数学基础

- 概率论基础
- 信息论基础
- 应用案例

概率论基础

1. 最大似然估计：

> 如果一个实验的样本空间是$\{S_1, S_2, \cdots, S_n\}$，在相同的情况下重复试验$N$次，观察到样本$s_k(1 \le k \le n)$的次数为$n_k(s_k)$，则$s_k$的相对概率为：
>
> $$q_N(s_k) = \frac{n_N(s_k)}{N}$$
>
> 由于$\sum_{k=1}^{n} n_N(s_k) = N$，因此$\sum_{k=1}^{n} q_N(s_k) = 1$。
> 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$P(s_k)$。事实上：
>
> $$\lim_{N \to \infty} q_N(s_k) = P(s_k)$$
>
> 因此，相对频率常用作概率的估计值。这种概率值的估计方法称为**最大似然估计**。

2. 贝叶斯决策理论：

> 假设研究的分类问题有c个类别，各类别的状态用$w_i(i = 1, 2, \cdots, c)$表示，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$\overline{x} = [x_1, x_2, \ldots, x_d]^T$是$d$维特征空间上的某一点，且条件概率密度函数$P(x|w_i)$是已知的。那么，利用贝叶斯公式可以得到后验概率：
>
> $$P(w_i|\overline{x}) = \frac{P(\overline{x}|w_i)P(w_i)}{\sum_{j=1}^{n} P(\overline{x}|w_j)P(w_j)}$$
>
> **基于最小错误律的贝叶斯决策规则**为：
>
> （1）如果$P(w_i|\overline{x}) = \max_{j=1,2,\cdots,c} P(w_j|\overline{x})$，则$\overline{x} \in w_i$；
>
> （2）或：如果$P(\overline{x}|w_i)P(w_i) = \max_{j=1,2,\cdots,c} P(w_j|\overline{x})P(w_j)$，则$\overline{x} \in w_i$；
>
> （3）或（$c = 2$）：如果$l(\overline{x}) = \frac{P(\overline{x}|w_1)}{P(\overline{x}|w_2)} > \frac{P(w_2)}{P(w_1)}$，则$\overline{x} \in w_1$，否则$\overline{x} \in w_2$。
>
> **贝叶斯决策理论在文本分类、词汇语义消歧等问题的研究中具有重要用途**。

3. 二项式分布：

> **在自然语言处理中，一般以句子为处理单位。假设一个句子独立于它前面的其他语句，句子的概率分布近似地认为符合二项式分布**。

信息论基础

1. 熵：

> 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：

> $$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$
>
> 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比特位（bit）。
>
> 熵又称为自信息（self-information），表示信源X每发一个符号（不论发什么符号）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需要越大的信息量用以确定其值。

2. 联合熵:

> 如果X, Y是一对离散型随机变量$(X, Y) \sim p(x, y)$，则X, Y的联合熵$H(X, Y)$为:
>
> $$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$$
>
> **联合熵实际上就是描述一对随机变量平均所需信息量的数量**。

3. 条件熵:

> 给定随机变量X的情况下，随机变量Y的条件熵定义为:
>
> $$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) H(Y|X = x) \\ &= \sum_{x \in X} p(x) [-\sum_{y \in Y} p(y|x) \log_2 p(y|x)] \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x) \end{aligned}$$
>
> - **条件熵与联合熵的关系**: $H(X, Y) = H(X) + H(Y|X)$（**连锁规则**）。
>   - tips:
>
> $$\begin{aligned} H(X, Y) &= -\sum_{x \in X} \sum y \in Y p(x, y) \cdot \log[p(x) p(y|x)] \\ &= -\sum_{x \in X} \sum y \in Y p(x, y) \cdot [\log p(x) + \log p(y|x)] \end{aligned}$$
>
> - $H(X|Y) \neq H(Y|X)$。

- 条件熵$H(Y|X)$用于衡量已知X的情况下，还需多少信息能够确定Y。

4. 熵率:

> 一般地，对于一条长度为n的信息，每一个字符或字的**熵率**为:
>
> $$H_{\text{rate}} = \frac{1}{n} \cdot H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})$$
>
> 其中，变量X_{1n}表示随机变量序列$(X_1, \cdots, X_n)$。$x_{1n} = x_1, \cdots, x_n$，亦作: $x_1^n = (x_1, \cdots, x_n)$。

5. 相对熵（或Kullback-Leibler divergence，KL距离）:

> 两个概率分布$p(x), q(x)$的**相对熵**为:
>
> $$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$
>
> 约定$0 \log(0/q) = 0$及$p \log(p/0) = \infty$。

> 相对熵常被用以衡量两个随机分布的差距。当两个随机变量相同时，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。

6. 交叉熵：

> 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，则随机变量X和**模型**q之间的交叉熵为：
>
> $$H(X,q) = H(X) + D(p||q) = -\sum_x p(x)\log q(x)$$
>
> **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**。
>
> 语言$L = (X_i) \sim p(x)$与其模型q的交叉熵为：
>
> $$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n} p(x_1^n)\log q(x_1^n)$$
>
> 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。
>
> 假设这一语言是“理想”的，即$n \to \infty$时，其全部“单词”的概率之和为1。则进一步有以下定理：
>
> > 假定语言L是**稳态**随机过程，x_1^n为L的样本，L与其模型q的交叉熵为：
> >
> > $$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n} p(x_1^n)\log q(x_1^n)$$
> > $$\Downarrow$$
> > $$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n)$$
>
> 由此，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$p(x)$。

7. 困惑度：

> 给定语言L的样本$l_1^n = (l_1, \cdots, l_n)$，$L$的**困惑度**$PP_q$为：
>
> $$PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n}\log q(l_1^n)} = [q(l_1^n)]^{-\frac{1}{n}}$$
>
> 在设计语言模型时，通常用困惑度来代替交叉熵衡量语言模型的好坏。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。

8. 互信息：

> 如果$(X,Y) \sim p(x,y)$，则X,Y之间的互信息$I(X;Y)$定义为：
>
> $$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$
>
> **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**。
>
> - 互信息、条件熵与联合熵的关系：
>
> $$\begin{aligned} H(X,Y) &= H(X|Y) + I(X;Y) + H(Y|X) \\ &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

> - 由于$H(X|X)=0$，所以：
>
> $$H(X)=H(X)-H(X|X)=I(X;X)$$
>
> 一方面说明了为什么熵又称**自信息**；另一方面说明了，**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**。
>
> - 汉语分词应用：
>   - 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，也可能成词。反之，断开的可能性越大**。
>   - 相邻词的两个词间字的互信息，应当小于词内部相邻字之间的互信息。
>   - 当两个汉字x,y的关联度较强时，其互信息$I(x;y)>0$；关联度较弱时，$I(x;y)\approx 0$；而当$I(x;y)<0$时，x,y称为**互补分布**。

9. 双字耦合度：

> 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：
>
> > 设c_i,c_{i+1}是两个连续出现的汉字，统计样本中c_i,c_{i+1}连续出现在一个词中的次数，和连续出现的总次数之比为二者的双字耦合度，即：
> >
> > $$\mathrm{couple}(c_i,c_{i+1})=\frac{N(c_i,c_{i+1})}{N(c_i,c_{i+1})+N(\cdots c_i|c_{i+1}\cdots)}$$
> >
> > 其中c_i,c_{i+1}是有序字对（即$c_ic_{i+1}\neq c_{i+1}c_i$）。$N(c_i,c_{i+1}$表示字符串$c_ic_{i+1}$构成词的频率；$N(\cdots c_i|c_{i+1}\cdots)$表示$c_i$作为上一词词尾，且$c_{i+1}$作为相邻下一词词头的频率。
>
> **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适合一些**。

10. 噪声信道模型：

> 二进制对称信道：
>
> - 过程：初始输入信号 -> 编码器 -> 噪声信道 -> 解码器 -> 根据输出尽量恢复初始输入信号。
> - 输入符号集$X:\{0,1\}$；输出符号集$Y:\{0,1\}$。设传输过程中输入符号被误传的概率为p。
> - **信道容量**：降低传输速率来换取高保真通讯的可能性，其可根据互信息定义为：
>
> $$C=\max_{p(x)}I(X;Y)$$
>
>   即设计一个输入编码X，其概率分布为$X\sim p(x)$。使输入与输出之间的互信息达到最大值，那么久达到了信道的最大传输容量。
>
> 自然语言处理：
>
> - 无需编码，只需进行解码，使系统输出更接近于输入。
> - 例如法译英的翻译过程：英语句子e -> 噪声信道模型 -> 法语句子f。
> - 原理：使贝叶斯概率最大，即：

> $$\hat{e} = \argmax_e p(e) \times p(f|e)$$
>
> - 统计翻译系统框架：e -> 语言模型$p(e)$ -> 翻译模型$p(f|e)$ -> 解码器 -> $\hat{e}$。

应用案例

1. 消歧方法：（1）**基于上下文分类的消歧方法**；（2）**基于最大熵的消歧方法**。
2. 基于上下文分类的消歧方法：（基于贝叶斯分类器的基本思路）

> 假设某多义词w所处的上下文语境为C，若w的多个语义记为$s_i(i \geq 2)$，则可通过：
>
> $$\argmax_{s_i} p(s_i|C)$$
>
> 来确定w的词义。由贝叶斯公式：$p(s_i|C) = \dfrac{p(s_i) \times p(C|s_i)}{p(C)}$。考虑分母不变性，并运用如下**独立性假设**：
>
> $$p(C|s_i) = \prod_{v \in C} p(v|s_i)$$
>
> 因此：
>
> $$\hat{s} = \argmax_{s_i} [p(s_i) \prod_{v \in C} p(v|s_i)]$$
>
> 其中概率$p(v|s_i)$和$p(s_i)$都可用最大似然估计求得：
>
> $$p(v|s_i) = \frac{N(v, s_i)}{N(s_i)}, p(s_i) = \frac{N(s_i)}{N(w)}$$
>
> 其中$N(s_i)$是词w用于语义s_i时的次数；$N(v, s_i)$为w用于语义s_i时，词v出现在w的上下文的次数；$N(w)$为多义词w出现的总次数。
>
> 在实际算法中，通常将概率$p(w|s_i), p(s_i)$的城际运算转换为**对数加法运算**：
>
> $$\hat{s} = \argmax_{s_i} \left[\log p(s_i) + \sum_{v \in C} \log p(v|s_i) \right]$$

3. 基于最大熵的消歧方法：
   - 基本思想：**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**。