第2章 数学基础

宗成庆
中国科学院自动化研究所
cqzong@nlpr.ia.ac.cn

本章内容

1. 概率论基础
2. 信息论基础
3. 应用举例
4. 习题
5. 附录

1. 概率论基础

- 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)
- 随机过程 (stochastic process)

“语言是稳态的可遍历性随机过程”

1. 概率论基础

- 随机过程 (stochastic process)

假设$\{\xi_t, t\in T\}$是一族随机变量，T是一个实数集合，如果对于任意实数 $t\in T$，ξ_t 是一个随机变量，则称$\{\xi_t, t\in T\}$为随机过程。

任意一个样本点t，ξ_t 都对应一个实数，而ξ_t 是随着试验结果不同而变化的一个变量，则称 ξ_t 为随机变量。

(有时将 ξ_t 写为：$\xi(t)$，或者用大写X 等表示。)

随机过程的**平稳性**(stationary)：在数学中平稳过程又称**严格平稳过程**或**强平稳过程**，是一种特殊的随机过程，在其中任取一段时间（$t = t_1 - t_2$）或空间里的联合概率分布，与将这段期间任意平移后的新时间段里的联合概率分布相等，即：

$$f_X(x_1,\dots,x_n,t_1+\Delta t,\dots,t_n+\Delta t) = f_X(x_1,\dots,x_n,t_1,\dots,t_n)$$

1. 概率论基础

随机过程的<u>**遍历性**</u>(ergodic)：对于一个平稳的随机变量*X*，如果它的所有样本函数在某一固定时刻的统计特性和单一样本函数在长时间内的统计特性一致，我们则称*X*为<u>**各态遍历**</u>，即随机变量单一样本函数随时间变化的过程可以包括该变量所有样本函数的取值经历。**隐藏含义是：如果一个随机变量是遍历性的，那么该随机变量的时间统计特性等于其空间统计信息。**

1. 概率论基础

解释： **"语言是稳态的可遍历性随机过程"**。

❶**稳态的**：从今天的《人民日报》和明天的《人民日报》中分别采样，计算得到的汉语单词的联合概率和其他数学特征是相等的，即在较短时间内**"语言是一个稳态的随机过程"**。当然从较长的时间跨度看，语言的分布和习惯也在随时间的变化而变化，并不满足稳态性要求。

❷**可遍历**：假设我们想计算某个语言的分布，可以采取两种方式：(1)让某个掌握该语言的人(单一样本函数)在很长时间一直书写这种语言，产生该语言的大量样本(时间统计特性)；(2)让很多个人(多个样本函数)在较短的时间里一起书写这个语言(空间统计特性)。上述两种方式所获取的概率分布是一样的(近似)。

本章内容

1. 概率论基础
2. 信息论基础
3. 应用举例
4. 习题
5. 附录

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. 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：

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

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)。

熵表示随机变量X 每个具体取值(如信源每发射一个信号)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. 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. 信息论基础

解：(1) 等概率出现情况：

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

$$= 27 \times \{-\frac{1}{27}\log_2 \frac{1}{27}\} = \log_2 27 = 4.75 \quad \text{(bits/letter)}$$

(2) 按表中概率计算：

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

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. 信息论基础

法语、意大利语、西班牙语、英语、俄语字母及汉字的熵：

语言	熵 (bits)
法语	3.98
意大利语	4.00
西班牙语	4.01
英语	4.03 (英语单词的熵约为10。)
俄语	4.35
汉字	9.71 (汉语词的熵约为11.46。)

[冯志伟，1989]

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 分别是从香港城市大学建立的语料库(LIVAC)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. 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：

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

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. 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：

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

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

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：

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log[\,p(x)p(y\mid x)]\\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)[\log p(x) + \log p(y\mid x)]\\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x)\\
&= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x)\\
&= H(X) + H(Y\mid X) \qquad (4)
\end{aligned}$$

(连锁规则)

2. 信息论基础

例2-2：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. 信息论基础

例2-2：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
$p(X)$	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. 信息论基础

类似地，可以计算$H(Y)$。

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

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

2. 信息论基础

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	

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

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

……

2. 信息论基础

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

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

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

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

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

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

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. 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：

$$H(P) = -\sum_{i \in \{p,t,k,a,i,u\}} P(i) \log P(i)$$

$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}] = 2\frac{1}{2} \text{(bits)}$$

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. 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 \ 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. 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. 信息论基础

求联合熵有几种方法，我们可采用连锁规则方法：

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

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

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

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

因此，

$$H(C,V) = H(C) + H(V\,|\,C) = \frac{9}{4} - \frac{3}{4}\log 3 + \frac{11}{8} \approx 2.44\,(\mathrm{bits})$$

2. 信息论基础

◆ 相对熵(relative entropy, 或称 Kullback-Leibler divergence, K-L 距离，或K-L散度）

两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为：

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{5}$$

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

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。

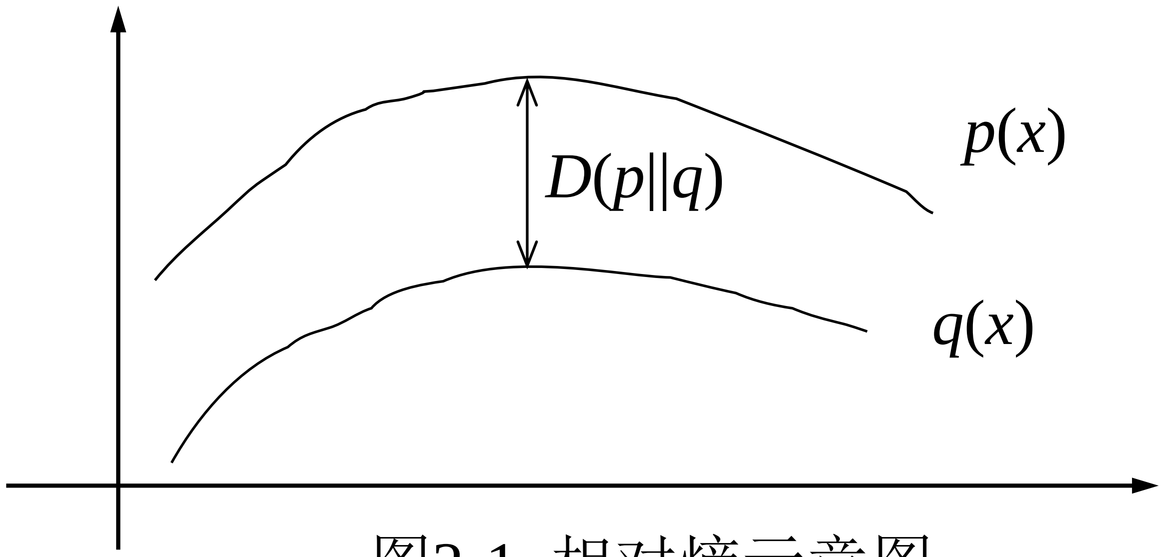


图2-1. 相对熵示意图

2. 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：

$$
\begin{aligned}
H(X,q) &= H(X) + D(p \| q) \\
&= \sum_{x \in X} p(x) \log p(x) + \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_{x \in X} p(x) \log q(x) \qquad (6)
\end{aligned}
$$

交叉熵用以衡量估计模型与真实概率分布之间的差异。

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:

$$H(L,q)=-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n}p(x_1^n)\log q(x_1^n) \qquad (7)$$

其中，$x_1^n = x_1 \ldots x_n$ 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. 信息论基础

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

$$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n) \tag{8}$$

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. 信息论基础

问题：相对熵和交叉熵的作用有什么区别呢？

根据定义，相对熵：$D(p\,\|\,q)=\sum_{x\in X}p(x)\log\frac{p(x)}{q(x)}$

交叉熵：$H(p,q)=-\sum_{x\in X}p(x)\log q(x)=\sum_{x\in X}p(x)\log\frac{1}{q(x)}$

$H(p,q)=H(p)+D(p\|q)$，即：**交叉熵 = 熵 + 相对熵**。

2. 信息论基础

说明：在机器学习中经常用$p(x)$表示**真实**数据的概率分布，由于真实数据的概率分布往往无法获得，所以一般通过大量的**训练数据**来近似。假设我们通过某个模型得到了训练数据的概率分布$q(x)$，**由于真实数据的概率分布$p(x)$往往是不变的，因此最小化交叉熵$H(p, q)$等效于最小化相对熵$D(p||q)$。**习惯上机器学习算法中通常采用**交叉熵**计算损失函数。

例如，在某机器学习任务中定义损失函数为**交叉熵**: Loss=$H(p, q)$，假设我们训练到得到一个非常好的模型，即$p(x)\approx q(x)$，此时Loss不会降低为0，而是一个很小的值，如Loss=2，它表示真实数据自身的熵为$H(p)$=2。如果选择**相对熵**作为损失函数，即Loss=$D(p||q)$，同样假设我们训练得到一个非常好的模型，即$p(x)\approx q(x)$，此时，Loss= 0，意味着两个概率分布几乎一样。

实际上，上述两种方法所得到的Loss仅仅是数值上的区别，训练得到的模型是完全一样的，即两个概念的作用一样。

2. 信息论基础

◆困惑度(perplexity)

在设计<u>语言模型</u>时，我们通常用<u>困惑度来代替交叉熵</u>衡量语言模型的好坏。给定语言L的样本$l_1^n = l_1 \ldots l_n$，L 的困惑度PP_q定义为：

$$PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n}\log q(l_1^n)} = [q(l_1^n)]^{-\frac{1}{n}} \quad (9)$$

> 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分布。

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：

$$I(X; Y) = H(X) - H(X \mid Y) \tag{10}$$

根据$H(X)$ 和 $H(X|Y)$ 的定义：

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

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

2. 信息论基础

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

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

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

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

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x) p(y)} \qquad (11)$$

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. 信息论基础

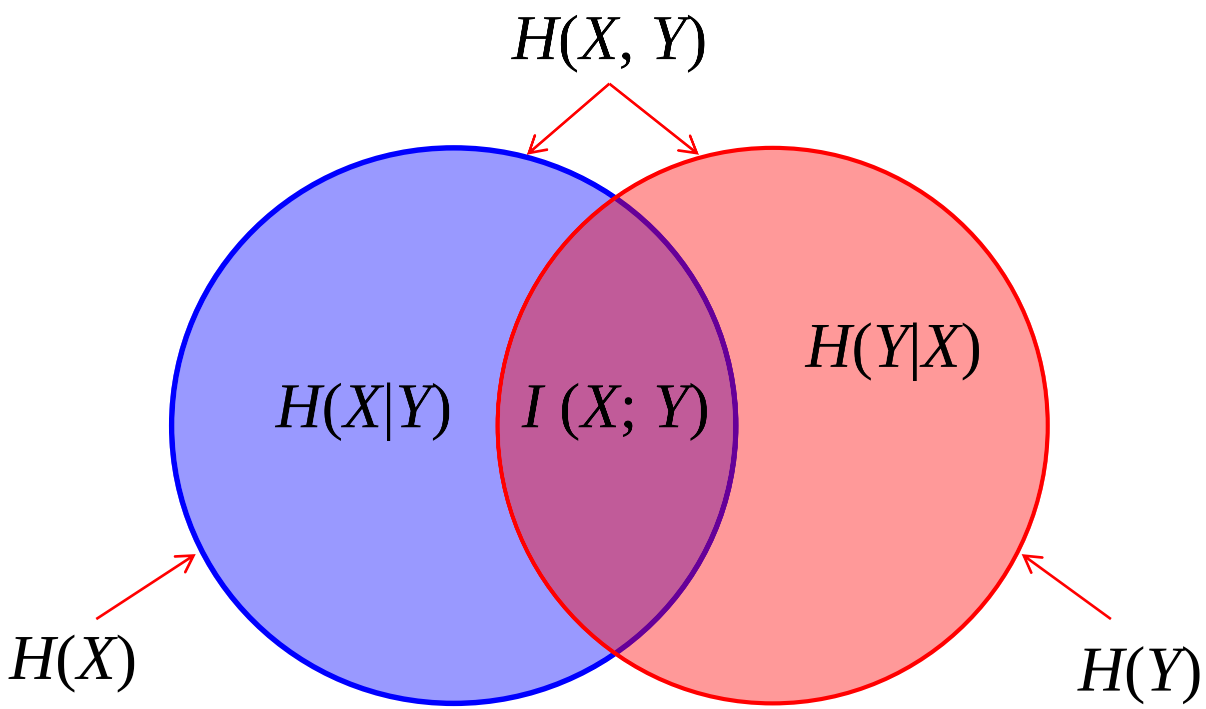


图 2-2. 互信息、条件熵与联合熵

2. 信息论基础

由于 $H(X|X) = 0$, 所以，

$$H(X) = H(X) - H(X|X) = I(X; X) \tag{12}$$

所以，熵又称自信息(self-information)。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. 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强度：

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

理论上，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。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. 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. 信息论基础

在汉语分词研究的实践证明中，用互信息作为衡量两个汉字之间是否作为切分边界点的判别依据，效果并不好。

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。教和务之间的互信息只有－0.5119。如果用互信息来判断的话，这两个字应该被切开。但实际上，教和务这两个字在文本集中出现的16次全部都是教务、教务长、教务处等词汇。也就是说，这两个字一旦连续同现，一定成词。因此，用两个邻近的字在训练样本中构成词的比率作为切分依据，比互信息效果更好。

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. 信息论基础

过程示意图：

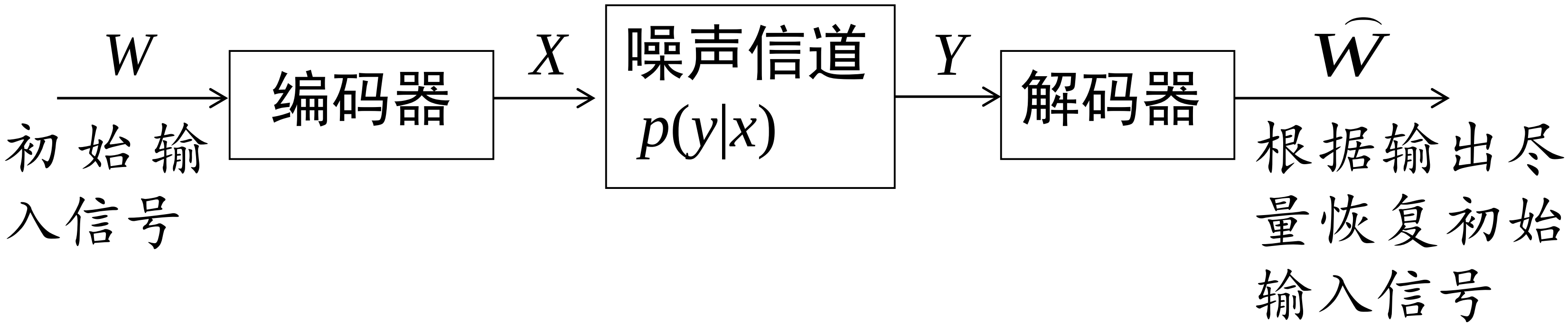


图 2-3. 噪声信道模型示意图

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，如机器翻译。

本章内容

3. 应用举例

例2-4: 词汇歧义消解

●问题提出

任何一种语言中，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。

3. 应用举例

以“打”字为例，用作动词时有24个意思：

(1) 他会打鼓。
(2) 他把碗打破了。
(3) 他在学校打架了。
(4) 他想打官司。
(5) 他用土打了一堵墙。
(6) 他会用木头打家具。
(7) 她用面打浆糊贴对联。
(8) 他打铺盖卷儿走人了。
(9) 她会用毛线打毛衣。
(10)他用尺子在纸上打了格子。
(11)他打开了井盖子。
(12)这种人打着灯笼也难找。
(13) 给他打个电话吧。
(14) 他把款打过去了。
(15) 你别打杈。
(16) 你打两瓶水去。
(17) 他想打车票回家。
(18) 他以打鱼为生。
(19) 他放学后去打猪草了。
(20) 你打个草稿再写。
(21) 八路军会打游击。
(22) 我们一起打扑克吧。
(23) 他给她打了个手势。
(24) 你别打官腔/马虎眼。

3. 应用举例

●解决思路

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。**词义消歧变成一个上下文分类任务。**

他/P 很/D 会/V 与/C 人/N 打/V 交道/N 。/PU

-2 -1 ↑ +1 +2

0

基本的上下文信息：词、词性、位置。

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 $\arg\max_{s_i} p(s_i \mid C)$ 来确定w 的词义。

3. 应用举例

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

考虑分母的不变性，并运用如下独立性假设：

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

出现在上下文中的词

因此，

$$\hat{s}_i = \arg\max_{s_i} \left[p(s_i) \prod_{v_k \in C} p(v_k \mid s_i) \right] \qquad (14)$$

概率 $p(v_k \mid s_i)$ 和 $p(s_i)$ 都可用最大似然估计求得。

3. 应用举例

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)} \tag{15}$$

其中，$N(s_i)$是在训练样本中词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 出现在上下文中的次数。

$$p(s_i) = \frac{N(s_i)}{N(w)} \tag{16}$$

$N(w)$ 为多义词 w 在训练数据中出现的总次数。

3. 应用举例

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

对于“打”字而言，用作实词的24个义项分别标记为：$s_1 \sim s_{24}$。假设 s_1 的词义为“敲击(beat)”。那么$N(s_1)$表示“打”字意为“敲击(beat)”时在所有标注样本中出现的次数；$N(v_k, s_1)$表示某个词 v_k 出现在“打”的上下文中时的次数。例如，句子：

他 对 打 鼓 很 在行 。 (取上下文：±2)

-2 -1 ↑ +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，

3. 应用举例

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

假若“打”在所有样本中总共出现了800次，那么，

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

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利用训练样本(已标注)计算：

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

②对于w 的每个语义 s_i 计算：

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

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：

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

当概率值最大时的参量值。

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：

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

3. 应用举例

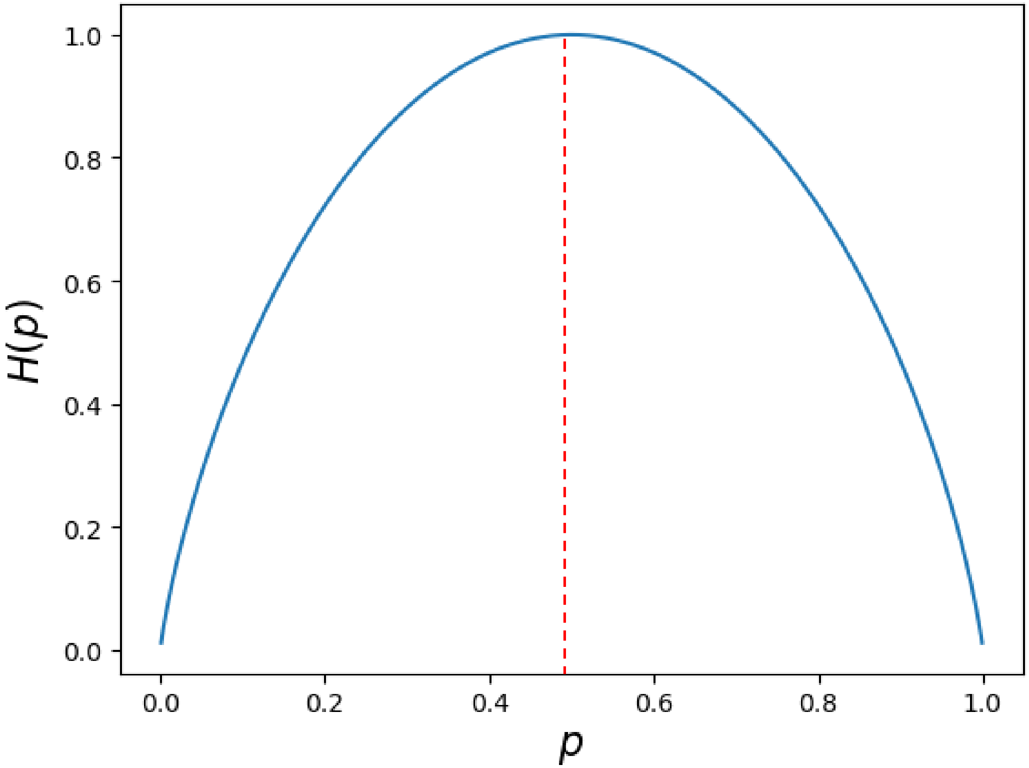
(2)基于最大熵的消歧方法

➢ **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，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。

3. 应用举例

假设 x 和 y 两个相互关联的事件，$p(x)+p(y)=1$，将$p(x)$ 简记为p，p的取值与其熵 $H(p)$ 的关系如下图所示：



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

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：

特征函数

$$p^*(a \mid b) = \frac{1}{Z(b)} \exp\left(\sum_{j=1}^{l} \lambda_j \cdot f_j(a,b)\right) \tag{17}$$

特征权重

其中，$$Z(b) = \sum_a \exp\left(\sum_{j=1}^{l} \lambda_j \cdot f_j(a,b)\right) \tag{18}$$

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

3. 应用举例

➢ 确定特征函数

对于词义消歧而言, 设 A 为某一个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”字的动词义项集合：$A=\{s_1, s_2, s_3, \ldots\ldots, s_{24}\}$,

$B=\{$“打”字出现的上下文$\}$。

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

3. 应用举例

上下文有两种表示方法：

①<u>位置无关</u>：目标词周围的词形、词性或其组合构成的集合，如取±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<u>位置有关</u>：词形(±2)：<与$_{-2}$，人$_{-1}$，交道$_{+1}$，。$_{+2}$>

≠<与$_{-2}$，。$_{+2}$，交道$_{+1}$，人$_{-1}$>

模板表示

3. 应用举例

他/P 很/D 会/V 与/C 人/N 打/V s_4 交道/N 。/PU

↑

假设以词形和词性等表示条件，可以构造特征函数：

$$f_1(a,b)=\begin{cases}1 & \text{If } a = s_4 \text{ and } b = \text{<(与, 人), (交道, 。)>} \\ 0 & \text{Otherwise}\end{cases}$$

$$f_2(a,b)=\begin{cases}1 & \text{If } a = s_4 \text{ and } b = \text{<(C, N), (N, PU)>} \\ 0 & \text{Otherwise}\end{cases}$$

… …

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、位置等又可以组合出很多种可能，可以构造出若干特征函数，因此需要对特征进行筛选。

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。**

3. 应用举例

回顾前面的公式(17)、(18)：

$$p^*(a\,|\,b)=\frac{1}{Z(b)}\exp\Big(\sum_{j=1}^{l}\lambda_j\cdot f_j(a,b)\Big) \qquad (17)$$

$$Z(b)=\sum_{a}\exp\Big(\sum_{j=1}^{l}\lambda_j\cdot f_j(a,b)\Big) \qquad (18)$$

此处 $l = k+1$。

3. 应用举例

- 获取λ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： $0 \sim C$。

3. 应用举例

- GIS算法描述

  (a) 初始化：$\lambda[1..\, l\,]=0$;

  (b) 计算每个特征函数 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(17)、(18)计算概率 p^*;

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

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

  ②若满足终止条件，则结束迭代；

  否则，修正 λ，继续下轮迭代。

  (d) 算法结束，确定 λ，算出每个 p^*。

λ 修正方法：

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

3. 应用举例

迭代终止条件：

a. 限定迭代次数；

b. 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。

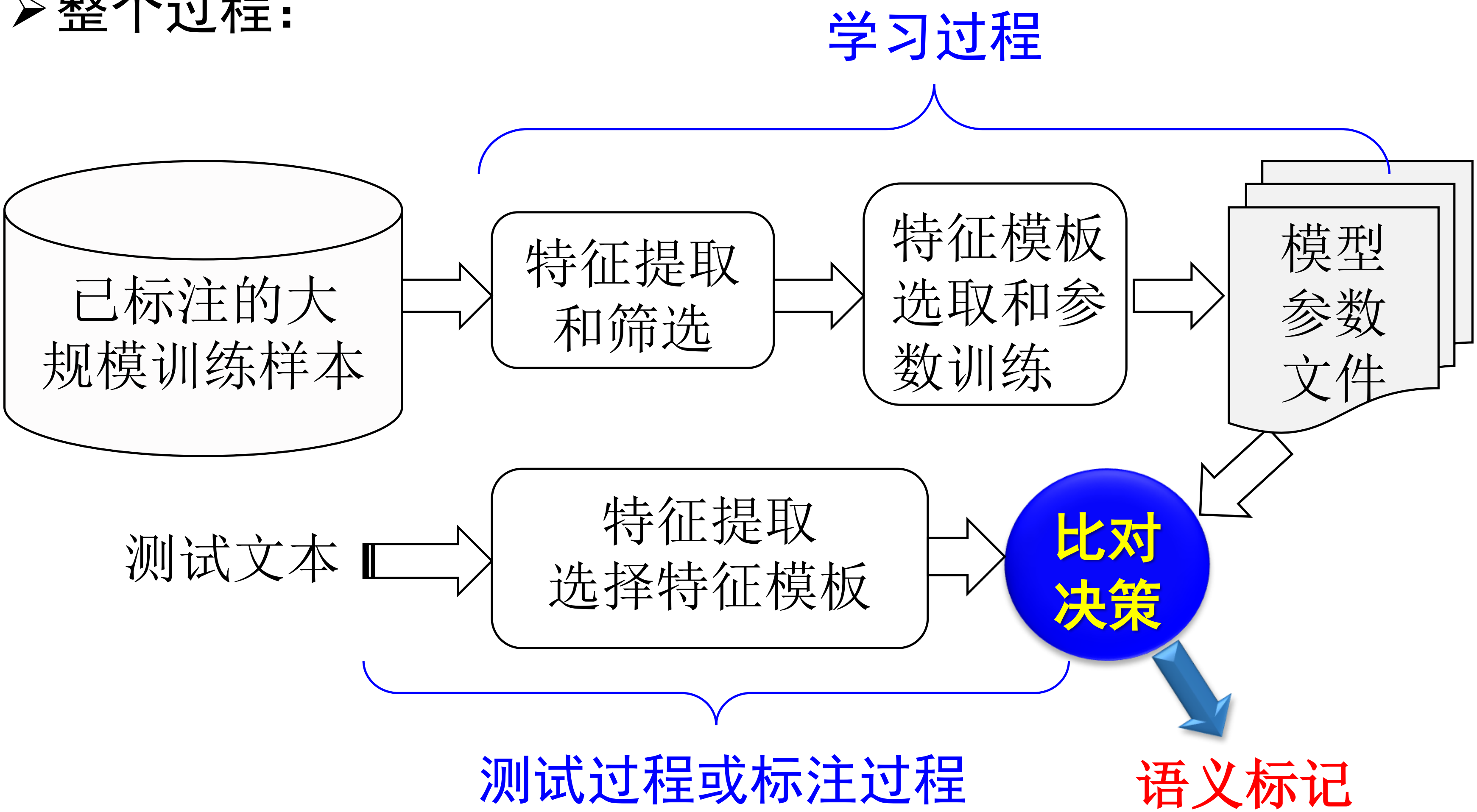
关于GIS算法的改进，请参阅如下文献：

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

3. 应用举例

➢ 整个过程：


学习过程
已标注的大规模训练样本
特征提取和筛选
特征模板选取和参数训练
模型参数文件
测试文本
特征提取选择特征模板
比对决策
测试过程或标注过程
语义标记

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%。

3. 应用举例

➢ 关于最大熵方法在NLP中的应用及GIS，请参阅：

[1] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. A Maximum Entry Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, 1996. Pages 39-71 **[ToT Award, ACL-IJCNLP'2021]**

[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] A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity Resolution, PhD Dissertation, UPenn., 1998

[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

➢ 最大熵开源工具：

✧ OpenNLP：http://incubator.apache.org/opennlp/

✧ Malouf： http://tadm.sourceforge.net/

本章内容

4. 习题

2-1. 收集尽量多的英语文本，编写程序分别计算这些文本中英语字符、单词和标点符号的熵。

2-2. 收集尽量多的汉语文本，编写程序分别计算这些文本中汉字的熵和任意两个汉字之间的点式互信息。

2-3. 根据上述两个题目收集的文本，分别计算汉字与英语字符之间的KL 距离，以及英语字符与汉字之间的KL距离，对比两个KL距离的大小。

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：$H(p, q) = H(p) + D(p \| q)$。请证明：

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

2-5. 根据本章内容设计并实现一种算法，计算任意两个句子之间的相似性。

本章小结

- ◆概率论基础
  - ➢基本概念和方法（复习）
- ◆信息论基础
  - ➢熵
  - ➢联合熵
  - ➢互信息
  - ➢相对熵
  - ➢交叉熵
  - ➢困惑度
  - ➢噪声信道模型
- ◆应用举例
  - ➢词义消歧

本章内容

1. 概率论基础
2. 信息论基础
3. 应用举例
4. 习题
5. 附录

5. 附录

1. 证明前面P31公式(8): $H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n)$

(又见《统计自然语言处理》P28，公式(2-39))

2. 证明：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

5. 附录

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

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

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

该定理的证明：

假设 $(x_1, x_2, \ldots, x_n)$ 符合独立同分布。根据熵率的定义, 左边:

$$H_{\text{rate}}(X) = \lim_{n\to\infty} \frac{1}{n} H(x_1^n)$$

$$= \lim_{n\to\infty} \frac{1}{n} H(x_1, x_2, ..., x_n)$$

$$= \lim_{n\to\infty} \{-\frac{1}{n} \sum_{x_1 x_2 \ldots x_n} p(x_1, x_2, ..., x_n) \log p(x_1, x_2, ..., x_n)\}$$

5. 附录

$$= \lim_{n\to\infty}\{-\frac{1}{n}\sum_{x_1x_2\ldots x_n} p(x_1,x_2,\ldots,x_n)(\sum_{x_i}\log p(x_i))\}$$

$$= \lim_{n\to\infty}\{-\frac{1}{n}[\sum_{x_1x_2\ldots x_n} p(x_1,x_2,\ldots,x_n)\log p(x_1) + \sum_{x_1x_2\ldots x_n} p(x_1,x_2,\ldots,x_n)\log p(x_2) + \ldots$$

$$+ \sum_{x_1x_2\ldots x_n} p(x_1,x_2,\ldots,x_n)\log p(x_n)]\}$$

$$= \lim_{n\to\infty}\{-\frac{1}{n}[\sum_{x_1x_2\ldots x_n} p(x_1)p(x_2)\ldots p(x_n)\log p(x_1) + \sum_{x_1x_2\ldots x_n} p(x_1)p(x_2)\ldots p(x_n)\log p(x_2)$$

$$+ \ldots + \sum_{x_1x_2\ldots x_n} p(x_1)p(x_2)\ldots p(x_n)\log p(x_n)]\}$$

5. 附录

$$
\begin{aligned}
&= \lim_{n\to\infty}\{-\frac{1}{n}[\sum_{x_1} p(x_1)\log p(x_1)(\underline{\sum_{x_2x_3\ldots x_n} p(x_2)p(x_3)\ldots p(x_n)})\\
&\quad + \sum_{x_2} p(x_2)\log p(x_2)(\underline{\sum_{x_1,x_3x_4\ldots x_n} p(x_1)p(x_3)p(x_4)\ldots p(x_n)})\\
&\quad + \ldots + \sum_{x_n} p(x_n)\log p(x_n)(\underline{\sum_{x_1x_2\ldots x_{n-1}} p(x_1)p(x_2)\ldots p(x_{n-1})})]\}
\end{aligned}
$$

由于$(x_1, x_2,\ldots, x_n)$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。

$$
\begin{aligned}
&= \lim_{n\to\infty}\{-\frac{1}{n}[\sum_{x_1} p(x_1)\log p(x_1) + \sum_{x_2} p(x_2)\log p(x_2) + \ldots + \sum_{x_n} p(x_n)\log p(x_n)]\}\\
&= \lim_{n\to\infty}\{-\frac{1}{n}[\sum_{x_1} p(x_1)\log p(x_1) + \sum_{x_2} p(x_2)\log p(x_2) + \ldots + \sum_{x_n} p(x_n)\log p(x_n)]\}\\
&= \lim_{n\to\infty}\{\frac{1}{n}[\mathrm{H}(x_1) + \mathrm{H}(x_2) + \ldots + \mathrm{H}(x_n)]\}\\
&= \lim_{n\to\infty}\{\frac{1}{n}\times n\times \mathrm{H}(x_i)\}\\
&= H(x_i)
\end{aligned}
$$

5. 附录

而定理右边：

$$-\lim_{n\to\infty}\frac{1}{n}\log p(x_1^n) = -\lim_{n\to\infty}\frac{1}{n}\log p(x_1, x_2, \ldots, x_n)$$

$$= -\lim_{n\to\infty}\frac{1}{n}\sum_{x_i}\log p(x_i) \qquad \text{(基于概率同分布)}$$

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μ)。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μ**，即

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

5. 附录

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

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

(依概率)

所以，左边等于等式右边。

参阅如下文献(英文版P58，中文版P33):

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

5. 附录

类似地，可以证明在 $(x_1, x_2, \ldots, x_n)$ 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。

根据布莱曼渐近均分性定理(Breiman's AEP)，可以推广到：

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

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

5. 附录

对于本章讲义公式(8),《统计自然语言处理》P28, 公式(2-39):

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

(利用布莱曼渐进均分性定理的推广)

证毕。

5. 附录

1. 证明前面P31公式(8): $H(L,q) = -\lim\limits_{n\to\infty}\frac{1}{n}\log q(x_1^n)$

(又见《统计自然语言处理》P28，公式(2-39))

2. 证明：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

5. 附录

2. 证明：两个随机变量之间的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：

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

其中，$f(x)$是凸函数，$g(x)$为任意函数。那么，

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

$$\geq -\log\left(\int_{x\in X}\int_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,\mathrm{p}(y)}\right)\mathrm{d}x\,\mathrm{d}y\right)$$

$$=0$$

证毕。

5. 附录

方法二：

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

那么，

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

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

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

5. 附录

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

$$\le\sum_{x_i\in X}\sum_{y_j\in Y}p(x_i,y_j)\left[\frac{p(x_i)p(y_j)}{p(x_i,y_j)}-1\right]\log_2 e$$

$$=\left[\sum_{x_i\in X}\sum_{y_j\in Y}p(x_i)\mathrm{p}(y_j)-\sum_{x_i\in X}\sum_{y_j\in Y}p(x_i y_j)\right]\log_2 e$$

$$=\left[\sum_{x_i\in X}p(x_i)\sum_{y_j\in Y}\mathrm{p}(y_j)-\sum_{x_i\in X}\sum_{y_j\in Y}p(x_i y_j)\right]\log_2 e=0$$

$\boxed{I(X;Y)\ge 0}$

证毕。

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

5. 附录

1. 证明前面P31公式(8): $H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n)$

(又见《统计自然语言处理》P28，公式(2-39))

2. 证明：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

5. 附录

3. 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 $$p^* = \arg\max_{p \in P} H(p) \qquad (19)$$

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。

5. 附录

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，

$$H(p) = -\sum_{a,b} \tilde{p}(b)\, p(a|b) \log p(a|b) \tag{20}$$

5. 附录

即求解使 $H(p)$ 值最大的条件概率 $p^*(a|b)$：

$$p^*(a|b) = \arg\max_{p \in P} H(p)$$

$$= \arg\max_{p \in P}\left(-\sum_{a,b} \tilde{p}(b)\, p(a|b) \log p(a|b)\right)$$

目标函数

5. 附录

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：

$$E_{\tilde{p}}(f_j) = \sum_{a,b} \tilde{p}(a,b) f_j(a,b) \qquad (21)$$

5. 附录

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：

$$E_p(f_j) = \sum_{a,b} p(a,b) f_j(a,b) \qquad (22)$$

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，(22)式变为：

$$E_p(f_j) = \sum_{a,b} \tilde{p}(b) p(a|b) f_j(a,b) \qquad (23)$$

5. 附录

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相同，即：

$$E_p(f_j) = E_{\tilde{p}}(f_j) \qquad (24)$$

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$f_i(i=1, 2, \ldots, k)$，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：

$$P = \{p \mid E_p(f_j) = E_{\tilde{p}}(f_j), j \in \{1,2,\cdots k\}\} \qquad (25)$$

5. 附录

根据以上阐述，归纳如下：

$$p^* = \arg\max_{p \in P} H(p) \tag{19}$$

$$H(p) = -\sum_{a,b} \tilde{p}(b) p(a \mid b) \log p(a \mid b) \tag{20}$$

$$E_{\tilde{p}}(f_j) = \sum_{a,b} \tilde{p}(a,b) f_j(a,b) \tag{21}$$

$$E_p(f_j) = \sum_{a,b} \tilde{p}(b) p(a \mid b) f_j(a,b) \tag{23}$$

$$P = \{ p \mid E_p(f_j) = E_{\tilde{p}}(f_j),\ j \in \{1, 2, \cdots k\} \} \tag{25}$$

5. 附录

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，从而证明，满足(24)式约束条件的解具有如下形式：

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

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

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

谢谢！

Thanks!