信息论基础

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

约定 $0\log 0 = 0$

$p(x)\downarrow \longrightarrow H(x)\uparrow \longrightarrow$ 不确定性↑

〔信源X每发一个符号所提供的平均信息量〕

〔在自然语言处理中的应用：
使熵H(p)值最大的模型因来推断某种语言现象存在的可能性〕

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

描述一对随机变量平均所需要的信息量

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

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

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

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

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

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

$$= H(X) + H(Y|X)$$ 同理 $H(Y) + H(X|Y)$

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

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

$$= -\left(-\frac{1}{2} + \left(-\frac{1}{2}\right) + \left(-\frac{6}{8}\right)\right) = \frac{7}{4}\ (bits)$$

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

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

$= \frac{1}{4}\cdot\left[-\sum_{i=1}^{4} p(x_i|y_1)\log p(x_i|y_1)\right) + \cdots$

$= \frac{1}{4}\cdot\left[-\left(\frac{1}{2}\times\log_2\frac{1}{2}\right) + \cdots\right] + \cdots$

结论: $H(Y|X) \neq H(X|Y)$

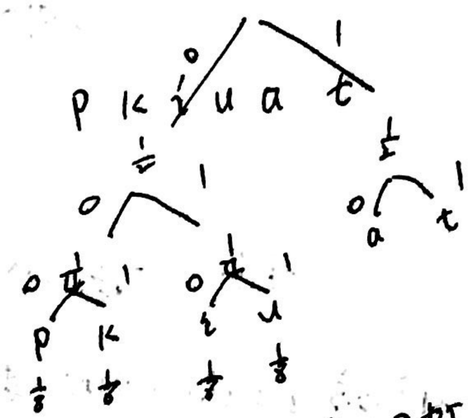
波利尼西亚语 $\frac{1}{8}$ $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$

p: $\frac{1}{8}$ t: $\frac{1}{4}$ k: $\frac{1}{8}$ a: $\frac{1}{4}$ i: $\frac{1}{8}$ u: $\frac{1}{8}$

$H(P) = 2.5$ bits

100 00 101 01 110 111

哈夫曼编码



带权路径长度最小的二叉树

元 V={a, i, u}

辅 C={p, t, k}

假定所有单词都由 CV (consonant − vowel) 音节序列（元—辅）组成，其联合概率与边缘概率分布如下：

V \ C	p	t	k	P(·, V)
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{2}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0	$\frac{1}{4}$
u	0	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
P(C, ·)	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	

求联合熵：

p: $\frac{1}{16}$ t: $\frac{3}{8}$ k: $\frac{1}{16}$ a: $\frac{1}{4}$ i: $\frac{1}{8}$ u: $\frac{1}{8}$

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

$H(C) = 1.061$ bits

$H(V|C) =$

熵率：对于一个长度为 n 的信息，每一个字符或字的熵为

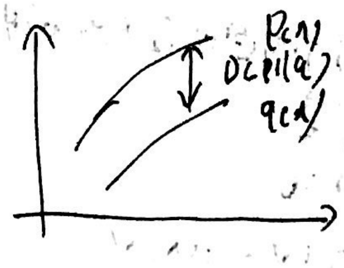
$H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$

X_{1n} 表示：$(X_1, \cdots, X_n)$ 有时将 X_{1n} 写作 X_1^n

相对熵：或称 KL 距离 kullback-Leibler divergence relative entropy

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

两个概率分布 p(x) 和 q(x) 的相对熵定义



其中 $0 \log (0/q) = 0$ $p \log (p/0) = \infty$

交叉熵：

$X \sim p(x)$　$q(x)$ 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：

$$H(X,q) = H(X) + D(p\|q)$$

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

推导：$-\sum_x p(x)\log_2 p(x) + \sum_x p(x)\log_2 \frac{p(x)}{q(x)} = -\sum_x p(x)\log_2 q(x)$

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

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

$x^n = x_1, \cdots, x_n$ 为语言 L 的前 n 个序列（样本）

$p(x^n)$ 为 x^n 的概率

$q(x^n)$ 为估计值

定理：假定语言 L 是熵意义下的历史性遍历过程，x_1^n 为 L 的样本，那么，有

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

困惑度：

在估计语言模型时，常用困惑度来代替交叉熵衡量模型的好坏

语言 L 的样本：$l^n = l_1 \cdots l_n$，

困惑度 PP_q 定义为 $PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n}\log q(l^n)}$

$$= \log_2 [q(l^n)]^{\frac{1}{n}}$$

互信息：$(X,Y) \sim p(x,y)$　X、Y 之间的互信息 $I(X;Y)$ 定义为：

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

$$= -\sum_x p(x)\log p(x) + \sum_x \sum_y p(x,y)\log p(x|y)$$

$$= \sum_x \sum_y p(x,y)(\log_2 p(x|y) - \log_2 p(x))$$

$$= \sum_x \sum_y p(x,y)\left(\log \frac{p(x|y)}{p(x)}\right)$$

$$= \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(x)p(y)}$$

表示知道了 Y 的值以后 X 的不确定性的减少量

$H(X|X) = 0$

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

$= I(X;X)$

可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

$$\begin{cases} I(x,y)>0 & \text{关联性强} \\ I(x;y)\approx 0 & \text{关联性弱} \\ I(x;y)<0 & \text{互斥分布} \end{cases}$$

22字耦合度

不考虑两个字不连续出现的情况

$$Couple(C_i,C_{i+1})=\frac{N(C_iC_{i+1})}{N(C_iC_{i+1})+N(\cdots C_i|C_{i+1}\cdots)}$$

统计样本中 C_i C_{i+1} 连续出现在一个词中的次数 / 连续出现的总次数

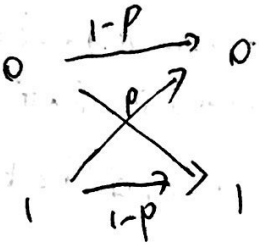
噪音信道模型

基本假设：一个信道的输出以一定的概率依赖于输入

目标：优化噪声信道中信号传输的吞吐量和准确率

输入 $\xrightarrow{W}$ 编码器 $\xrightarrow{X}$ 噪声信道 $p(y|x)$ $\xrightarrow{Y}$ 解码器 $\xrightarrow{W}$ 输出

二进制对称信道：



p：传递概率

信道容量：$C=\max_{p(x)} I(X;Y)$ ：输入与输出的互信息达到最大值

词义消歧问题：(word sense disambiguation, WSD)

① 基于上下文分类的消歧方法

(1) 基于贝叶斯分类器

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 n 个词义记作 $S=\{S_1,S_2,\cdots,S_n\}$，那么，可通过计算 $\arg\max_{S_i} p(S_i|C)$ 确定 w 的词义

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

考虑分母的不变性，并运用如下独立性假设

$$p(C|S_i)=\prod_{V_n\in C} p(V_n|S_i)$$

因此 $\hat{S}_i=\arg\max_{S_i}\left[p(S_i)\prod_{V_n\in C} p(V_n|S_i)\right]$

概率 $P(V_{k1}|S_1)$ 和 $P(S_1)$ 都可用最大似然估计求得

$$P(V_{k1}|S_1)=\frac{n(V_{k1},S_1)}{n(S_1)} \qquad P(S_1)=\frac{n(S_1)}{n(w)}$$

$n(S_1)$：训练数据中词 w 用于语义 S_1 时的次数

$n(V_{k1},S_1)$：w 用于语义 S_1 时词 V_{k1} 出现在 w 的上下文的次数

$n(w)$：语义词 w 在训练语料中出现的总次数

② 基于最大熵的消歧方法

在给定条件 $b\in B$ 下（已知的信息），估计某个事件（未知分布）的概率 $p(a|b)$，以概率使熵值（$P(A|B)$）最大

$$p^*(a,b)=\frac{1}{z(b)}\exp\left(\sum_j \lambda_j \cdot f_j(a,b)\right)$$

λ_j：特征权重

$f_j(a,b)$ 特征函数

j 代表有 j 个特征

其中 $z(b)=\sum_a \exp\left(\sum_j \lambda_j \cdot f_j(a,b)\right)$

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

确定特征函数：

对于语义消歧而言，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，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。

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

上下文条件 b 包括：词形信息（实词数、时态等）

词性信息

词形+词性

表示方法：① 顺序无关（词袋模型）（通常用向量表示）

取上下文窗口范围内的词形

$\{$与，人，交道，。$\}$

② 位置有关（模板表示）

$$f(a,b)=\begin{cases}1 & \text{if } a=S_j \text{ and } b=\langle(C,N),(N,PU)\rangle\\ 0 & \text{otherwise}\end{cases}$$

特征选择
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
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、利用遗墨式特征选择方法

选取n个，则 L=n+1

λ的确定：

GIS算法：

$(a,b) \in A \times B$ 所特征函数之和为一常量C，即：

$$\sum_i f_i(a,b) = C$$

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

举例如一个修正特征 f_t：$f_t(a,b) = C - \sum_{i=1}^{n} f_i(a,b)$

$f_t(a,b)$的取值范围为 0～C

1. 初始化
2. 计算每个 f_i 的期望 $E_{\tilde{P}}(f_i)$
3. 迭代
4. 终止

限定迭代次数

前后似然 $L(P)$ 的变化极小

$$|L_{n+1} - L_n| < \varepsilon$$

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

修正为：

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