

$$P(A) = P\left(\bigcup_{i}^{n} AB_i\right) = \sum_{i}^{n} \mathrm{P}(AB_i) = \sum_{i}^{n} P(B_i)P(A|B_i)$$

✦ 贝叶斯法则(Bayesian theorem)

假设A为样本空间Ω的事件，$B_1, B_2, \cdots, B_n$为Ω的一个划分，如果

$\mathrm{A} \subseteq \bigcup_{i}^{n} B_i$，$P(A) > 0$，并且$i \neq j, B_i \cap B_j = \emptyset, P(B_i) > 0 (i = 1,2, \cdots, n)$，则：

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)} = \frac{P(B_j)P(A|B_j)}{\sum_{i}^{n} P(B_i)P(A|B_i)}$$

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

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表示，$i=1,2,\cdots,c$；对应于各类别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$\vec{x}=[x_1,x_2,\cdots,x_d]^T$是$d$维特征空间上的某一点，且条件概率密度函数$P(\vec{x}|w_i)$是已知的。那么，利用贝叶斯公式我们可以得到后验概率

$$P(w_i|\vec{x})=\frac{P(w_i)P(\vec{x}|w_i)}{\sum_{j=1}^{c}P(w_j)P(\vec{x}|w_j)}$$

2.1 概率论基础

✦ 贝叶斯决策理论(Bayesian decision theory)

① 如果$P(w_i|\vec{x}) = \max\limits_{j=1,2,\cdots,c} P\left(w_j\middle|\vec{x}\right)$，那么$\vec{x} \in w_i$

② 或者：如果$P(w_i|\vec{x})P(w_i) = \max\limits_{j=1,2,\cdots,c} P\left(w_j\middle|\vec{x}\right)P(w_j)$，那么$\vec{x} \in w_i$

③ 或者($c = 2$时)：如果$l(\vec{x}) = \dfrac{p\left(\vec{x}\middle|w_1\right)}{p\left(\vec{x}\middle|w_2\right)} > \dfrac{P(w_2)}{P(w_1)}$，那么$\vec{x} \in w_1$，否则

$\vec{x} \in w_2$

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

✦ **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，那么，$\overset{\frown}{s} = \underset{s}{\operatorname{argmax}} P(S|A)$

✦ **解**：

语言模型

声学模型

$$\overset{\frown}{s} = \underset{s}{\operatorname{argmax}} \frac{P(S)P(A|S)}{P(A)}$$

- **例2-2**：假设某一种特殊的句法结构很少出现，平均大约每100000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**：假设G表示事件“句子中存在该特殊的句法结构”，T表示“程序判断句子含有该特殊句法结构”，则有

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

$$P(T|G)=0.95,\ P(T|\bar{G})=0.005$$

于是，可得

$$P(G|T)=\frac{P(G)P(T|G)}{P(T)}=\frac{P(G)P(T|G)}{P(T\cap G)+P(T\cap\bar{G})}$$

$$=\frac{P(G)P(T|G)}{P(G)P(T|G)+P(\bar{G})P(T|\bar{G})}\approx 0.002$$

✦ 二项式分布(binomial distribution)

假设某一事件A在一次试验中发生的概率为p，先把试验独立地重复进行n次。如果用变量X来表示A在这n次试验中发生的次数，那么， X的取值可能为$0,1,\cdots,n$。为了确定其分布情况，考虑事件$\{X=i\}$，如果这个事件发生，必须在这n次记录中有i个A，$n-i$个$\bar{A}$。

由于在n次试验中每次A出现与否与其他各次试验的结果无关，因此，根据乘法定律可以得出：每个这样的结果序列$A\bar{A}AA\cdots\bar{A}$发生的概率为

$$p^i(1-p)^{n-i}$$

✦ 二项式分布(binomial distribution)

又因为A可能出现在n个位置中的任何一处，因此，结果序列有$\binom{n}{i}$或C_i^n种可能，由此可得

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, i = 0,1,\cdots,n$$

X所遵从的这种概率分布成为二项式分布，并记为$B(n,p)$。如果随机变量X服从某种特定的分布F时，常用$X \sim F$表示。如果X服从二项式分布，可记为$X \sim B(n,p)$

✦ 期望 (expectation)

期望值是指随机变量所取值的概率平均。假设X为一随机变量，其概率分布为$P(X=x_k)=p_k, k=1,2,\cdots$，若极数$\sum_{k=1}^{\infty} x_k p_k$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。如果X为一随机变量，其方差$var(X)$为

$$var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E^2(X)$$

平凡根$\sqrt{var(x)}$成为X的标准差

2.2 信息论基础

2.2 信息论基础

- **基本概念**
  - 熵
  - 联合熵
  - 条件熵
  - 熵率
  - 相对熵
  - 交叉熵
  - 困惑度
  - 互信息
  - 噪声信道模型

✦ 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。

✦ 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：

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

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。

✦ **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

✦ **解：**

(1) 等概率出现情况：

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) = 27 \times \left\{ -\frac{1}{27} \log_2 \frac{1}{27} \right\}$$

$$= \log_2 27 = 4.75\ (bits)$$

(2) 实际情况

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

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：

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

Ref. 冯志伟，1989

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(bits)。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、台北香港、三地汉语词的熵：

北京5年		台北5年		香港5年		京、台、港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。

专用名词主要指：人名、地名、组织机构名，又称命名实体(named entity)。

Ref. Tsou, 2003

✦ 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：

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

➢ 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。

✦ 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：

$$H(Y|X)=\sum_{x\in X}p(x)H(Y|X=x)$$

$$=\boxed{\sum_{x\in X}p(x)[-\sum_{y\in Y}p(y|x)\log_2 p(y|x)]}$$

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

2.2 信息论基础

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

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

$$=-\sum_{x\in X}\sum_{y\in Y}p(x,y)[\log p(x)+\log p(y|x)]$$

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

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

$$=\boxed{H(X)+H(Y|X)}$$

连锁规则

✦ **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?

✦ 解：

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

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) = \frac{7}{4}$$

✦ 解:

Y \ 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

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

✦ 解：

Y \ 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	

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

…

✦ **解:**

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

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

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

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

类似地，$H(Y|X)=\frac{13}{8}\ (\text{bits})$

可见，$H(X|Y)\neq H(Y|X)$

✦ 解:

$$H(X,Y) = -\sum_{x=1=1}^{4}\sum^{4} p(x,y)\log_2 p(x,y)$$

$$= -\left(\frac{1}{8}\times\log_2\frac{1}{8}+\cdots+0\right)$$

$$= \frac{27}{8}$$

✦ **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：

字母	p	t	k	a	i	u
概率	1/8	1/4	1/8	1/4	1/8	1/8

那么，每个字符的熵为：

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

$$= \frac{5}{2}\ (\text{bits})$$

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：

字母	p	t	k	a	i	u
编码	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$P(\bullet,C)$如下表所示：

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

字母	p	t	k	a	i	u
概率	1/16	3/8	1/16	1/4	1/8	1/8

✦ 求联合熵$H(C,V)$为多少?

- **解**：采用连锁规则方法计算

$$\boldsymbol{H}(\boldsymbol{C}) = -\sum_{c=p,t,k} p(c)\log p(c) = -2\times\frac{1}{8}\times\log\frac{1}{8} - \frac{3}{4}\times\log\frac{3}{4}$$

$$= \frac{9}{4} - \frac{3}{4}\log 3 \approx 1.061\ (\text{bits})$$

$$\boldsymbol{H}(\boldsymbol{V}|\boldsymbol{C}) = \sum_{c=p,t,k} p(C=c)H(V|C=c)$$

$$= \frac{1}{8}H\left(\frac{1}{2},\frac{1}{2},0\right) + \frac{3}{4}H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) + \frac{1}{8}H\left(\frac{1}{2},0,\frac{1}{2}\right) = \frac{11}{8} = 1.375\ (\text{bits})$$

$$\boldsymbol{H}(\boldsymbol{C},\boldsymbol{V}) = H(C) + H(V|C)$$

✦ 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：

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

这个数值我们也成为熵率。其中，变量X_{1n}表示随机变量序列$(X_1, \cdots, X_n)$，$x_{1n} = (x_1, \cdots, x_n)$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，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）
- $x_{1n} = (\text{为}, \text{传}, \cdots, \text{动}, \text{。})$
- $H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{66} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$

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

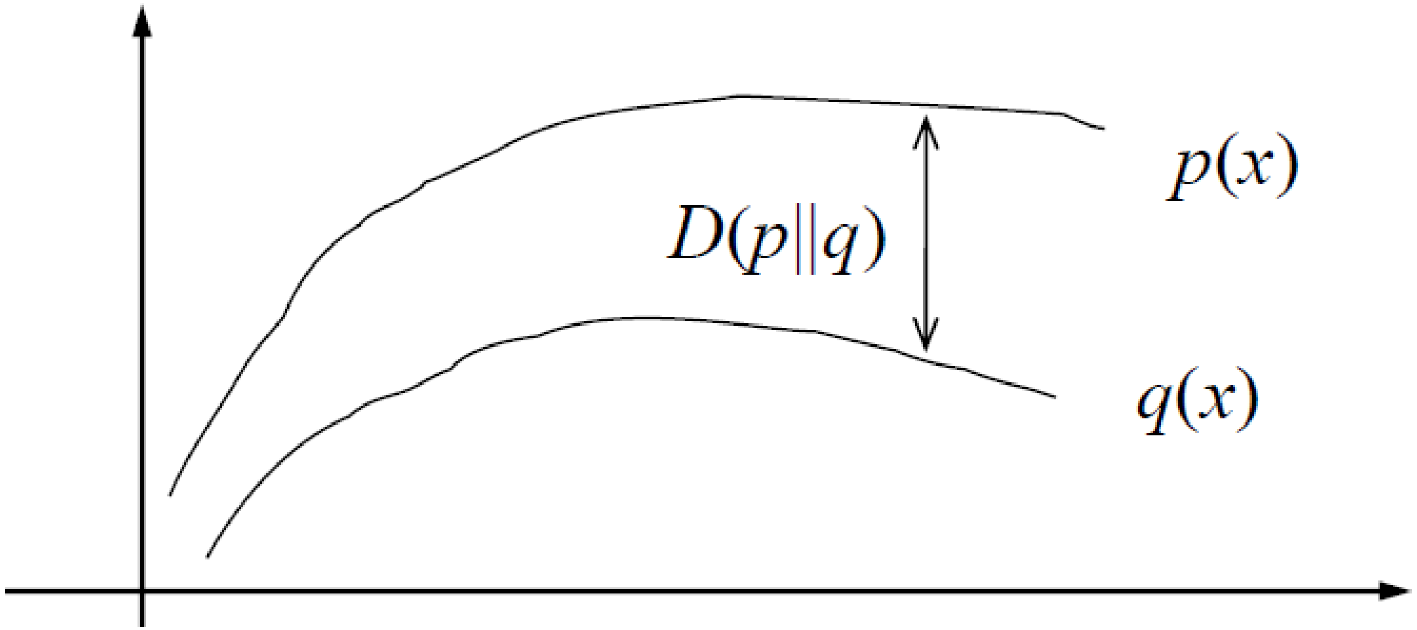
两个概率分布$p(x)$和$q(x)$的相对熵定义为：

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

该定义中约定：

$$0 \log\left(\frac{0}{q}\right) = 0\,, \qquad p \log \frac{p}{0} = \infty$$

- 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。



✦ 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：

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

✦ 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：

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

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

$p(x_1^n)$为语言L中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**：

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

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

➢ 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$p(x)$。

✦ 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：

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

- 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。

✦ 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：

$$I(X;Y)=H(X)-H(X|Y)$$

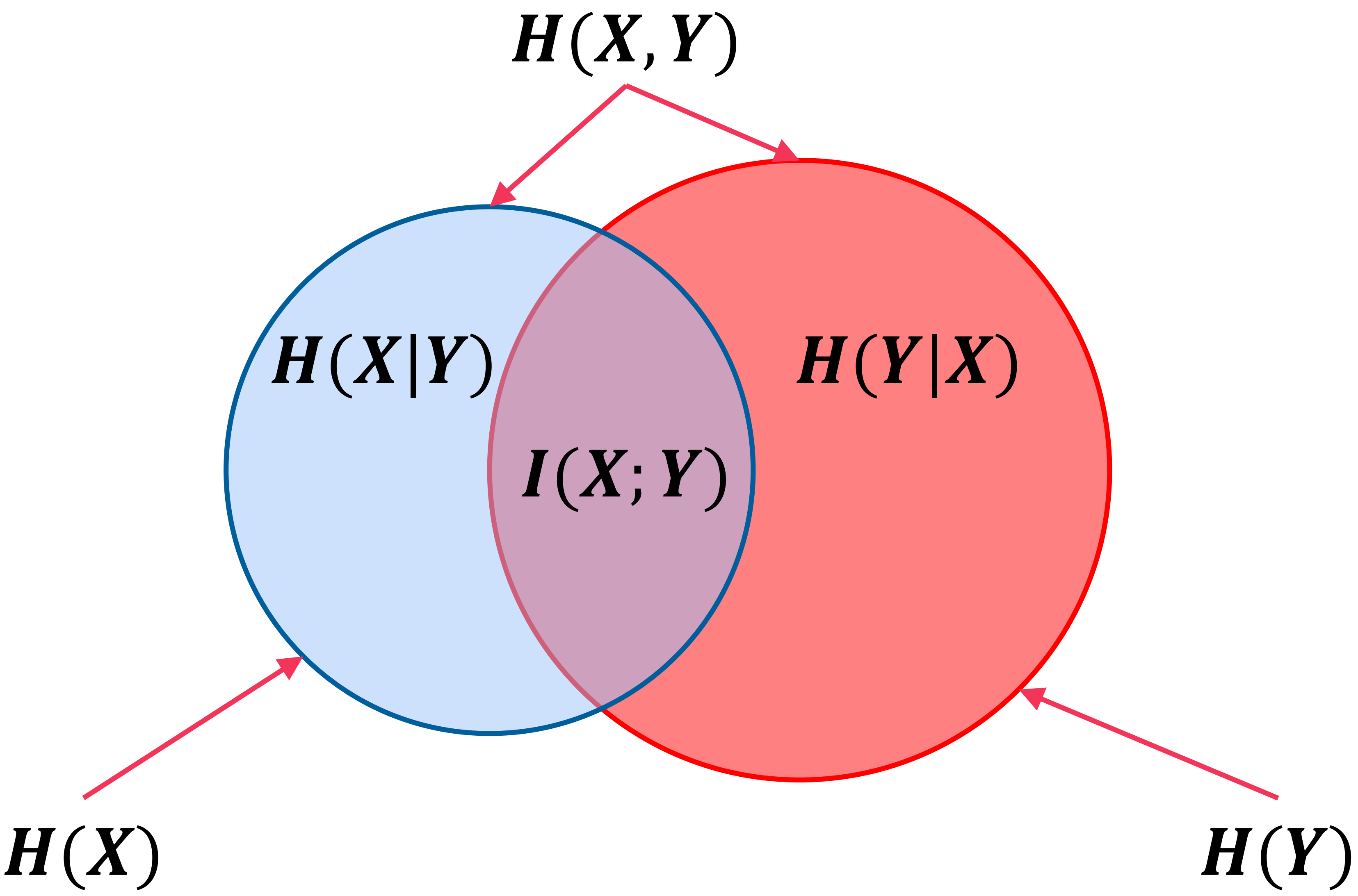
$$=-\sum_{x\in X}p(x)\log_2 p(x)+\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2 p(x|y)$$

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

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

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



2.2 信息论基础

- 由于$H(X|X)=0$，所以，

$$H(X)=H(X)-H(X|X)=I(X;X)$$

- 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。

✦ 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强度：

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

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：

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

- 其中，c_i，c_{i+1}是一个有序字对，表示两个连续的汉字，且c_i，c_{i+1}不等于$c_i c_{i+1}$。$N(c_i c_{i+1})$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。
- 注：此处"|"不表示条件概率。
- 例如："为人"出现5次，"为人民"出现20次，那么，

$$Couple(为, 人) = 0.2$$

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

- **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14945次，"务"字出现了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。

✦ 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。

✦ **过程示意图：**

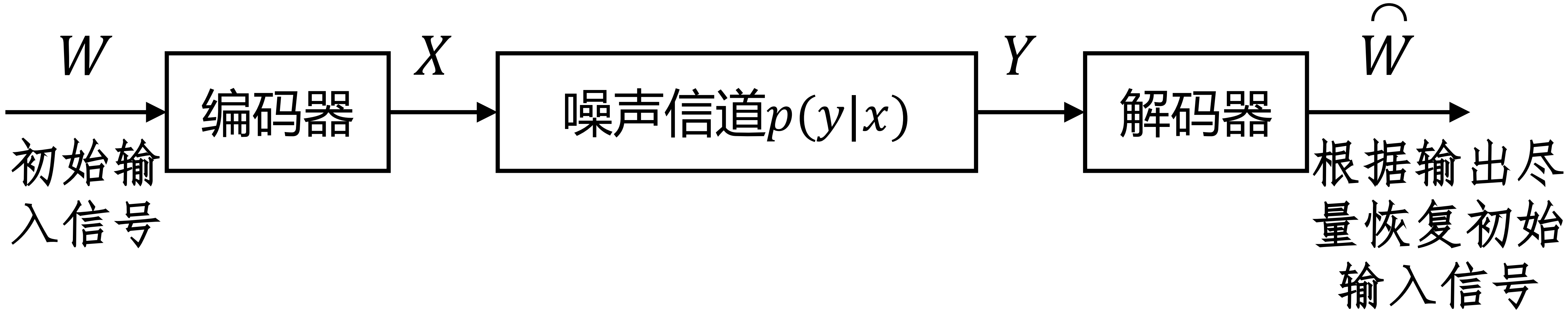


图. 噪声信道模型示意图

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

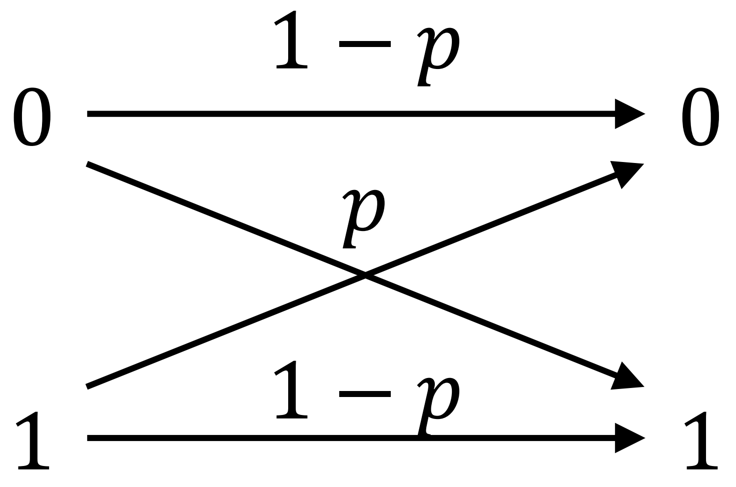


图. 对称信道

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：

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

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

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。

- **例如**：将法语翻译成英语

英语句子e → 噪声信道 → 法语句子f

根据贝叶斯公

$$p(e|f) = \frac{\underbrace{p(e)}_{\text{语言模型}}\underbrace{p(f|e)}_{\text{翻译模型}}}{p(f)}$$

2.3 应用实例

2.3 应用实例

- **例2-6：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 应用实例

- 例如：

(1) 他**打**鼓很在行。

(2)他会**打**家具。

(3)他把碗**打**碎了。

(4)他在学校**打**架了。

(5)他很会与人**打**交道。

(6)他用土**打**了一堵墙。

(7)用面**打**浆糊贴对联。

(8)他**打**铺盖卷儿走人了。

(9) 她会用毛线**打**毛衣。

(10) 他用尺子**打**个格。

(11) 他**打**开了箱子盖。

(12) 她**打**着伞走了。

(13) 他**打**来了电话。

(14) 他**打**了两瓶水。

(15) 他想**打**车票回家。

(16) 他以**打**鱼为生。

2.3 应用实例

- 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

✦ 基于上下文分类的消歧方法

——基于贝叶斯分类器 (Ref. 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，可通过计算$\underset{s_i}{\operatorname{argmax}}\, p(s_i|C)$确定$w$ 的词义。

2.3 应用实例

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：

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

因此，

$$\hat{s_i} = \underset{s_i}{\operatorname{argmax}}[p(s_i) \prod_{v_k \in C} p(v_k|s_i)]$$

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

✦ **描述基于上下文分类的消歧算法：**

① 对于多义词w 每个语义s_i 执行如下循环：对于词典中所有的词v_k 计算：

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

② 对于多义词w 的每个语义$\boldsymbol{s_i}$ 计算：

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

利用已标注的大规模数据训练的过程

2.3 应用实例

③ 对于多义词w 每个语义$\boldsymbol{s}_i$ 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：

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

$\boxed{\text{标注过程或称测试过程}}$

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：

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

(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。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

- 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$\boldsymbol{a}$，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$\boldsymbol{b}$。利用条件熵$\boldsymbol{H(a|b)}$ 最大时概率$\boldsymbol{p(a|b)}$ 推断多义词使用某一义项的可能性。

2.3 应用实例

- 用A表示某一多义词所有义项的集合，B表示所有上下文的集合。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。

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

- 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**$\tilde{p}(a,b)$的计算式为：$\tilde{p}(a,b)=\frac{Count(a,b)}{\sum_{A,B} Count(a,b)}$

2.3 应用实例

- f_j 在训练样本中关于经验概率分布的数学期望为：

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

- 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：

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

- 由于$p(a,b)=p(b)p(a|b)$，且所建立模型应符合已知样本中的概率分布，即$p(b)=\tilde{p}(b)$，则有：

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

2.3 应用实例

- 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：

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

- 因此，该式称为该问题建模的**约束方程**，简称**约束**。

- 假设存在k 个特征f_j ($j = 1,2,\cdots,k$)，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：

$$P = \left\{p \middle| E_p\left(f_j\right) = E_{\tilde{p}}\left(f_j\right), j \in \{1,2,\cdots,k\}\right\}$$

2.3 应用实例

- 根据条件熵的定义：

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

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

$$\hat{p} = \underset{p\in P}{\operatorname{argmax}}\, H(p)$$

- 就是使熵最大且符合约束的概率分布

谢 谢

2.3 应用实例

- 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(未知分布) 的概率$p(a|b)$，该概率使熵$H(p(A|B))$最大。

经推导，有:

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

其中，$Z(b) = \sum_{a} \exp\left(\sum_{j=1}^{l} \lambda_j \cdot f_i(a,b)\right)$

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