第 3 章　概率论&信息论基础知识

中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

内 容 提 要

3.1 概率论基本概念

3.2 信息论基本概念

3.1 概率论基础

★ 概率(probability)

➢ 概率的统计定义

1，频率：事件A在n次重复随机试验中出现n_A次，则比值 n_A/n 为事件A发生的频率,记为$f_n(A)$，即

$$f_n(A)=\frac{n_A}{n}$$

2，概率定义：在同一组条件下所作的大量重复试验中，如果事件A发生的频率总是在一个确定的常数 p 附近摆动，并且逐渐稳定于p，那末数 p 就表示事件A发生的可能性大小，并称它为事件A的概率，记作 $P(A)$ 。

$$\text{频率}\xrightarrow{\text{稳定}}\text{概率}$$

3.1 概率论基础

➢ 概率的公理化定义

设E是随机试验，Ω是E的样本空间，对于E 的每一个事件A赋予一个实数值，表示事件发生的可能性（记为 $P(A)$），则 $P(A)$ 为事件A的概率。概率函数必须满足如下公理：

（1）**非负性**：$P(A)\geq 0$

（2）**规范性**：$P(\Omega)=1$

（3）**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$(A_i\cap A_j=\Phi)$，

则有：

$$P\left(\bigcup_{i=0}^{\infty} A_i\right)=\sum_{i=0}^{\infty} P(A_i)$$

3.1 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 $\{s_1, s_2, \ldots, s_n\}$，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 $\sum_{i=1}^{n} n_N(s_k) = N$ 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$P(s_k)$。

$$\lim_{N \to \infty} q_N(s_k) = P(s_k)$$

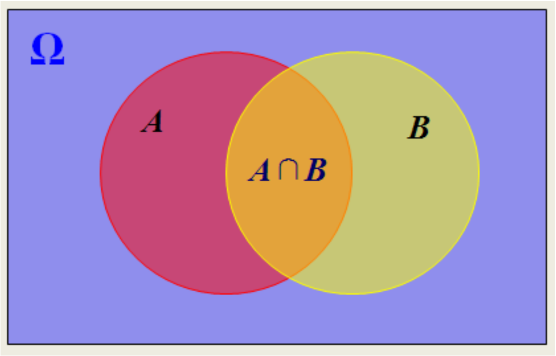
因此，**相对频率**常被用作概率的估计值。这种概率值的估计方法称为**最大似然估计**。

3.1 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(B)>0，那么在给定 B 时 A 的条件概率 P(A|B) 为：

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



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 发生的概率。

一般地，P(A|B) ≠ P(A)。

3.1 概率论基础

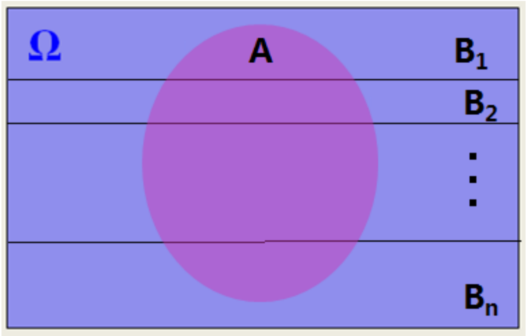
★ 全概率公式

设 Ω 为实验 E 的样本空间，$B_1, B_2, \dots B_n$ 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) $B_i \cap B_j = \Phi$ （$i \neq j$; $i, j = 1, 2, \dots, n$）

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

则称 $B_1, B_2, \dots B_n$ 为样本空间Ω的一个划分。



设A为Ω 的事件，$B_1, B_2, \dots B_n$ 为Ω的一个划分，且 $P(B_i)>0$ （$i=1, 2, \dots, n$），

则 全概率公式为：

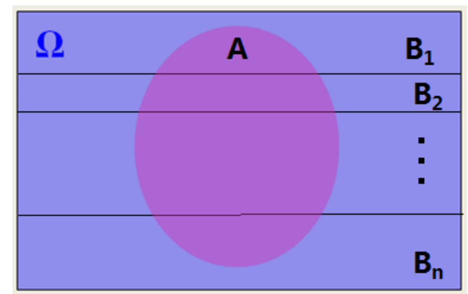
$$P(A) = P(\bigcup_{i=1}^{n} AB_i) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

3.1　概率论基础

★ 贝叶斯法则(Bayes' theorem)

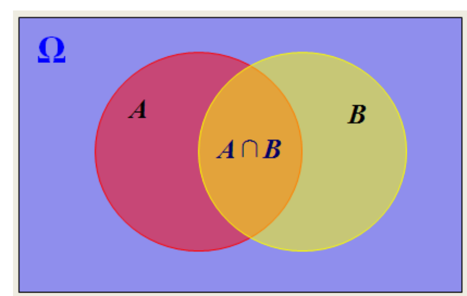
如果 A 为样本空间 Ω 的事件， $B_1, B_2, \ldots, B_n$ 为 Ω 的一个划分，且 $P(A)>0$，$P(B_i)>0$ ($i=1, 2, \ldots, n$) ，那么，

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



当 n=1 时,

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$



3.1　概率论基础

例1：

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？

3.1 概率论基础

解： **假设*G*表示事件“句子确实存在该特殊句法结构”，**

***T*表示事件“程序判断的结论是存在该特殊句法结构”。**

有：

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

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

求：P(G|T)=？

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

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

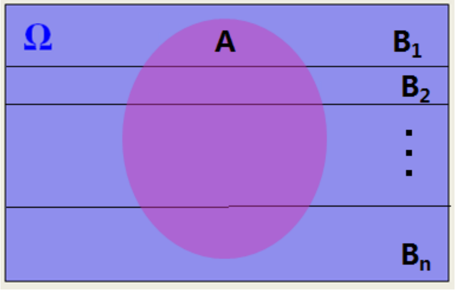
3.1　概率论基础

★ 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, ..., 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 $\bar{x} = [x_1, x_2, \cdots, x_d]^T$ **是 *d* 维特征空间上的某一点，且条件概率密度函数** $p(\bar{x} \mid w_i)$ **是已知的。**

那么，利用贝叶斯公式可以得到后验概率：

$$p(w_i \mid \bar{x}) = \frac{p(\bar{x} \mid w_i)\, p(w_i)}{\sum_{j=1}^{c} p(\bar{x} \mid w_j)\, p(w_j)}$$



3.1 概率论基础

基于最小错误率的贝叶斯决策规则为：

(1) 如果 $p(w_i \mid \bar{x}) = \max\limits_{j=1,2,\cdots,c} p(w_j \mid \bar{x})$ **则** $\bar{x} \in w_i$

(2) 或者：如果 $p(\bar{x} \mid w_i)\, p(w_i) = \max\limits_{j=1,2,\cdots,c} p(\bar{x} \mid w_j)\, p(w_j)$ **则** $\bar{x} \in w_i$

(3) 或者(*c*=2时)：如果

$$l(\bar{x}) = \frac{p(\bar{x} \mid w_1)}{p(\bar{x} \mid w_2)} > \frac{p(w_2)}{p(w_1)}$$

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

3.1 概率论基础

例2：

- 假设在某地区切片细胞中正常(ω_1)和异常(ω_2)两类的先验概率分别为 $P(\omega_1)=0.9$，$P(\omega_2)=0.1$。
- 现有一待识别细胞呈现出状态x，由其类条件概率密度分布曲线查得 $p(x|\omega_1)=0.2$，$p(x|\omega_2)=0.4$，
- 试对细胞x进行分类。

3.1 概率论基础

解：

利用贝叶斯公式，分别计算出状态为x时ω1与ω 2 的后验概率

$$P(\omega_1|X)=\frac{p(x|\omega_1)P(\omega_1)}{\sum_{j=1}^{c}p(X|\omega_j)P(\omega_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(w_2|X)=1-P(w_1|X)=0.182$$

– 根据贝叶斯决策有

$P(\omega 1|x)=0.818>P(\omega 2|x)=0.182$

– 分析:错误概率是多少？

- 判断为正常细胞，错误率为0.182
- 判断为异常细胞，错误率为0.818

因此判定该细胞为正常细胞比较合理。

3.1 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），令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 k 表示A 在这 n 次实验中发生的次数，若随机变量X 的分布律为

$$P(X = k) = C_n^k p^k q^{n-k} \qquad k = 0,1,2,\cdots,n$$ 其中 $0 < p < 1$

则称X服从参数为n，p的二项分布，记为 $X \sim B(n, p)$

在自然语言处理中，一般以句子为处理单位。假设一个语句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。

3.1 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, … 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：

$$\mathrm{EX} = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k + \ldots$$

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

3.1　概率论基础

★方差(variance)

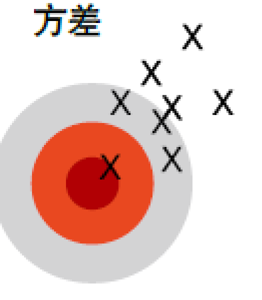
一个随机变量的方差描述的是该随机变量的值偏离其期望值的程度。

设 X 为一随机变量，其方差为：

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

$$= (x_1 - EX)^2 p_1 + (x_2 - EX)^2 p_2 + \ldots + (x_k - EX)^2 p_k + \ldots$$

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



3.1 概率论基础

★ 偏置(Bias)（偏置-方差分解）

假设有K个数据集，每个数据集都是从一个分布p(t,x)中独立的抽取出来的(t代表要预测的变量，x代表特征变量)。对于每个数据集D，可以在其基础上根据学习算法来训练出一个模型y(x;D)来。在不同的数据集上进行训练可以得到不同的模型。学习算法的性能是根据在这K个数据集上训练得到的K个模型的平均性能来衡量的，亦即：

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2\right] \\
&= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]-h(\mathbf{x})\}^2}_{(\text{bias})^2}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.
\end{aligned}
$$

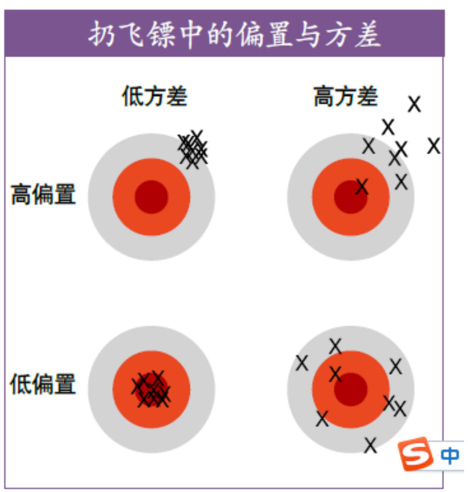
其中的h(x)代表生成数据的真实函数，亦即t=h(x)。

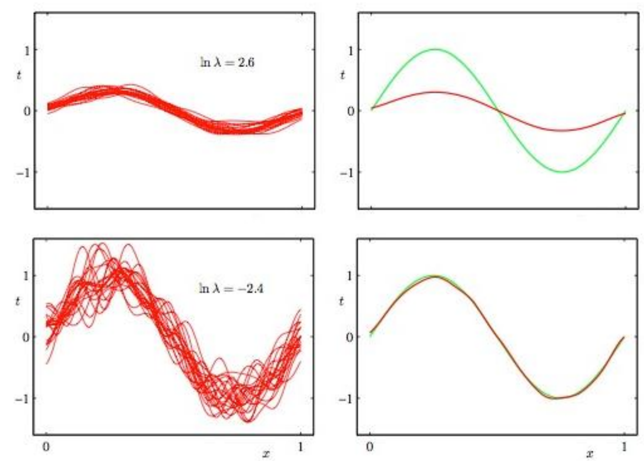
可以看到，给定学习算法在多个数据集上学到的模型的和真实函数h(x)之间的误差，是由**偏置(Bias)**和**方差(Variance)**两部分构成的。

3.1 概率论基础

★ 偏置(Bias)（偏置-方差分解）

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2\right] \\
&= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]-h(\mathbf{x})\}^2}_{(\text{bias})^2}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.
\end{aligned}
$$



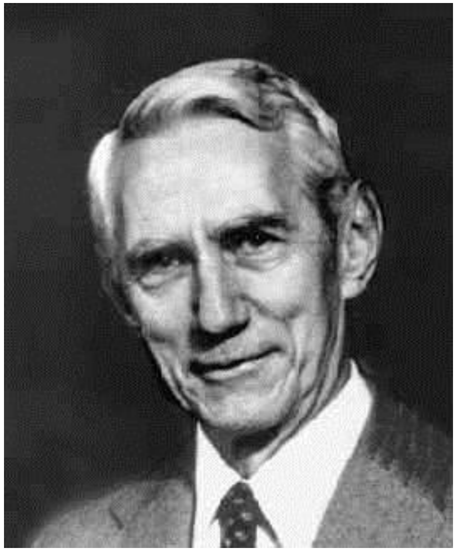


内 容 提 要

3.1 概率论基本概念

3.2 信息论基本概念

3.2 信息论基础



Claude Shannon

信息论创立的标志是1948年Claude Shannon （香农）发表的论文 “A Mathematical Theory of Communication”

在这篇文章中香农创造 性的采用**概率论**的方法来研究通信中的问题，并且对信息给予了科学的**定量描述**，第一次 提出了**信息熵**的概念。

3.2　信息论基础

（1）**狭义信息论：又称香农信息论**。主要通过数学描述与定量分析，研究通信系统从信源到信宿的全过程，包括信息的测度、信道容量以及信源和信道编码理论等问题，强调通过编码和译码使收、发两端联合最优化，并且以定理的形式证明极限的存在。这部分内容是**信息论**的**基础理论**。

（2）**一般信息论：也称工程信息论**。主要也是研究信息传输和处理问题，除香农信息论的内容外，还包括噪声理论、信号滤波和预测、统计检测和估计、调制理论、信息处理理论以及保密理论等。

（3）**广义信息论**：不仅包括上述两方面内容，而且包括所有与信息有关的自然和社会领域，如模式识别、计算机翻译、心理学、遗传学、神经生理学、语言学、语义学甚至包括社会学中有关信息的问题。

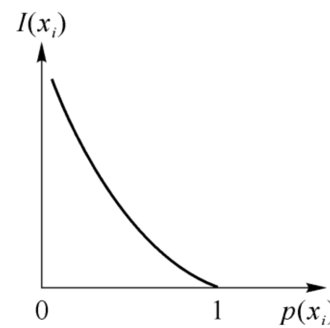
3.2 信息论基础

3.2.1 信息的度量的几个重要的概念：

1. 自信息：

一个事件（消息）本身所包含的信息量，它是由事件的不确定性决定的: 随机事件的**自信息量**定义为该事件发生概率的对数的负值; 如，设事件 x_i 的概率为 $p(x_i)$，则它的**自信息**定义为

$$\boxed{I(x_i) \overset{def}{=} -\log p(x_i) = \log \frac{1}{p(x_i)}}$$



自信息量

自信息量单位

1）取对数底为2，信息量的单位为比特（bit，binary unit）。

当 $p(x_i)$ =1/2时，$I(x_i)$ =1比特。

2）若取自然对数e为底，自信息量的单位为奈特（nat，natural unit）

1奈特= $\log_2 e$ 比特=1.443比特

3）工程上用以10为底较方便。若以10为对数底，则自信息量的单位为哈特莱（Hartley）。1哈特莱= $\log_2 10$ 比特=3.322比特

3.2 信息论基础

2．信息熵（平均自信息）

随机变量 X有A_1，A_2，…，A_n共n个可能的结局，每个结局出现的机率分别为p_1,p_2 ，…，p_n，则随机变量 X 的 **平均自信息量**

$$H(X) = -\sum_{1}^{n} p_i \log p_i$$

定义为X 的**信息熵**,记为**H(X)** 或 **H(p)** 。

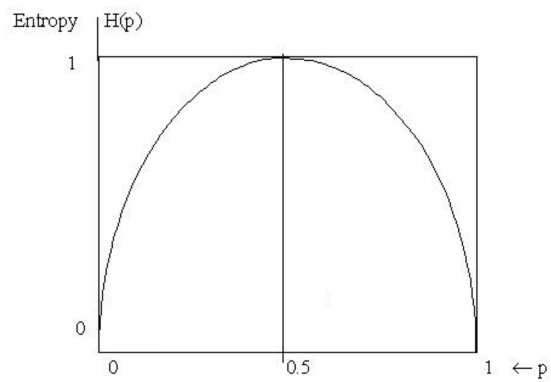
通常熵的单位为二进制位比特 (bit)

这里q 为的所有 X可能取值的个数； 其中，约定 $0 \log 0 = 0$

X的具体内容跟信息量无关，我们只关心概率分布

3.2　信息论基础

◆ 熵的图形



$$H(X) = -\sum_{1}^{n} p_i \log p_i$$

◆ 熵的性质

$$0 \le H(X) \le \log |X|$$

第一个等号在X为确定值的时候成立（没有变化的可能）

第二个等号在X均匀分布的时候成立。

均匀分布的时候，熵最大

例: 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

3.2 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）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。

3.2 信息论基础

3. 联合熵(joint entropy)

一个随机变量的不确定性可以用熵来表示，这一概念可以方便地推广到多个随机变量。

如， 二维随机变量 XY 的概率空间表示为

$$\begin{bmatrix} XY \\ P(XY) \end{bmatrix} = \begin{bmatrix} x_1 y_1 & \cdots & x_i y_j & \cdots & x_n y_m \\ p(x_1 y_1) & \cdots & p(x_i y_j) & \cdots & p(x_n y_m) \end{bmatrix}$$

其中 $p(x_i, y_j)$ 满足概率空间的非负性和完备性：

$$0 \le p(x_i y_j) \le 1, \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i y_j) = 1$$

3.2 信息论基础

离散型二维随机变量 X Y 的**联合熵** *H (X,Y)* 定义为：

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

联合熵实际上就是描述一对随机变量平均所需要的信息量。

是二维随机变量*XY* 的不确定性的度量。

3.2　信息论基础

4. 互信息

一个事件 y_j 所给出关于另一个事件 x_i 的信息定义为**互信息**，用 $I(x_i;y_j)$ 表示。

$$I(x_i;y_j) \overset{def}{=} I(x_i) - I(x_i|y_j) = \log\frac{p(x_i|y_j)}{p(x_i)}$$

互信息 $I(x_i;y_j)$ 是已知事件 **y_i** 后所消除的关于事件 **x_i** 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。

如，自然语言中互信息值越大，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。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。

3.2 信息论基础

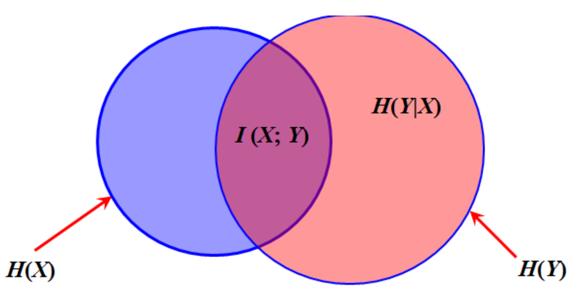
5. 条件熵(conditional entropy)

有两个变量：x,y。它们不是独立的。给定随机变量 X 的情况下，随机变量 Y 的条件熵定义为：

$$H(Y \mid X) = \sum_i p(x_i) H(Y \mid x_i) = -\sum_i \sum_j p(x_i) p(y_j \mid x_i) \log p(y_j \mid x_i)$$

$$= -\sum_i \sum_j p(x_i y_j) \log p(y_j \mid x_i)$$

其中， H(Y| X) 表示已知 X 时，Y 的平均不确定性



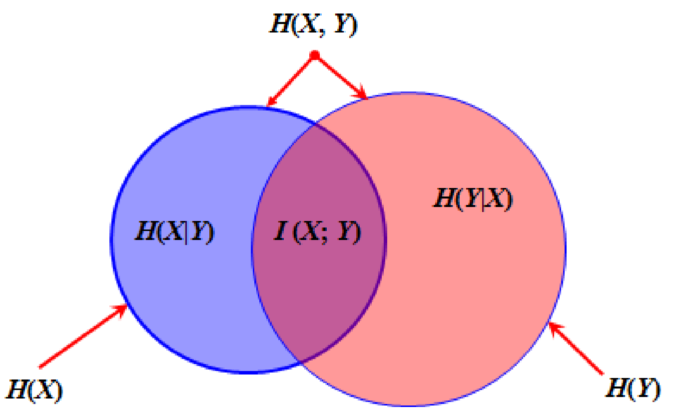
含义：

知识（X）减少（Y）的不确定性

$$H(Y \mid X) \le H(Y)$$

3.2 信息论基础

4. 联合熵与信息熵、条件熵的关系：



互信息、条件熵与联合熵

推论：

$$H(XY) \le H(X) + H(Y)$$

当二维随机变量*X*、*Y* 相互独立时等号成立。

3.2 信息论基础

例：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？

3.2 信息论基础

解： $H(X)$：

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

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

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

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

$H(Y)$：

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

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

3.2 信息论基础

$H(X|Y)$: $$H(X|Y) = -\sum_j \sum_i p(y_j x_i) \log p(x_i | y_j)$$

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	

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

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

$p(x_1 | y_2)$ ▪▪▪ $p(x_4 | y_2)$ ▪▪▪ $p(x_1 | y_4)$ ▪▪▪ $p(x_4 | y_4)$

$H(X|Y)$:

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

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

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

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

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

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

同理:

$H(Y|X)$:

$$H(Y|X)=-\sum_i\sum_j p(x_i y_j)\log p(y_j|x_i)$$

$H(Y|X)$=13/8 (bits)

可见，$H(Y|X)\neq H(X|Y)$

3.2　信息论基础

***H*(*X*, *Y*):**

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

$H(X, Y)$=27/8 (bits)

3.2　信息论基础

5. 熵率(entropy rate)

一条长度为 n 的信息(X_1, ..., X_n) ，每一个字符或字的熵为：

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

其中，变量 X_{1n}表示随机变量序列 (X_1, ..., X_n)，

有时写成：$x_1^n = (x_1, x_2, \cdots, x_n)$

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

6. 相对熵(relative entropy)--Kullback-Leibler divergence, KL 距离

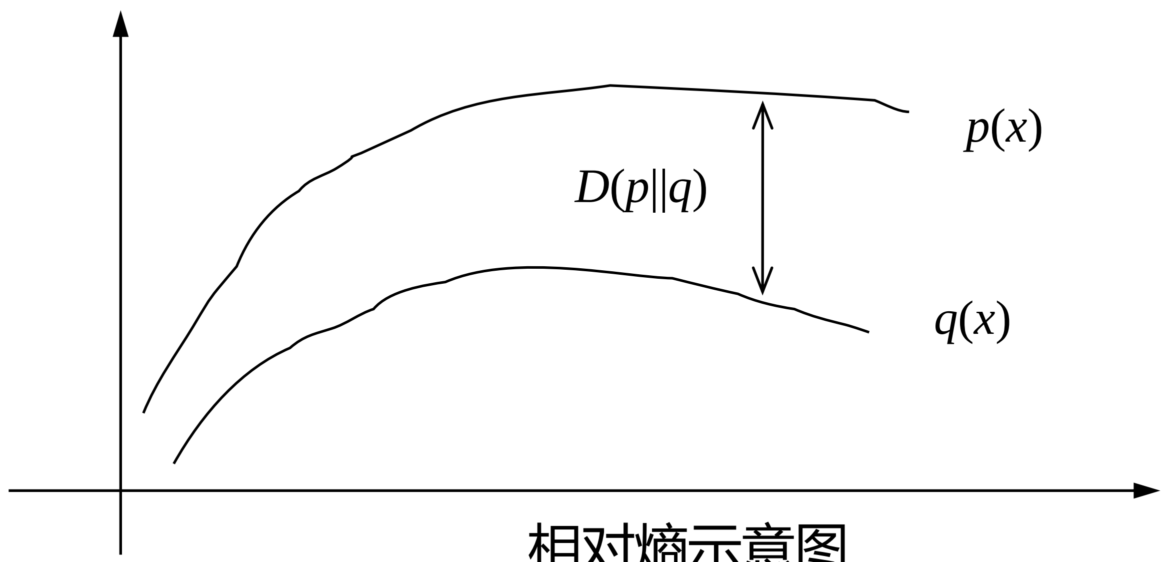
两个概率分布 p(x) 和 q(x) 的相对熵定义为：

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

该定义中约定 0 log (0/q) = 0, p log (p/0) = ∞。

3.2 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。



相对熵示意图

3.2 信息论基础

7. 交叉熵(cross entropy)

一个随机变量 X ~ p(x)，q(x)为近似 p(x) 的概率分布，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。

3.2 信息论基础

8. 语言与其模型 交叉熵

对于语言 L = (Xi) ~ 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。

在设计模型 q 时，目的是使交叉熵最小，从而使模型最接近真实的概率分布 p(x)。

3.2 信息论基础

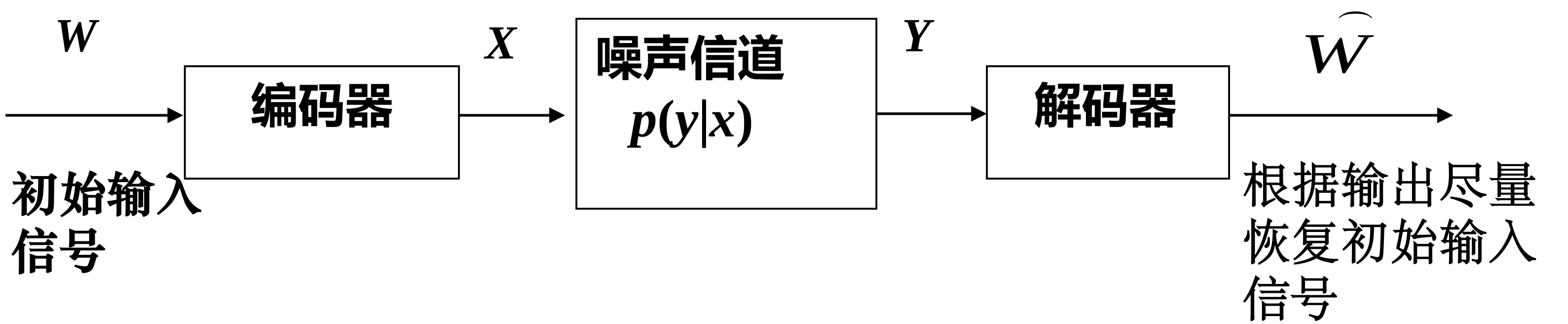
9. 困惑度(perplexity)

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

3.2.2 噪声信道模型(noisy channel model)



过程示意图：

3.2 信息论基础

在自然语言处理中，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：

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

3.2 信息论基础

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

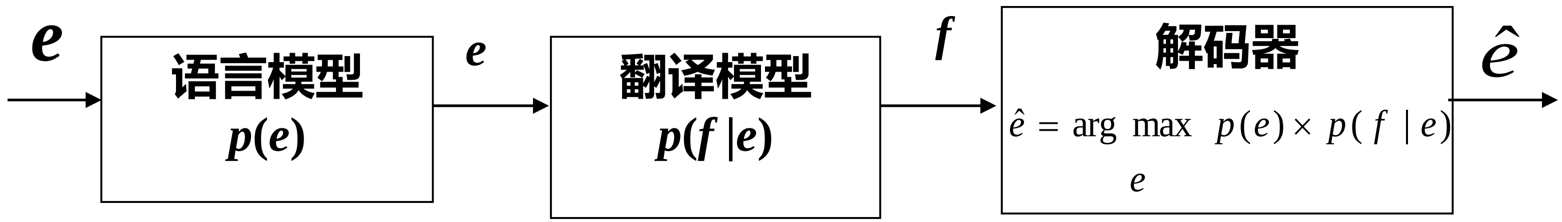
$p(e) \times p(f|e)$ 最大，即：

$$\hat{e} = \arg\max_{e} p(e) \times p(f|e)$$

语言模型　　**翻译模型(translation model)**

3.2 信息论基础

统计翻译系统框架：



法语句子 *f* □□ ⟹ **英语句子** $\hat{e}$

建立一个源语言 f 到目标语言 e 的统 计翻译系统，

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）设计有效快速的搜索算法求解 $\hat{e}$ 使得 p(e)×p(f | e)最大。

参考文献：

宗成庆，统计自然语言处理（第2版）课件

王莉，信息工程基础（课件），北京科技大学

在 此 表 示 感 谢 ！

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

谢谢各位！