计算机自然语言处理

数学基础

初等概率理论

基于统计的语言处理技术已经成为语言处理技术的主流。统计语言处理的目的在于以自然语言为处理对象进行统计推导。包括两个步骤

1. 收集自然语言词汇（或者其他语言单位）的分布情况，即统计语言单位出现的概率。
2. 根据这些分布情况进行统计推导。

基本概念

概率论（probability theory） 是研究随机现象的数学分支。随机现象的实现和对它的观察成为随机试验（random experiment / trail）。随机试验的每一个可能结果称为一个基本事件（elementary event）一个或一组基本事件又统称为随机事件（random event），简称为事件（event）。事件的概率（probability）是衡量该事件发生的可能性的度量。在大量重复试验或观察中所呈现的固有规律性，称为随机现象的统计规律性（statistical rule）.

全体基本事件构成的集合称为样本空间（sample space）, 记做Ω。随机变量（random variable）X 是定义于Ω上的函数。随机变量的取值可以是离散的(discrete random variable)，也可以是连续的（continuous random variable）。

在语言处理中，通常将语言单位视为一个离散型的随机变量。

概率的统计定义：

频率：描述事件出现的频繁程度。

若事件A在相同条件下进行的n次试验中出现了r次，则称

$W_n(A) = \frac{r}{n}$

为事件A在n次试验中出现的频率（frequency），称r为事件A在n次试验中出现的频数（frequence）。

如果随着试验次数n的增大，事件A出现的频率$W_n(A)$总在区间$[0,1]$上的某个数字p附近摆动，那么定义事件A的概率为

$P(A) = p$

称为概率的统计定义（statistical definition of probability），由此确定的概率称为统计概率（statistical probability）。

除统计概率之外，还有古典概率，几何概率等多种定义方式。

通过公理化体系，给出概率的数学定义：

定义　设随机试验的样本空间为Ω，如果对于每一个事件$A \subset \Omega$，总有一实数$P(A)$与之对应，此实值函数$P(A)$满足如下公理：

公理１　（非负性）对于任一事件A，有

$0 \leq P(A) \leq 1$

公理２　（规范性）

$P(\Omega)=1, P(\Phi)=0$

公理３　（完全可加性）对于任意有限个两两互斥的事件$A_1, A_2, \ldots A_n$，有

$P(A_1 \cup A_2 \cup \ldots \cup A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$

则称$P(A)$为事件A的概率，这个定义称为概率的公理化定义（axiomatic definition of probability）。公理３又称概率的加法定理（addition formula）

推论：

1. $P(B-A)=P(B)-P(A \cap B)$
2. $A \subseteq B \Rightarrow P(A) \leq P(B)$
3. $P(\overline{A})=1-P(A)$
4. $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

   $B-A$ 差事件（differential event）

   $A \cap B$ 积事件（product event）

   $A \subseteq B$ 子事件（sub-event）

   $\overline{A}$ 逆事件（inverse event）

   $A \cup B$ 和事件（additive event）

条件概率与独立

$P(A|B)$：　事件Ｂ出现条件下事件Ａ的条件概率（conditional probability）

先验概率（prior probability）

后验概率（posterior probability）

在事件Ｂ已知出现的情况下，事件Ａ出现的条件概率

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

由上式可得概率的乘法定理（production rule）

$P(A \cap B)=P(B)P(A|B)=P(A)P(B|A)$

链规则(chain rule)

$P(A_1 \cap A_2 \cap \ldots \cap A_n=P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots P(A_n| \cap_{i=1}^{n-1} A_i)$

链规则在统计自然语言处理技术中有着广泛应用，是构造统计语言模型的理论基础之一。

两个事件Ａ和Ｂ是相互独立的（independent）,当且仅当

$P(A \cap B)=P(A)P(B)$

否则，称事件Ａ和Ｂ是相互依赖的（dependent）

类似地，称事件Ａ和事件Ｂ在事件Ｃ发生的条件下相互独立，当且仅当

$P(A \cap B|C)=P(A|C)P(B|C)$

全概率公式和贝叶斯公式

定义　满足如下条件的一组事件$B_1, B_2, \ldots B_n$，称为样本空间Ω的一个划分（partition）：

1. $B_i B_j = \phi, i \neq j; i, j = 1, 2, \ldots, n$
2. $B_i \cup B_2 \cup \ldots \cup B_n = \Omega$

定理　如果$B_1, B_2, \ldots B_n$构成样本空间Ω的一个划分，且$P(B_i) > 0 (i = 1, 2, \ldots, n)$,则对任一事件A，有

$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n)$

$= \sum_{i=1}^{n} P(B_i) \cdot P(A|B_i)$

称为全概率公式（breakdown law）

贝叶斯公式（Bayesian formula）在统计语言处理中占据举足轻重的地位。

由条件概率公式得

$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$

$arg_B max P(B|A) = arg_B max \frac{P(A|B)P(B)}{P(A)} = arg_B max P(A|B)P(B)$

贝叶斯定理　如果存在一组事件B_i是事件A的划分，$A \subseteq B_1 \cup B_2 \cup \ldots \cup B_n$，且当$i \neq j$时，$B_i \cap B_j = \phi$,于是，

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

数学期望与方差

数学期望（expectation）和方差（variance）是随机变量的两个重要的数值特征，数学期望反映了随机变量的平均取值，方差反映了随机变量取值的分散程度。

常用分布

二项分布

离散型随机变量只有两个取值

伯努利试验（Bernoulli trials）

泊松分布

离散型随机变量的取值范围为$k = 0, 1, 2, \ldots$

正态分布

连续型随机变量最重要的概率分布是正态分布，又称高斯分布（Gaussian distribution）