第2章　数学基础

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)

2.1 概率论基础

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

如果一个实验的样本空间是$\{s_1, s_2, \ldots, s_n\}$，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：

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

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

2.1 概率论基础

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

当N越来越大时，相对频率 $q_N(s_k)$ 就越来越接近 s_k的概率$P(s_k)$。事实上，

$$\lim_{N\to\infty} q_N(s_k) = P(s_k) \qquad (3)$$

因此，$\boxed{\text{相对频率常被用作概率的估计值}}$。这种概率值的估计方法称为<u>最大似然估计</u>。

2.1 概率论基础

◆链式法则

2个事件同时发生的概率：

P(a, b) = P(a | b) * P(b)

推广到N个事件，概率链式法则：

P(X1, X2, ... Xn) = P(X1 | X2, X3 ... Xn) *
P(X2 | X3, X4 ... Xn) ... P(Xn-1 | Xn) * P(Xn)

2.1 概率论基础

◆全概率公式

如果 A 为样本空间 Ω 的事件， $B_1, B_2, \ldots, B_n$ 为样本空间 Ω 的一个划分，且 $P(B_i) > 0$ （ $i = 1, 2, \ldots, n$ ），则全概率公式为：

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

2.1 概率论基础

◆贝叶斯定理

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

当 $n = 1$ 时，$\boxed{P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}}$

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)

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)。

2.2 信息论基础

熵表示信源 X 每发一个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：

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

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

(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。

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：

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

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。

2.2 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：

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

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

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

$$=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2 p(y\mid x) \qquad (3)$$

条件熵表示在已知 X 的情况下，Y的不确定性

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：

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\boxed{\log[p(x)p(y\mid x)]}$$

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

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

全概率公式

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

$$=H(X)+H(Y\mid X)\qquad(4)$$

注意：$H(Y|X)\neq H(X|Y)$。

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：

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

这个数值我们也称为 熵率(entropy rate)。

X_{1n}表示随机变量序列 $(X_1, \ldots, X_n)$，$x_{1n} = (x_1, \ldots, x_n)$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}=(为，传，播，……，活，动，。)
- $$H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{66} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$$

2.2 信息论基础

◆相对熵(relative entropy, 或称 Kullback-Leibler divergence, 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：

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

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。

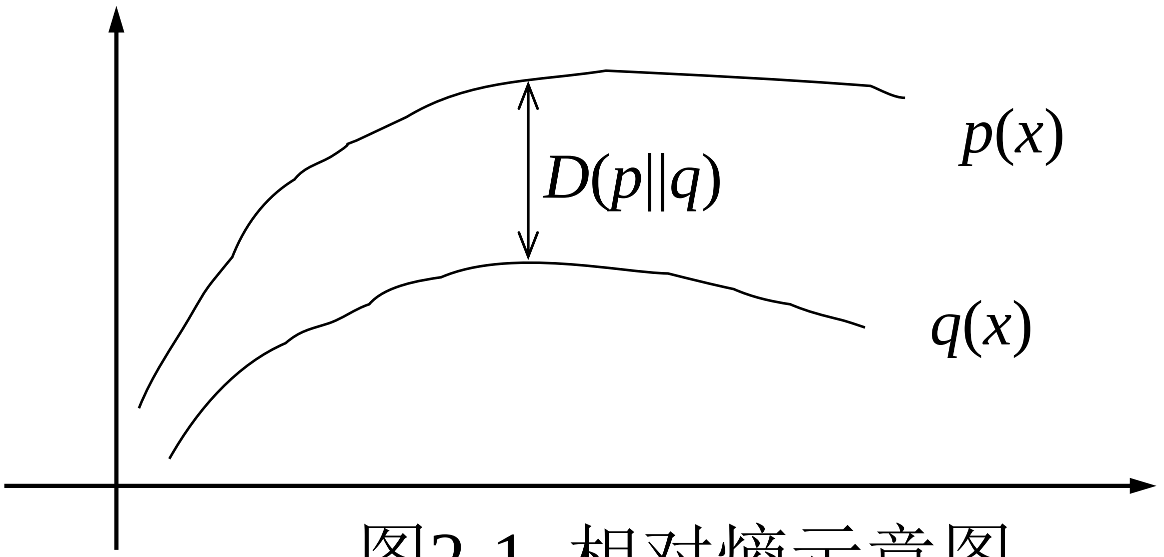


图2-1. 相对熵示意图

2.2 信息论基础

◆交叉熵(cross entropy)

如果一个随机变量 $X \sim p(x)$ (真实值)， $q(x)$为用于近似 $p(x)$ 的概率分布(估计值)，那么，随机变量 X 和模型 q 之间的交叉熵定义为：

$$\begin{aligned} H(X,q) &= H(X) + D(p \| q) \\ &= -\sum_{x} p(x) \log q(x) \end{aligned} \qquad (7)$$

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

说明： 在机器学习中经常用$p(x)$表示**真实**数据的概率分布（真实数据的概率分布往往无法获得，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$D(p||q)$。**

机器学习算法中通常采用**交叉熵**作为模型优化目标，即损失函数Loss。

2.2 信息论基础

例如：利用交叉熵衡量一个语言模型q的好坏

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

其中，

$p(x_1^n)$ 为 L 中 x_1^n 的概率（真实值）；

$q(x_1^n)$ 为模型 q 对 x_1^n 的概率估计值。

2.2 信息论基础

◆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 $l_1^n = l_1 \ldots l_n$，L 的困惑度 PP_q 定义为：

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

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：

$$I(X; Y) = H(X) - H(X \mid Y) \qquad (11)$$

根据$H(X)$ 和 $H(X|Y)$ 的定义：

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

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

2.2 信息论基础

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

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

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

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

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \qquad (12)$$

互信息 $I(X;Y)$ 是在知道了 Y 后 X 的(熵)不确定性减少量，即 Y 的值透露了多少关于 X 的信息量。

2.2 信息论基础

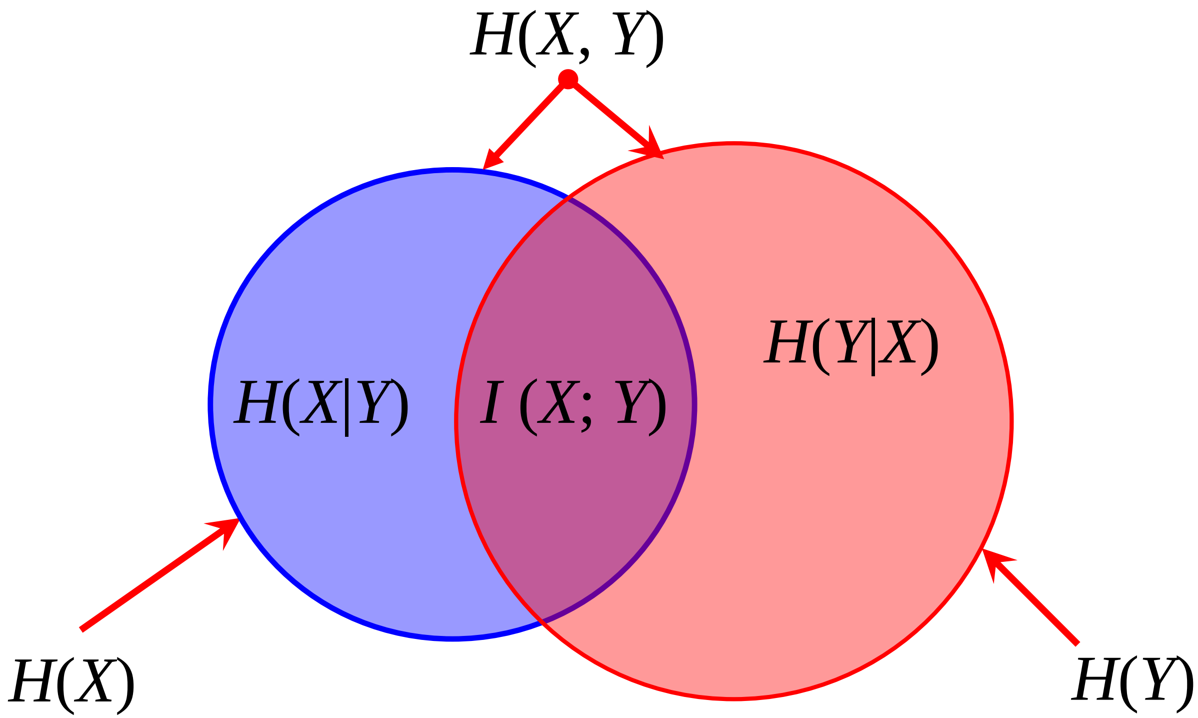


图 2-2. 互信息、条件熵与联合熵

2.2 信息论基础

例如：利用互信息解决汉语分词问题

为||人||民 服 务。

?

利用互信息值估计两个汉字结合的强度：

I(为;人) or I(人;民)

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

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：

2.2 信息论基础

真连续次数

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

假连续次数

c_i，c_{i+1} 是一个有序字对，表示两个连续汉字，$N(c_i c_{i+1})$表示字符串$c_i\, c_{i+1}$构成的词出现的频率，$N(\ldots c_i|c_{i+1}\ldots)$ 表示c_i 作为上一个词的词尾 且 c_{i+1} 作为相邻下一个词的词头出现的频率。例如：“为人”出现5次，“为|人民”出现 20次，那么，$Couple$(为, 人)＝0.2。

注意：此处“|”不表示条件概率！

2.2 信息论基础

<u>理由</u>：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。

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。

区别：分母不同。

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

◆噪声信道模型(noisy channel model)

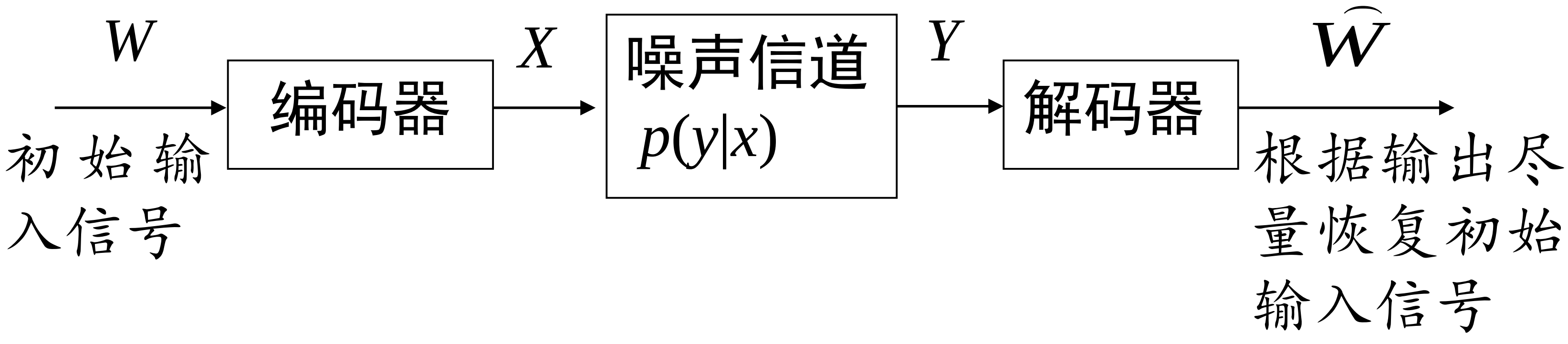
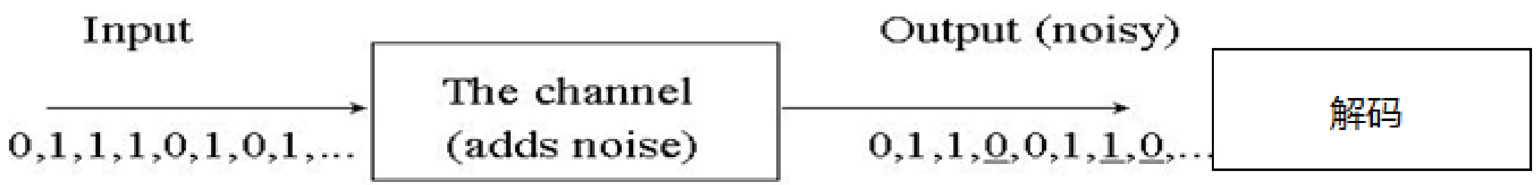


图 2-3. 噪声信道模型示意图

2.2 信息论基础

噪声信道模型在NLP中的应用



解码器的目标是试图通过带噪声的输出信号恢复输入信号，形式化定义为：

语言模型

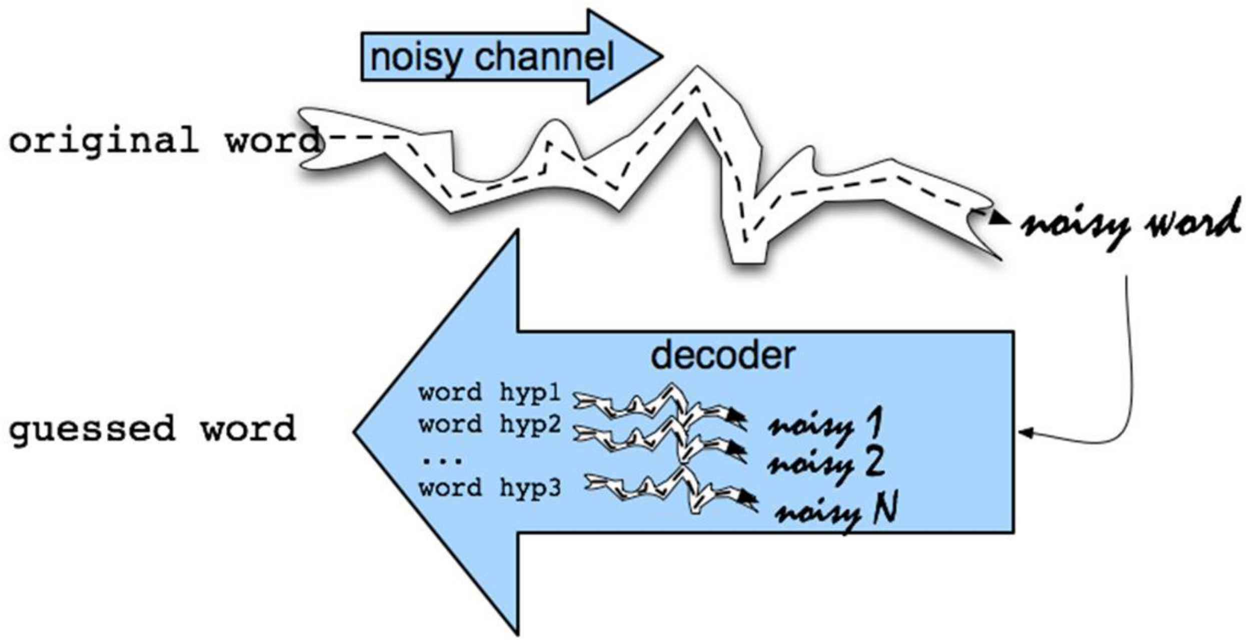
$$\hat{I} = \arg\max_{I} P(I \mid O) = \arg\max_{I} \frac{P(O \mid I)P(I)}{P(O)} = \arg\max_{I} P(O \mid I)P(I)$$

翻译模型

2.2 信息论基础

噪声信道模型在NLP中的应用

应用场景：语音识别、机器翻译、拼写纠错、密码破译等



$$\underset{i}{\operatorname{argmax}}\ P(\text{错误的拼写}|\text{正确的拼写}i)\ .\ P(\text{正确的拼写}i)$$

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

以“打”字为例，用作动词时有24个含义：

(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。

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

2.3 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，

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

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

2.3 应用举例

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)} \tag{16}$$

其中，$N(s_i)$ 是在训练数据中语义s_i 出现的次数，而 $N(v_k, s_i)$ 为 语义s_i 与词 v_k 共现的次数。

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

$N(w)$ 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：

对于“打”字而言，由于其有24个义项，因此我们分别计算24次s_i，并选出其中下列概率值最大的：

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

假设第一个待计算 的词义s_1为“敲击(beat)”，其所在的句子为：

他 对 打 鼓 很 在行 。 (取上下文： ±2)

-2 -1 ↑ +1 +2

2.3 应用举例

他 对 打 鼓 很 在行 。 (取上下文： ±2)

-2 -1 ↑ +1 +2

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

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

假若“打”在所有样本中总共出现了800次，那么，

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

2.3 应用举例

- 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

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

②对于w 的每个语义 s_i 计算：

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

利用已标注的大规模数据－训练过程

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：

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

称测试过程 标注过程或

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：

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

Thanks

谢谢!