熵和语言模型评价

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

最优编码

- 有一个房间中有时没有人，有时甲在房间中，有时乙在房间中，有时甲乙都在房间中，房间状态服从下面的概率分布：

房间状态	房间没有人	甲在房间	乙在房间	甲乙均在房间
概率	0.5	0.125	0.125	0.25

- 定时记录房间状态(消息)，将房间状态编码，并通过通信设备发送出去。如何编码，使得连续发送消息时，编码长度最短？
- 定长编码　2个二进制位
  发送一个消息，平均2个二进制位。

最优编码

- 变长编码：给小概率信息赋以较长的编码，而给大概率消息赋以较短的编码。

消息	编码
房间没有人	0
甲在房间	110
乙在房间	111
甲乙均在房间	10

- 发送一个消息，平均需要1.75个二进制位。

$$0.5\times1+0.125\times3+0.125\times3+\times0.25\times2=1.75$$

最优编码

- 随机变量 X 服从概率分布 P，如果消息 x 的分布密度为 $p(x)$，则给其分配一个长度为 $\lceil -\log_2 p(x) \rceil$ 个二进制位的编码。
- 发送一个消息平均需要 $-\sum p(x)\log_2 p(x)$ 个二进制位。
- 消息的编码长度大，可理解为消息所含信息量大。
  消息的编码长度小，则消息所含信息量小。
  平均信息量即为发送一个消息的平均编码长度 。
- 信息论中用熵描述随机变量平均信息量。

熵(entropy)

- 定义**1** 熵 设 X 是取有限个值的随机变量，它的分布密度为

  $p(x) = P\{X=x\}$，　且$x \in \mathcal{X}$

  则，X的熵定义为

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

- 熵描述了随机变量的不确定性。

- 规定 $0\log_a 0 = 0$
- 通常a=2，此时熵的单位为比特。
- 熵的基本性质:

1. $H(X) \geqslant 0$，等号表明确定场(无随机性)的熵最小。
2. $H(X) \leqslant \log|\mathcal{X}|$，等号表明等概场的熵最大。

熵

例子 1：假定有一种语言 P 有 6 个字母 p、t、k、a、i、u，字母的分布密度为：

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

则随机变量 P 的熵为：

语言的字母熵

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

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

$$= 2\frac{1}{2}\ \text{bit}$$

联合熵、条件熵

- 定义2 联合熵 设X、Y是两个离散型随机变量，它们的联合分布密度为$p(x,y)$，则X,Y的联合熵定义为：

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

- 定义3 条件熵 设X、Y是两个离散型随机变量，它们的联合分布密度为$p(x,y)$，则给定X时Y的条件熵定义为：

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

- 链式规则　$H(X,Y) = H(X) + H(Y|X)$

熵率(entropy rate)

- 信息量的大小随着消息长度的增加而增加，为了便于比较，一般使用熵率的概念，熵率一般也称为字符熵(per-letter entropy)或词熵(per-word entropy)。
- 定义**4**: 熵率，对于长度为*n*的消息，熵率定义为：

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

- 语言L的熵

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots X_n)$$

互信息(mutual information)

- 根据链式规则，有：

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

  可以推导出:

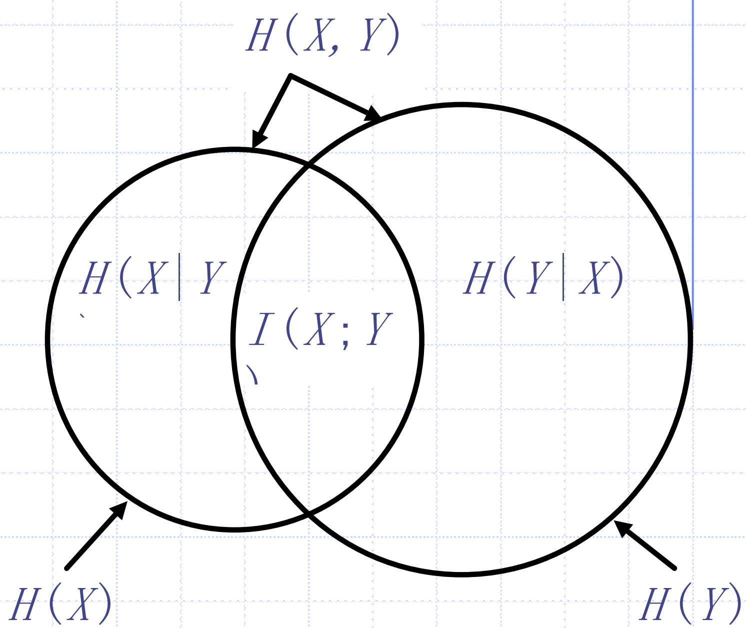
  $$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- $H(X)$和 $H(X|Y)$ 的差称为互信息，一般记作$I(X;Y)$
- $I(X;Y)$ 描述了包含在X中的有关Y的信息量，或包含在Y中的有关X的信息量

互信息

- 定义5: 互信息，随机变量*X*,*Y*之间的互信息定义为:

$$I(X;Y)=\sum_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$



- 互信息的性质:
  - $I(X;Y)\geqslant 0$ 等号成立当且仅当*X*和*Y*相互独立。
  - $I(X;Y) = I(Y;X)$ 说明互信息是对称的。

点间互信息(pointwise mutual information)

- 在计算语言学中，更为常用的是两个具体事件之间的互信息，一般称之为点间互信息。
- 定义6: 点间互信息，事件x, y之间的互信息定义为:

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

- 点间互信息度量两个具体事件之间的相关程度
  - 当 $I(x,y)>>0$ 时，x和y高度相关。
  - 当 $I(x,y) = 0$ 时，x和y高度相互独立。
  - 当 $I(x,y)<<0$ 时，x和y呈互补分布。

相对熵(relative entropy)

- 定义7: 相对熵，设$p(x)$、$q(x)$是随机变量X的两个不同的分布密度，则它们的相对熵定义为：

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

- 相对熵一般也称谓*Kullback-Leibler* 发散度或*Kullback-Leibler* 距离。
- 度量同一个随机变量的不同分布的差异。
- 相对熵描述了因为错用分布密度而增加的信息量。

交叉熵(cross entropy)

- 定义**8**: 交叉熵，设随机变量X的分布密度为$p(x)$，在很多情况下$p(x)$是未知的，人们通常使用通过统计手段得到的X的近似分布$q(x)$，则随机变量X的交叉熵定义为：

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

语言模型评价

- 语言L的交叉熵

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

- 如果假定语言是稳态具各态遍历性质的随机过程，则可以作如下计算：

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\log m(x_{1n})$$

- 如果 n 足够大，则

$$H(L,m) \approx -\frac{1}{n}\log m(x_{1n})$$

语言模型评价

- 令T为测试语料
  $T = (t_1 t_2 \ldots t_n)$
- 测试语料的概率

$$p(T) = \prod_{i=1}^{n} p(t_i)$$

- 利用测试语料计算交叉熵

$$H_p(T) = -\frac{1}{W_T} log_2\, p(T)$$

测试语料中的词数

- 一般而言，交叉熵越小，模型性能会越好。

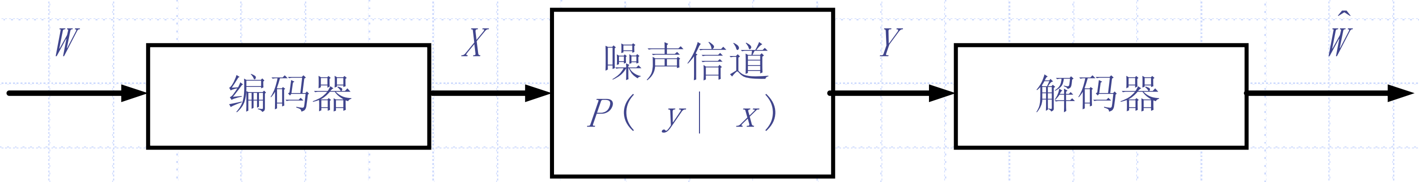
困惑度(perplexity)

- 语言模型的评价也可以计算困惑度，困惑度定义如下：

$$PP_p(T) = 2^{H_p(T)}$$

- 同交叉熵的度量结果没有区别
  交叉熵 9.9 →9.1
  困惑度 950→540

噪音信道模型

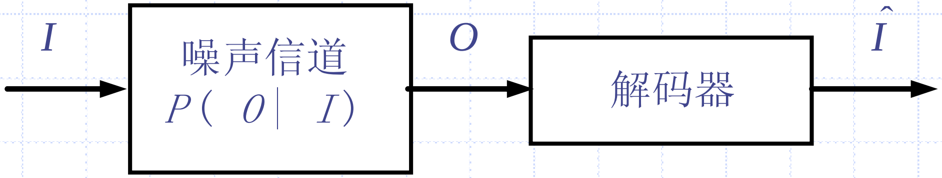


W是欲经信道传输的消息，在传输之前，首先进行编码使其适于信道传输，编码后的消息为X，由于信道噪声的存在，在信道末端，人们并不能精确接收到X，而是接收到有噪声在内的编码Y，信道概率$p(y|x)$描述了编码x因噪声而变成y的概率，当接收方接到含有噪声的编码后，其任务就变为将Y解码，得到最为可能的消息$\hat{W}$。

作为通信系统而言，人们最为关心的是，如何将消息编码，以便消息在有噪声存在的情况下有效可靠地发送到接收方。

噪音信道模型

- 在利用噪声信道处理语言问题时，人们并不关心编码问题，而更多关心的是，在有噪声存在的情况下，如何解码将输出还原为信道输入。



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

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

$$\hat{I} = \arg\max_{I} p(I)\,p(O \mid I)$$

语言模型

信道模型

噪音信道模型的应用

- 机器翻译

$$\hat{T} = \arg\max_{T} p(T)p(S|T)$$

- 利用信道模型，人们为翻译问题找出了一个整齐的数学描述。
- 词性标注 音字转换 字音转换等