论文

第三种贝叶斯定理用于语义通讯和统计学习

鲁晨光

Email： lcguang@foxmail.com

摘要 第一种贝叶斯定理是贝叶斯提出的，描述两个逻辑概率之间的对称关系；第二种贝叶斯定理是 Shannon 信息论中使用的，描述两个统计概率之间的对称关系. 本研究提出第三种贝叶斯定理，它描述一个统计概率和一个逻辑概率之间的不对称关系，据此可以简单地从真值函数(反映语义)得到似然函数，也可以简单地从似然函数或条件样本分布得到真值函数. 语义信息测度被定义为对数标准似然度. 一组真值函数构成一个语义信道，它反映样本的多类别多标签逻辑分类，可以用作预测模型. 用样本训练预测模型就是让语义信道匹配 Shannon 信道；用语义信息准则或似然准则选择假设或标签就是让 Shannon 信道匹配语义信道. 两种信道相互匹配和迭代可以更方便实现最大似然分类、检验、估计和混合模型.

关键词 贝叶斯定理，语义信息，真值函数，似然函数，多标签分类，检验，估计，混合模型

1 引言

频率主义和贝叶斯主义之间的矛盾[1]涉及许多领域，比如假设检验、语义通信和统计学习. 频率主义认为概率就是事件发生的频率或频率的极限，是客观的. 贝叶斯主义认为概率是主观预期的频率或对事件或假设的相信程度. Shannon 信息论[2]显示了频率主义的巨大成功. 然而，如果不考虑语句的逻辑概率或真值，信息理论就不能度量语义信息[3-6]，比如度量自然语言、各种经济指标和各种仪表(包括时钟和 GPS)提供的信息. 对于统计学习，我们需要考虑假设检验、语句的真假，并根据语义预测，因而也需要考虑逻辑概率.

频率主义根据一个假设 y_j 和预测模型 θ 预测 X 的概率分布，即产生似然函数 $P(X|y_j, \theta)$(其中 X 是随机变量，y_j 是随机变量 Y 的一个取值). 然而，当 X 的先验概率 $P(X)$发生变化, 比如变为 $P'(X)$ 时，旧的似然函数就会失效. 所以，我们需要概率分布函数 $P(y_j, \theta|X)$(y_j 不变 X 变)，使得收信人收到 y_j 时，可以根据贝叶斯定理，从 $P(y_j, \theta|X)$和 $P'(X)$得到新的似然函数 $P'(X|y_j, \theta)$(后面称之为贝叶斯预测). 根据流行的贝叶斯(主义)推理(Bayesian Inference)[7], 虽然我们可以得到 $P(y_j, \theta|X)$，然而，当样本足够大时，通过这样的 $P(y_j, \theta|X)$和 $P'(X)$得到的预测 $P'(X|y_j, \theta)$和通过 $P(y_j|X)$和 $P'(X)$得到的预测 $P'(X|y_j)$（根据贝叶斯定理 2）是不相等的. 如果不相等，流行的贝叶斯推理就和贝叶斯定理不兼容.

然而，贝叶斯定理本身也有局限性，需要推广或补充. 为此，我们需要重新审视频率主义和贝叶斯主义之争和贝叶斯定理. 本研究争论说，概率的客观频率解释和主观期望解释都是需要的；实际存在三种概率和三种贝叶斯定理：

1)逻辑概率(LP)(参看 Jaynes 对概率的解释[1])，即一个假设为真的概率，它等价于一个随机变量 X 在一个集合中的概率. 相应的是第一种贝叶斯定理，即贝叶斯提出的定理[8]，简称贝叶斯定理 1， 其中两个概率是一个随机变量 X 出现在一个论域的两个集合中的概率.

2)统计概率(SP)，即频率(或频率的极限)，是一个随机变量 X 取某个值 x_i 的概率. 相应的贝叶斯定理如 Shannon 信息论中用到贝叶斯公式所示[2]，简称贝叶斯定理 2，其中两个概率是两个随机变量 X 和 Y 分别取值于两个论域中两个元素 x_i 和 y_j 的概率.

3)主观预测的概率(PP，即 Predictive Probability)，比如似然度. 它是逻辑概率和统计概率的杂交. 相应的贝叶斯定理我们称之为贝叶斯定理 3，其中两个概率是一个统计概率和一个逻辑概率. 我们需要解决: a)已知 X 在某集合 A 中和 $P(X)$，求 X 被预测的概率分布 $P(X|X\in A)$; b)已知预测的概率分布和 $P(X)$，求集合 A 的特征函数或集合模糊时的隶属函数[9]——它也就是逻辑分类函数.

看来可以说，频率学派理解的概率包括 SP 和 PP，而贝叶斯学派理解的概率包含 LP 和 PP.

当我们把 X 取值的论域划分成 n 个子集，则 X 在 n 个子集中的概率(逻辑概率)P_j, j=1, 2, …, n 就和统计概率一样，是归一化的. 这时候 P_j 也可以解释为统计概率. 但是更一般情况下，比如天气预报，可选择预报并不是相互排斥的(比如包含“明天无雨”,“明天有雨”,“明天有小雨”和“明天有小到中雨”)，使预报为真的若干降水量集合构成论域的一个覆盖，而不是一个划分. 这时候，已有的两种贝叶斯定理就不够用了.

一般情况下，逻辑概率不是归一化的；而统计概率是归一化的. 我们需要严格区分两者. 比如对于天气预报，y_1=”明天有雨”， y_1 有被选择的概率(SP)， 也有为真的概率(LP). 两者是不同的. 考虑预报“明天没有暴雨”，其逻辑概率很大， 而被选择的概率很小. 一个永真句“明天可能有雨也可能没雨”，它的逻辑概率是 1， 但是被选择的概率可能是 0. 笔者以为， 流行的贝叶斯(主义)推理之所以存在种种问题，主要原因是没有很好区分两种概率.

作者之所以提出第三种贝叶斯推理，是因为在过去的语义信息和统计学习研究中[6, 10-13]发现了相关公式. 现在提出第三种贝叶斯定理，是把特殊推广到一般，希望能大大提高统计学习的效率和可靠性.

本文首先介绍第三种贝叶斯定理及其应用于多类别多标签逻辑分类(类别是清晰的)，继而把它推广到类别模糊的场合；然后再介绍，如何用对数标准似然度(normalized likelihood)定义语义信息(即广义信息[6])；讨论在样本并不足够大时，如何使用语义信息准则优化真值函数，把语义信息方法——特别是语义信道和 Shannon 信道相互匹配的算法(即 CM 算法)——应用于统计学习.

2 推广贝叶斯定理

2.1 统计概率和逻辑概率的定义和性质

定义 2.1.1 设论域 U 中有元素 x_1, x_2, …, x_m; X 是取值于 U 中某个元素的随机变量，即 $X\in U=\{x_1, x_2, \ldots, x_m\}$. 再设论域 V 中有元素 y_1, y_2, …, y_n; Y 是取值于 V 中某个元素的随机变量，即 $Y\in V=\{y_1, y_2, \ldots, y_n\}$. 对于每个假设 y_j, 存在一个集合 $A_j\in 2^U$, y_j=“$X\in A_j$”.

定义 2.1.2 用等号“=”表示的随机事件的概率——比如 $P(X=x_i)$——是统计概率，后面简写为 $P(x_i)$; 如果 X 的值没有给定，我们用 $P(X)$表示 $P(X$=any)(any 意指 U 中任一元素). 用属于符号“∈”表示的随机事件的概率——比如 $P(X\in A_j)$——是逻辑概率，后面简记为 $P(A_j)$或 $T(A_j)$.

我们把 $P(X\in A_j)$称之为逻辑概率，是因为根据 Tarski 的真理论[14]，$P(X\in A_j)=P(\text{“}X\in A_j\text{”}$是真的)=$P(y_j$ 是真的). 于是，一个假设 y_j 有两种概率: 统计概率 $P(y_j)$和逻辑概率 $P(y_j$ 是真的)=$P(A_j)$. 为了更清楚区分两者，后面我们用 $T(A_j)$表示 y_j 的逻辑概率，即

$$T(A_j)= P(y_j \text{是真的})= P(X\in A_j) \tag{2.1}$$

以 X 为条件的 y_j 的逻辑概率就是集合 A_j 的特征函数或 y_j 的真值函数，记为 $T(A_j|X)$， 于是

$$T(A_j)=\sum_i P(x_i)T(A_j\mid x_i) \tag{2.2}$$

根据 Davidson 的真值条件语义学[15]，上述真值函数确定了假设 y_j 的语义.

统计概率分布——比如 $P(Y)$和 $P(Y|x_i)$——是归一化的, 即

$$P(y_1)+P(y_2)+\ldots+P(y_n)=1，\ P(y_1|x_i)+P(y_2|x_i)+\ldots+P(y_n|x_i)=1 \tag{2.3}$$

而逻辑概率不是归一化的，比如在$\{A_1, A_2, \ldots, A_n\}$是 U 的一个覆盖的情况下，

$$T(A_1)+T(A_2)+\ldots+T(A_n)\geqslant 1 \tag{2.4}$$

只有在$\{A_1, A_2, \ldots, A_n\}$是 U 的划分并且 Y 的使用总是正确的情况下，两种概率才相等.

注意：$P(y_j|X)$和 $P(Y|x_i)$不同，$P(y_j|X)$ (y_j 不变而 X 变)可谓转移概率函数，可用作贝叶斯预测，产生 X 的后验概率分布 $P(X|\text{yj})$. 它也不是归一化的，即一般情况下

$$P(y_j|x_1)+P(y_j|x_2)+\ldots+P(y_j|x_m)\neq 1 \tag{2.5}$$

而真值函数(即集合的特征函数或隶属函数)$T(A_j|X)$的最大值是 1(也考虑到集合是模糊的情况). 可以说统计概率分布是横向归一化的，而真值函数是纵向归一化的. 即：

$$\max(T(A_j|x_1), T(A_j|x_2), \ldots, T(A_j|x_m))=1 \tag{2.6}$$

这一重要性质将给求解真值函数带来方便. 注意：逻辑概率分布——即后面的 $T(\theta)$，很像是流行的贝叶斯推理中的 $P(\theta)$[7]——和转移概率函数 $P(y_j|X)$一样，既不是横向归一化的也不是纵向归一化的，所以求解两者比较困难. 流行的贝叶斯推理就似乎遇到这样的困难[7].

2.2 三种贝叶斯定理

贝叶斯定理 1：设集合 $A, B\in 2^U$. A’是 A 的补集，B’是 B 的补集. $T(A)=P(X\in A)$, $T(B)$等同理. 则：

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)= T(A|B)T(B)+ T(A|B')T(B') \tag{2.7}$$

$$T(A|B)=T(B|A)T(A)/T(B),\ T(B)= T(B|A)T(A)+ T(B|A')T(A') \tag{2.8}$$

如果$\{B_1, B_2, \ldots, B_K\}$构成 U 的一个划分，则

$$T(A)=\sum_{k=1}^{K} T(A\mid B_k)T(B_k) \tag{2.9}$$

贝叶斯定理 2：设事件是 $X=x_i$ 和 $Y=y_j$. $P(x_i)=P(X=x_i)$， $P(y_j)$等同理. 则

$$P(x_i\mid y_j)=P(x_i)P(y_j\mid x_i)/P(y_j),\ P(y_j)=\sum_i P(x_i)P(y_j\mid x_i) \tag{2.10}$$

$$P(y_j\mid x_i)=P(y_j)P(x_i\mid y_j)/P(x_i),\ P(x_i)=\sum_j P(y_j)P(x_j\mid y_j) \tag{2.11}$$

上面 x_i 换成变量 X， 公式同样成立， 即：

$$P(X|y_j)=P(y_j|X)P(X)/P(y_j) \tag{2.12}$$

$$P(y_j|X)=P(X|y_j)P(y_j)/P(X) \tag{2.13}$$

其中 $P(X|y_j)$是 X 的后验概率分布，是归一化的，可以表达一个样本的条件概率分布. 而 $P(y_j|X)$是 y_j 的转移概率函数，n 个转移概率函数构成一个 Shannon 信道(后面谈及). 上面每个定理中的两个公式是对称的，分母都是归一化常数.

贝叶斯定理 3：设两个事件是 X=any 和 $P(X\in A)$，$P(X)=P(X=\text{any})$, $T(A_j)=P(X\in A_j)$, 则

$$P(X|A_j)=P(X)T(A_j|X)/T(A_j),\ T(A_j)=\sum_i P(x_i)T(A_j|x_i) \tag{2.15}$$

$$\begin{aligned} &T(A_j|X)=T(A_j)P(X|A_j)/P(X),\ T(A_j)=1/[P(x_j^*|A_j)/P(x_j^*)] \\ &x_j^*=\arg\max_{x_i}[P(x_i|A_j)/P(x_i)] \end{aligned} \tag{2.16}$$

解释：(2.15)中 $T(A_j)$是 $P(X|A_j)$的横向归一化系数；而在(2.16)中，$T(A_j)$是 $T(A_j|X)$的纵向归一化系数，它使 $T(A_j|X)$的最大值等于 1；x_j^*是函数 $P(X|A_j)/P(X)$曲线最高点(或多个最高点之一)下面的 x_i.

如果仿照**贝叶斯定理 1 和 2**， (2.16)中应有

$$P(X)=\sum_j T(A_j)P(X|A_j)$$

然而，逻辑概率不是归一化的，用上式求出的 $P(X)$也不会是归一化的，即 $P(x_1)+P(x_2)+\ldots+P(x_m)>1$. 所以上式是不成立的. 贝叶斯定理 3 中两个公式是不对称的，这是因为 $P(X|A_j)$是横向归一化的，而 $T(A_j|X)$是纵向归一化的.

贝叶斯定理 3 证明：设联合概率 $P(X=\text{any}，X\in A_j)$ (X=any 就是 X 出现)，则

$$P(X=\text{any}, X\in A_j)= P(X=\text{any}|\ X\in A_j)P(X\in A_j)= P(X|A_j)T(A_j)$$

$$P(X=\text{any}, X\in A_j\ )=P(X\in A_j\ |\ X=\text{any})P(X=\text{any})= T(A_j|X)P(X)$$

于是有

$$P(X|A_j)=P(X)T(A_j|X)/T(A_j),\ T(A_j|X)=T(A_j)P(X|A_j)/P(X)$$

因为 $P(X|A_j)$是横向归一化的，所以 $T(A_j)=\sum_i P(x_i)\ T(A_j|x_i)$. 因为 $T(A_j|X)$是纵向归一化的，把(2.6)代入上式，可以得到

$$1= T(A_j)P(x_j^*|A_j)/P(x_j^*)，\text{即 } T(A_j)=1/[P(x_j^*|A_j)/P(x_j^*)]$$

证毕.

所以**贝叶斯定理 3** 的第二个公式也可以直接写成：

$$T(A_j|X)=[P(X|A_j)/P(X)]/[P(x_j^*|A_j)/P(x_j^*)] \tag{2.17}$$

图 1 直观显示了 $T(A_j|X)$，$P(X|A_j)$和 $P(X)$三者之间的关系.

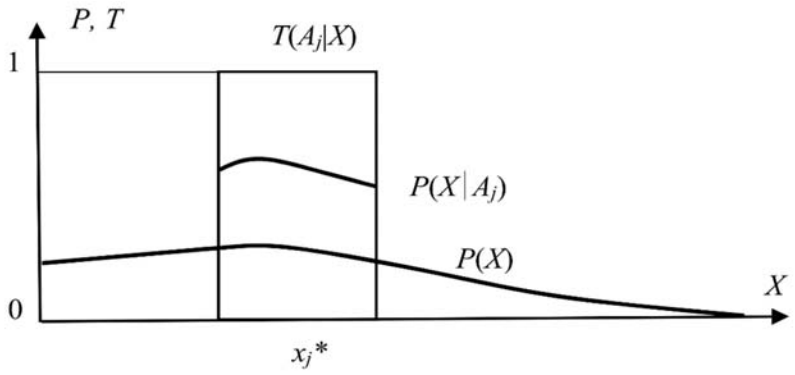


图 1 贝叶斯定理 3 中 $T(A_j|X)$，$P(X|A_j)$和 $P(X)$之间的关系

Figure 1 Relationships between $T(A_j|X)$，$P(X|A_j)$ and $P(X)$ in Bayes' Theorem 3

可以这样产生概率分布等于 $P(X|A_j)$样本：按 $P(X)$产生一个样本，如果一个样本点落在 A_j 中就保留；保留的样本点在 U 上的概率分布就是 $P(X|A_j)$. 如果样本足够大, 由 $P(X|A_j)$和 $P(X)$就可以求出 A_j 的真值函数.

2.3 多类别多标签分类——集合清晰时的逻辑分类和选择分类

设 U 包含不同年龄，U 上有集合 A_1={未成年人}={$X|X$<18}, A_2={成年人}={$X|X$≥18}, A_3={年轻人}={$X|14\leq X<30$}，它们构成 U 的一个覆盖. 三个真值函数 $T(A_1|X)$，$T(A_2|X)$，$T(A_3|X)$分别反映了 y_1, y_2, y_3 的语义，参看图 2.

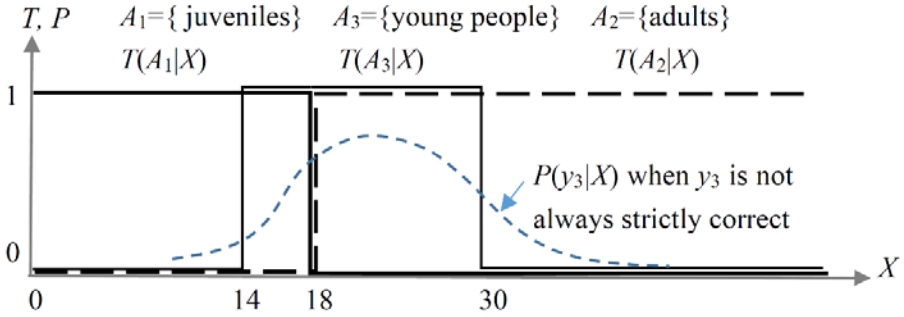


图 2 年龄上的三个集合构成 U的一个覆盖，反映三个谓词 y_1, y_2, y_3 的语义

Figure 2. Three sets form a cover of U, indicating semantic meanings of y_1, y_2, and y_3.

在这个例子中， $T(A_1)+T(A_2)=1$, 设 $T(A_3)=0.3$，则逻辑概率之和是 1.3. 而 Y 的统计概率是归一化的，之和为 1. $P(y_3)$可大可小. 比如 $P(y_3)=0.2, P(y_1)=0.3, P(y_2)=0.5$,.下面考虑如何从联合概率分布 $P(X,Y)$得到 A_1，A_2，A_3 的特征函数或 y_1, y_2, y_3 的真值函数.

对于分类，Y 表示标签，X 表示样本点或实例，一个数据对$(x(t), y(t))$(t 表示数据对序号)或(x_i, y_j)表示一个样例. 从语义通信的角度看，收信人的分类和发信人的分类是不同的，可选择标签集合也是不同的。收信人从以往带标签样本中学习得到不同 Y 的语义即真值函数，然后根据每次收

到的 y_j 的语义和 $P(X)$得到概率预测 $P(X|A_j)$(即理解)，以便决策. 假设原子句构成复合句，收信人训练模型只需要训练 n 个原子句的真值函数，而发信人最多可使用 2^n 个复合句。 比如对于图 2 例子，训练模型只要求出 y_1 和 y_3 的真值函数(n=2)，因为 $T(A_2|X)$=1-$T(A_1|X)$; 而发信人使用的标签最多可以是 2^2=4 个。发信人每次尽可能选择信息量大或内涵丰富的 y_j. 在 y_j 正确的前提下，逻辑概率越小越好. 比如对于 17 岁，选择复合句 y_4=y_1 且 y_3 = "X 是不到 18 岁的年轻人" 信息量最大. 只有在不很确定哪个 X 发生时，发信人才选择语义比较模糊或外延较大的 y_j.

多标签分类已有很多研究成果[16]. 上面分类方法和已有研究中把多类别分类化为多个二元分类是类似的[17, 18]. 不同之处主要是：

1） 逻辑分类函数就是真值函数，反映语义，不因 $P(X)$变化而变化；而已有的分类方法没有求逻辑分类函数，而是直接求发信人的分类，分类函数不能很好反映语义。

2） 发信人选择分类和 $P(X)$有关（参看后面(2.21)和(3.14)），用真值函数产生似然函数就已经考虑了类别不平衡 [26]; 而用已有的分类方法，$P(X)$变化后，旧的模型就会失效或不理想。

多标签逻辑分类使用的样例可以是单标签的，如果样例是多标签的，我们可以采用已有的 First-order-strategy [17, 19]把它分开，比如$(x_1; y_1, y_2)$可以分成(x_1, y_1)和(x_1, y_2). 对于多实例单标签样例也做分拆处理[20]，比如把$(x_1, x_2; y_1)$分拆成(x_1, y_1)和(x_2, y_1). 然后通过统计得到联合概率分布 $P(X, Y)$. 有了 $P(X,Y)$，就可以得到多标签逻辑分类. 另外，流行的多标签分类可能是因为 X 是多维的；而这里，即使 X 是一维的，也存在多标签逻辑分类(参看图 2). 后面我们只考虑 X 是一维的，因为无论是一维还是多维，逻辑分类在原理上是一样的.

定义 2. 3. 1 如果一个条件样本，其概率分布

$$P(X|y_j)=P(X|y_j \text{ 是真的})=P(X|A_j) \tag{2.18}$$

则我们称它是窗口样本——以 A_j 为窗口选出来的，称样本概率分布是窗口分布.

如果 $P(X|y_j)$是窗口分布，根据(2.17)和(2.18), 有

$$T(A_j \mid X)=[P(X \mid y_j) / P(X)] / [P(x_j^* \mid y_j) / P(x_j^*)] \tag{2.19}$$

根据**贝叶斯定理 2**，从上式可以得到

$$T(A_j \mid X)=P(y_j|X) / P(y_j|x_j^*)\text{，} j=1, 2, \ldots, n \tag{2.20}$$

其中 x_j^*也是函数曲线 $P(y_j|X)$最高点(可能不止一个)下的 x_i. 上式就是多标签逻辑分类函数. 我们这样可以理解(2.20)：$P(X|y_j)=P(X|A_j)$时，真值函数正比于转移概率函数.

容易证明，改变 $P(X)$并不影响逻辑分类, 因为 $P(y_j|X)$不变. 也容易证明，如果 A_j 是清晰集合，则窗口样本会使 $P(X|y_j)/P(X)$等于 0 或常数 c=1/$T(A_j)$，并且 $P(y_j|X)$= $P(y_j)/T(A_j)$.

定理 2.3.1 如果一个条件样本分布 $P(X|y_j)$使得 $P(X|y_j)/P(X)$等于 0 或常数 c，则可以找到一个清晰集合(窗口)；否则，可以找到一个模糊集合(模糊窗口)，使得 $P(X|A_j)=P(X|y_j)$.

证明：先考虑 A_j 是模糊集合. 令 $T(A_j|X)$= $[P(X|y_j)/P(X)]/c$, 则 $T(A_j|X)$等于 0 或 1，所以它可以作为清晰集合的真值函数. 根据贝叶斯定理 3，

$$P(X|A_j)=T(A_j|X)P(X)/T(A_j)=[P(X|y_j)/c]/T(A_j)\text{，}\ T(A_j)=1/c$$

将 $T(A_j)$=1/c 代入上面左边等式可得到 $P(X|A_j)$= $P(X|y_j)$.

再考虑 A_j 是模糊集合. 使用(2.19)和(2.20), 可以得到 $T(A_j|X)\in[0,1]$，所以 A_j 是模糊的. **证毕**.

上面讨论的求真值函数的方法是收信人的逻辑分类方法，是多标签的. 而发信人的选择分类

是单标签的(包括两个或多个标签通过逻辑乘合并为一个标签， 比如 $y_4=y_1$ 且 y_3)，即 Y 是 X 的函数. 如果几个 y_j 同样真，并且集合都是清晰的，我们选择逻辑概率最小的一个 y_j. 即选择 y_j*使得

$$y_j^*(X)=\arg\min_{y_j,\,T(A_j|X)=1} T(A_j) \tag{2.21}$$

当样本不够大时，直接用贝叶斯定理 3 得到的真值函数是不连续的，用来做贝叶斯预测(产生似然函数)是不行的. 下面讨论在样本并不足够大、Y 的真假不很确定即集合模糊时，如何用参数构造真值函数，并用最大语义信息准则(它兼容最大似然准则)做逻辑分类和选择分类.

3. 语义信息方法用于统计学习

3.1 从 Shannon 信道到语义信道

Shannon 信息论[2]中称 $P(X)$为信源， 称 $P(Y)$为信宿， 称下面转移概率矩阵为信道：

$$P(Y|X)\Leftrightarrow\begin{bmatrix} P(y_1|x_1) & P(y_1|x_2) & \ldots & P(y_1|x_m)\\ P(y_2|x_1) & P(y_2|x_2) & \ldots & P(y_2|x_m)\\ \ldots & \ldots & \ldots & \ldots\\ P(y_n|x_1) & P(y_n|x_2) & \ldots & P(y_n|x_m)\end{bmatrix}\Leftrightarrow\begin{bmatrix} P(y_j|X)\\ P(y_j|X)\\ \ldots\\ P(y_n|X)\end{bmatrix} \tag{3.1}$$

其中双向箭头表示等价. 我们称其中一行 $P(y_j|X)$ 为转移概率函数. 于是，一组转移概率函数构成一个 Shannon 信道. 转移概率函数有一个重要性质：在信源 $P(X)$变为 $P'(X)$后，可以用它和 $P'(X)$做贝叶斯预测, 得到 X 的后验概率分布 $P'(X|y_j)$; 而且 $P(y_j|X)$ 乘上一个系数 k，预测不变, 即

$$\frac{P'(X)kP(y_j|X)}{\sum_i P'(x_i)kP(y_j|x_i)}=\frac{P'(X)P(y_j|X)}{\sum_i P'(x_i)P(y_j|x_i)}=P'(X|y_j) \tag{3.2}$$

Shannon 互信息被定义为

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i|y_j)}{P(x_i)}=\sum_j\sum_i P(x_i,y_j)\log\frac{P(y_j|x_i)}{P(y_j)} \tag{3.3}$$

其中用到贝叶斯定理 2.

现在我们用随机变量 θ 表示一个模糊集合，其隶属函数 $T(\theta_j|X)$是用参数构造的，也像 $P(y_j|X)$可以用于概率预测，所以 θ 同时也是一个预测模型或一组模型参数. 我们再用 θ_j 表示 θ 的一个取值，并且 y_j=“X 在 θ_j 中”；用 $y_j(X)$表示一个谓词, 其真值函数就是 X 在 θ_j 上的隶属函数.

对比流行的似然方法，上述方法使用子模型 θ_1 , θ_2, …, θ_n 而不是只用一个模型 θ. $P(X|\theta_j)$等价于流行的似然方法中的 $P(X|y_j, \theta)$. 用来检验 y_j 的样本也是一个子样本，或条件样本，其概率分布就是 $P(X|y_j)$. 这些改变将使新的似然方法(即语义信息方法)更加灵活，更加兼容 Shannon 信息论.

当 $X=x_i$ 时，谓词 $y_j(X)$变成命题 $y_j(x_i)$, 其真值便是 $T(\theta_j|x_i)$. 于是，一个语义信道由若干真值或真值函数构成：

$$T(\theta \mid X) \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid x_1) & T(\theta_1 \mid x_2) & \ldots & T(\theta_1 \mid x_m) \\ T(\theta_2 \mid x_1) & T(\theta_2 \mid x_2) & \ldots & T(\theta_2 \mid x_m) \\ \ldots & \ldots & \ldots & \ldots \\ T(\theta_n \mid x_1) & T(\theta_n \mid x_2) & \ldots & T(\theta_n \mid x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid X) \\ T(\theta_2 \mid X) \\ \ldots \\ T(\theta_n \mid X) \end{bmatrix} \tag{3.4}$$

贝叶斯定理 3 变成：

$$P(X \mid \theta_j) = P(X)T(\theta_j \mid X) / T(\theta_j), \quad T(\theta_j) = \sum_i P(x_i)T(\theta_j \mid x_i) \tag{3.5}$$

$$T(\theta_j \mid X)=[P(X \mid \theta_j) / P(X)] / [P(x_j^* \mid \theta_j) / P(x_j^*)] \tag{3.6}$$

其中 x_j^*是函数曲线 $P(X|\theta_j)/P(X)$最高点下面的 x_i. 由式(3.2)可知，当真值函数和转移概率函数成正比时，语义贝叶斯预测和传统的贝叶斯预测——用 $P(y_j|X)$和 $P(X)$产生 $P(X|y_j)$——等价.

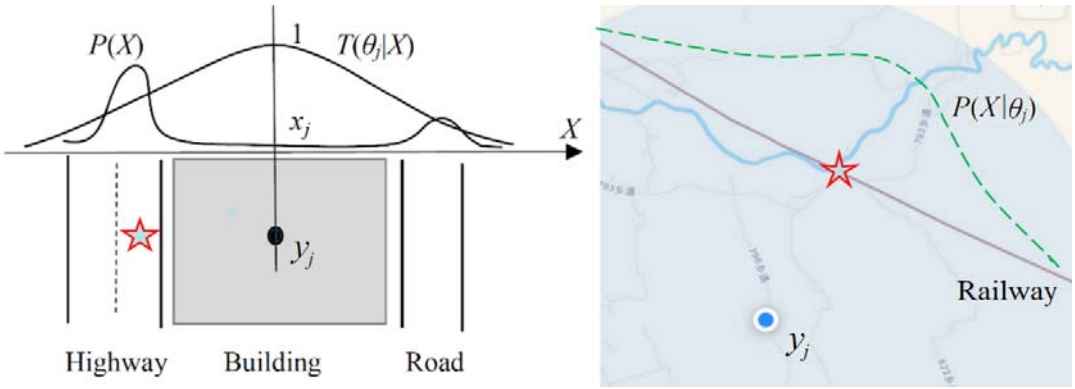
一个语义信道后面总有一个 Shannon 信道. 以天气预报为例，转移概率函数 $P(y_j|X)$反映预报语句 y_j 的选择规律，因预报员而异——有人错的少，有人错的多. 而 $T(\theta_j|X)$反映听众理解的语义，可能来自语言的定义，也可能来自过去的样本的训练. 不同的人理解的语义 $T(\theta_j|X)$是大体相同的.

3.2 理解 GPS 定位——似然函数还是真值函数？

考虑全球定位系统(GPS)显示屏上的定位(小圆圈)的语义. 它表示实际位置大概在某处. 即 y_j="$X \approx x_j$". 一个时钟、一个秤、一个温度表，含义类似. 具有这样含义的 y_j 可谓无偏估计. 一个无偏估计的语义信道可以表示为：

$$T(\theta_j|X)=\exp[-|X-x_j|^2/(2d^2)], \quad j=1, 2, \ldots, n \tag{3.7}$$

其中 d 是标准差. $|X-x_j|$是实际位置和预测位置之间的距离. 考虑 GPS 定位的特殊环境如图 3(a)所示, 其中定位指在高楼上，楼的左边是高速公路，右边是普通公路. 请问小车在哪里可能性最大？



(a) **(b)**

图 3 GPS 定位图解. 圆点是指针位置，五角星是人最为可能的位置. (a) 显示人在小车上，$P(X)$不断变化且指针有误差；(b) 显示人在高铁列车上，$P(X)$是一条轨迹而定位有误差.

Figure 3 Illustration of GPS's positioning. (a) shows that the user is in a car, $P(X)$ is changing, and the indicator has deviation; (b) shows that the user is in a high-speed train, $P(X)$ is a line, and the indicator has deviation. The round point is indicated position and the star is the user's most possible position.

参看图 3(a)，如果认为 GPS 提供似然函数预测小车位置，则小车在楼顶上的概率最大——这是不对的. 根据语义贝叶斯预测或贝叶斯定理 3，小车在高速公路上的概率最大. 再看图 2(b), 人在高铁列车上，根据语义贝叶斯预测，似然函数如虚线曲线所示，五角星表示最为可能的位置. 上述结果和人脑推理结论一致. 从 GPS 的例子可以看出，语义信道比 Shannon 信道简单，更易于理解. 图 3 显示我们可以利用先验概率分布 $P(X)$纠偏 Y; 后面说明，通过优化语义信道也可以纠偏 Y.

3.3 从似然度到语义互信息

笔者早在 1990 的文章[21]中就提出广义熵和广义互信息（就是语义互信息），广义熵的对数左边仍然是统计概率，而右边变为似然度或逻辑概率. 后来其他学者也提出这种熵，并称之为交叉熵[22]. 所以下面语义互信息也可以称之为交叉互信息. 交叉熵方法[22]和笔者的语义信息方法[6, 10-12]有很多类似之处，但是侧重点不同. 语义信息方法重在语义通信，和 Shannon 信息论及语义（即真值函数）结合比较紧密.

在 Shannon 信息论中，只有统计概率，没有逻辑概率，也没有预测的概率(似然度). 下面语义信息测度同时用到这三种概率[6]. y_j 提供关于 x_i 的信息量就是对数标准似然度：

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.8}$$

其中用到贝叶斯定理 3，并假设先验似然函数等于先验概率分布 $P(X)$. 对于无偏估计，真值函数和信息之间的关系如图 4 所示.

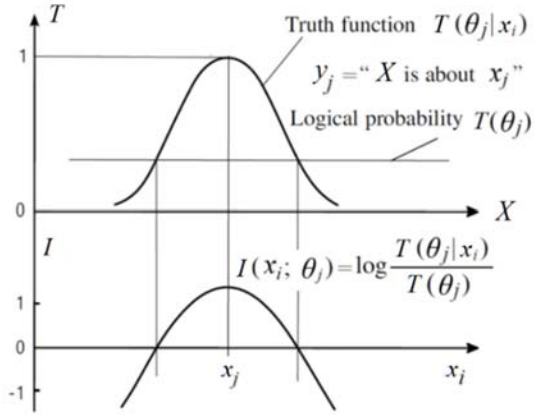


图 4 语义信息量图解. 偏差越大，信息越少；逻辑概率越小，信息量越大；错误预测提供负的信息.

Figure 4 Illustration of semantic information measure. The larger the deviation is, the less information there is; the less

the logical probability is, the more information there is; and, a wrong estimation may convey negative information.

这个公式就能反映 Popper 的思想[23]：(先验)逻辑概率越小，并能经得起检验(后验逻辑概率越大) ，信息量就越大; 永真句在逻辑上不能被证伪，因而不含有信息.

把式(3.7)中的 $T(\theta_j|X)$代入式(3.8), 就得到

$$I(x_i;\theta_j)=\log[1/T(\theta_j)]-\left|X-x_j\right|^2/\left(2d^2\right) \tag{3.9}$$

其中 $\log[1/T(\theta_j)]$ 就是 Bar-Hillel 和 Carnap 定义的语义信息测度[3]. 上述语义信息测度还考虑了偏差——语义信息量随偏差增大而减小.

对 $I(x_i; \theta_j)$求平均，就得到广义 Kullback-Leibler (KL)信息：

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.10}$$

其中对数左边是统计概率 $P(x_i|y_j)$，i=1, 2, …，它们构成样本概率分布 $P(X|y_j)$, 是用以检验 θ_j 的.

对 $I(X; \theta_j)$求平均，就得到广义或语义互信息公式：

$$\begin{aligned}&I(X;\theta)=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\\&=H(\theta)-H(\theta\mid X)\\&H(\theta)=-\sum_j P(y_j)\log T(\theta_j),\ H(\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i)\end{aligned} \tag{3.11}$$

容易证明，在语义贝叶斯预测和样本分布一致时， 即 $P(x_i|\theta_j)=P(x_i|y_j)$ (对于所有 i, j)时，上述语义互信息达到其上限，等于 Shannon 互信息. 从式(3.9)和(3.11)可见， 语义互信息准则和流行的误差加正则化准则是类似的. $H(\theta|X)$就是误差项，$H(\theta)$就是正则化项. $I(X; \theta)$就是负的损失函数.

3.4 似然度和语义信息之间的关系

Akaike 揭示了似然度和 KL 信息之间的联系[24]. 然而，上述广义 KL 信息和似然度之间的关于更加简单. 假设相应 y_j 有 N_j 个样本点 $x(1), x(2), \ldots, x(N_j)\in U$，它们来自 N_j 个独立同分布随机变量，其中 x_i 有 N_{ij} 个; 当 N_j 很大时，就有 $P(X|y_j)= N_{ij}/N_j$. 因此就有 log(标准似然度)和广义 KL 信息之间关系：

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_j I(X;\theta_j) \tag{3.12}$$

对不同的 y_j 求平均，就得到平均 log(标准似然度)，它和语义互信息的关系是：

$$\sum_j \frac{N_j}{N}\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}$$
$$=I(X;\theta)=H(X)-H(X\mid\theta) \tag{3.13}$$
$$H(X\mid\theta)=-\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log P(x_i|\theta_j)$$

其中 $H(X|\theta)$是 X 的广义后验熵[6](属于交叉熵[22]). 因为优化模型 θ_j 时 $P(X)$不变，所以最大似然准则等价于最大语义互信息准则. 容易证明：Shannon 互信息是语义互信息的上限；似然函数和样本分布符合时，两者相等.

3.5 语义信道优化——匹配 Shannon 信道

优化一个语义信道等价于优化一个预测模型 θ 或一组子模型(θ_1，θ_2，…, θ_n). 给定 Shannon 信道时优化子模型 θ_j，也就是优化隶属函数或逻辑分类函数 $T(\theta_j|X)$. 于是有

$$T^*(\theta_j\mid X)=\arg\max_{T(\theta_j|X)} I(X;\theta_j)=\arg\max_{T(\theta_j|X)}\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.14}$$

$I(X;\theta_j)$可以写成两个 KL 距离的差，

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i)}-\sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i\mid\theta_j)} \tag{3.15}$$

因为当 $P(X|\theta_j)=P(X|y_j)$时，后一项为 0，所以这时 $I(X;\theta_j)$最大，等于 KL 信息 $I(X;y_j)$. 根据贝叶斯定理 3 可以得到

$$T^*(\theta_j|X)=P(y_j|X)/P(y_j|x_j^*)=[P(X|y_j)/P(X)]/[P(x_j^*|y_j)/P(x_j^*)] \tag{3.16}$$

等式(3.16)适合有大样本时的非参数估计. 式(3.14)也适合只有小样本时的参数估计. 当样本足够大时，用 $T^*(\theta_j|X)$做语义贝叶斯预测和用 $P(y_j|X)$做贝叶斯预测，结果相同. 所以和流行的贝叶斯推理相比，上述方法更加兼容贝叶斯定理 2.

3.6 语义信息准则用于多类别多标签分类——逻辑分类和选择分类

当条件样本足够大时， 我们可以通过(3.16)得到 n 个优化的真值函数(或一个语义信道)，这些真值函数就是多类别多标签逻辑分类函数. 如果样本不够大， 我们可以通过(3.14)得到这些真值函数. 用这样的真值函数，收信人可以通过贝叶斯定理 3 得到语义贝叶斯预测.

在实例 X 可见时，发信人用信息准则做选择分类

$$y_j^*(X)=\arg\max_{y_j}\log\frac{T(\theta_j\mid X)}{T(\theta_j)} \tag{3.17}$$

当所有集合是清晰的时候，上面信息准则就退化为最小逻辑概率准则，如(2.21)所示.

如果实例是不可见的(比如我们根据人的声音把人分成男女，或者根据西瓜的外观把西瓜分为好瓜和差瓜)，我们只知道观察数据 Z，则可用平均语义信息准则选择标签， 即

$$y_j^*(Z)=\arg\max_{y_j}\sum_i P(x_i\mid Z)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.18}$$

因为考虑到预测不准，这样选出来的 y_j^*, 其逻辑概率未必最小.

因为可以由转移概率函数得到真值函数，并允许 $T(\theta_1|X)+T(\theta_2|X)+\ldots+T(\theta_n|X)\neq 1$，多类别分类就可以简单地转化为单类别隶属函数求解，设计每个标签的真值函数的参数形式时不用考虑其他标签真值函数的参数形式. 比如，X 表示年龄，y_1=“X 是小孩”，类别是模糊的，其隶属函数可用 Logistic 函数表示；y_2=“X 是成年人”(相应的集合是清晰的，$A_2=\{X|X\geqslant 18\}$)，y_3=“X 是年轻人”(隶属函数可用没有系数的高斯分布表示)，…. 在待分类样本中， 无论是两种标签，还是 n 种标签，隶属函数的参数形式都是一样的. 换一种场合，学习得到些隶属函数同样有效——这正是迁移学习所需要的.

通过比较隶属函数，我们可以求出不同假设 Y 之间的蕴含关系. 当对所有 X，$T(\theta_j|X)\leq T(\theta_k|X)$ 时，y_j 蕴含 y_k.

实例不可见的分类，在本质上和检验及估计相同，下面详细讨论.

4. 信道匹配算法用于估计、检验和混合模型

4.1 医学检验和 GPS 的语义信道优化

我们用随机变量 Z 表示观察条件，$Z\in C=\{z_1, z_2, \ldots, z_w\}$. 发信人根据 Z，选择 Y；收信人根据 Y 预测 X. 从假设-检验的角度看，X 是证据或样本点，Y 是假设; 我们用样本的概率分布评价和检验一个假设. 比如对于天气预报，X 是日降雨量，Y 是预报语句，Z 是预测依据的气象数据；对于医学检验，X 是真有病或真没病的测试者， Y 是阳性或阴性，Z 是化验数据. 对于 GPS，X 是带有 GPS 装置的小车的实际位置，Y 是 GPS 指针所指位置，Z 是 GPS 设备到三个卫星的距离.

对于西瓜分类，我们可以用 X 表示瓜瓤质量——由口感、色泽、沙瓤与否决定，包含价值判断，是连续的，在很好和很差之间变化，比如在 0 和 10 之间变化. 给定 $P(X, Z)$(根据过去的样例得到)和 C 的一个划分$\{C_0, C_1\}$，我们也可以求出关于瓜瓤好坏的预测 y_0 和 y_1, 从而可以求出逻辑分类函数 $T(\theta_0|X)$和 $T(\theta_2|X)$和语义贝叶斯预测 $P(X|\theta_0)$和 $P(X|\theta_1)$. 如果瓜瓤好坏简单用 x_1 和 x_2 表示，那么根据外观 Z 预测瓜瓤好坏就和医学检验一样，是检验问题. 从分类的角度看，x_1 和 x_2 是真的分类标签，y_1 和 y_2 是根据 Z 预测的分类标签. 优化预测就是改变划分$\{C_0, C_1\}$，最大化 Shannon 互信息 $I(X; Y)$或语义互信息 $I(X; \theta)$(两者这时等价).

对于医学检验(参看图 5)，这时 $U=\{x_0, x_1\}$， $V=\{y_0, y_1\}$. 其中 x_0 是真没病者，x_1 是真有病者；y_0=检验呈阴性，y_1=检验呈阳性；$Z\in C=\{z_1, z_2, \ldots\}$，$z$’决定了划分$\{C_0, C_1\}$， 它和判决函数 $Y=f(Z)$ 相互确定.

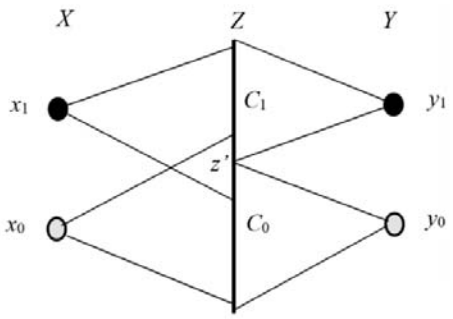


图 5 医学检验图解(二元有噪声 Shannon 信道，互信息随判决分界点 z'改变)

Figure 5 Illustrating the medical test. The test can be abstracted as a 2×2 Shannon nosy channel. The Shannon mutual information changes with the dividing point z'.

医学检验中把 x_1 检验为阳性的概率叫做敏感性(sensitivity), 把 x_0 检验为阴性的概率叫做特异性(specificity) [25]. 检验的敏感性和特异性构成 Shannon 信道，如表 1 所示.

表 1 医学检验的敏感性和特异性构成 Shannon 信道 $P(Y|X)$

Table 1 The sensitivity and Specificity of Medical Tests Form a Shannon's Channel $P(Y|X)$

	真有病 x_1	真没病 x_0
检验是阳性 y_1	$P(y_1\|x_1)$=敏感性	$P(y_1\|x_0)$=1-特异性
检验是阴性 y_0	$P(y_0\|x_1)$=1-敏感性	$P(y_0\|x_0)$=特异性

如果我们相信阳性表示绝对有病，阴性表示绝对无病，那么就有非模糊命题的真值 $T(y_1|x_1)=T(y_0|x_0)=1$, $T(y_1|e_0)=T(y_0|x_1)=0$. 但是采用这样的语义信道，有一个反例存在，就会有负无穷大信息. 为此，我们需要考虑预测和检验的置信度(confidence level)——用 b 表示，并且用 b'=1-|b|表示不信度(no-confidence–level, 或显著性水平). y_j 的真值函数可定义为：

$$T(\theta_j|X)= b' + bT(y_j|X) \tag{4.1}$$

设阳性 y_1 的置信度 b_1, 不信度是 b_1'; 阴性 y_0 的置信度是 b_0, 不信度是 b_0'. 则医学检验的语义信道如表 2 所示.

表 2 医学检验的语义信道——含有两个不信度 b_1' 和 b_0'

Table 2 Two No-confidence Levels of a Medical Test Form a Semantic Channel $T(\theta|X)$

	真有病 x_1	真没病 x_0
检验是阳性 y_1	$T(y_1\|x_1)$=1	$T(y_1\|x_0)= b_1'$
检验是阴性 y_0	$T(y_0\|x_1)= b_0'$	$T(y_0\|x_0)$=1

根据式(3.16), 两个优化的不信度是

$$b_1'^*= P(y_1|x_0)/P(y_1|x_1); \quad b_0'^*= P(y_0|x_1)/P(y_0|x_0) \tag{4.2}$$

医学界用似然比 LR^+和 LR^-表示检验有多好[24]. 上式来自最大语义信息检验， 它和最大似然

比检验是兼容的，因为：

$$\mathrm{LR}^{+}=P(y_1|x_1)/P(y_1|x_0)=1/b_1'^{*}=1/(1-b_1^{*});\quad \mathrm{LR}^{-}=P(y_0|x_0)/P(y_0|x_1)=1/b_0'^{*}=1/(1-b_0^{*}) \tag{4.3}$$

其中 b_1^*是优化的 y_1 的置信度，在 0 和 1 之间变化，和 LR^+正相关且一一对应. b_0^*同理. 从(4.2)和(4.3)看，最大似然准则接近最小相对误差准则，报错相对于报对的比例越小越好；而不是总体报错的比例越小越好. 如果假设的逻辑概率较小，比如地震预报，我们就更要减少漏报.

有了上述语义信道，发信人就可以根据 Z, 选择最优 Y，即根据坐标 Z 上的两个信息(广义 Kullback-Leibler 信息)曲线：

$$I(X;\theta_j|Z)=\sum_i P(x_i\mid Z)I(x_i;\theta_j)\text{，}\ j=0,\ 1. \tag{4.4}$$

选择 Y. $I(X;\ \theta_1|Z)>I(X;\ \theta_0|Z)$则选 y_1, 否则选 y_0. 这样就会产生一个新的分界点 z'*.

困难的是，z'^*取决于 z'， 而 z'是否合理？z'能保证 z'^*是全局最优的？我们最初可以根据 $I(x_i;\ z_k)=\log[P(x_i|z_k)/P(x_i)]$选择 Y，即 $I(x_1;\ z_k)>I(x_0;\ z_k)$则选择 y_1, 否则选择 y_0. 但是这样得到的分界点 z' 并不能确保减少小概率事件的漏报，也不能保证 z'^*能最大化互信息 $I(X;\ \theta)$. 为此，我们要用到迭代方法，后面谈及.

对于 GPS, 观察数据主要是 GPS 设备到三个卫星的距离. 使用语义信息方法，我们可以得到隶属函数——它可以消除来自其他因素的系统误差. 假定 GPS 的 Shannon 信道是

$$P(y_j\mid X)=K\exp[-|\,X-x_j-\Delta x\,|^2/(2d^2)],\quad j=1,\ 2,\ \ldots,\ n \tag{4.5}$$

其中 x_j 是所指位置，即 y_j="$X=x_j$"中的 x_j；K 是系数，Δe 是系统偏差，d 是标准偏差， 那么根据式(3.16)就有优化的语义信道

$$T^*(\theta_k\mid X)=\exp[-|\,X-x_k\,|^2/(2d^2)],\quad k=1,\ 2,\ \ldots,\ n \tag{4.6}$$

其中 $x_k=x_j+\Delta x$. 温度表、秤、指数预测…提供的语义信道优化类似.

4.2 语义贝叶斯决策和翻译

在图 2(a)中，有了指针位置 y_j, 我们需要判断小车具体在哪里？ 在高速公路上，普通道路上，还是楼顶上？假设 y_1=“小车在高速公路上”和 y_2=“小车在普通路上”属于不同于 V 的集合，其真值函数是 $T(\theta_1|X)$和 $T(\theta_2|X)$ (覆盖范围更小的无系数高斯分布), 我们就可以把 $I(x_i;\ \theta_k),\ k=1,\ 2$ 当做奖惩函数(即负的损失函数)用于贝叶斯决策，得到 y_1 和 y_2 的平均语义信息：

$$I(X;\theta_k|y_j)=\sum_i P(x_i\mid\theta_j)\log\frac{T(\theta_k\mid x_i)}{T(\theta_k)},\quad j=1,\ 2 \tag{4.7}$$

其中 $P(x_i|\theta_j)$来自语义贝叶斯预测. 决策在这里是选择使 $I(X;\ \theta_k|y_j)$达最大的 y_k.

自然语言翻译是类似的. y_j 就源语句，y_k, $k=1,\ 2$ 就是目标语句. 如果最为接近 $P(X|\theta_j)$的是 $P(X|\theta_{k^*})$，则提供平均信息量最大的 y_{k^*}就是最优翻译.

以上讨论的 Shannon 信道都是给定的，即 $y_j=f(Z|Z\in C_j)$，划分$\{C_1,\ C_2,\ldots,\ C_n\}$是给定的. 下面考虑 Shannon 信道或 C 的划分可变时，如何用迭代方法求出使最优划分，得到使平均对数似然

度达最大的 Shannon 信道.

4.3 用信道匹配算法(CM 算法)求检验，估计和预测的最大互信息和最大似然度

在语义互信息公式(3.11)中，固定 $P(Y|X)$改变 $T(X|\theta)$并最大化 $I(X;\theta)$, 这一过程可谓“语义信道匹配 Shannon 信道”(匹配 I). 在 $P(X|\theta_j)=P(X|y_j)$或 $T(X|\theta_j)\propto P(y_j|X)$对所有 j 成立时，语义互信息 $I(X;\theta)$达到最大值, 等于 Shannon 互信息 $I(X;Y)$. 但是反过来，令 $P(y_j|X)\propto T(\theta_j|X)$(对所有 j)未必能增加 Shannon 互信息或语义互信息. 给定语义信道，可能存在更好的 Shannon 信道(比如噪声更少)，传递更多的语义信息. 寻找这样 Shannon 信道可谓让 Shannon 信道匹配语义信道(匹配 II). 信道匹配算法(即 CM 算法)就是重复匹配 I 和匹配 II 的迭代算法.

笔者曾推广 Shannon 的信息率失真函数 $R(D)$[27, 28]得到 $R(G)$函数[6, 10]，其中 G 是语义互信息的下限，R 是给定 G 时的最小 Shannon 互信息. 我们最近的研究表明：通过语义信道和 Shannon 信道相互匹配和迭代可以找到使 Shannon 互信息最大化的 Shannon 信道，它决定了语义互信息和似然度的上限. 迭代收敛可以通过 $R(G)$函数得到直观解释(参看图 6).

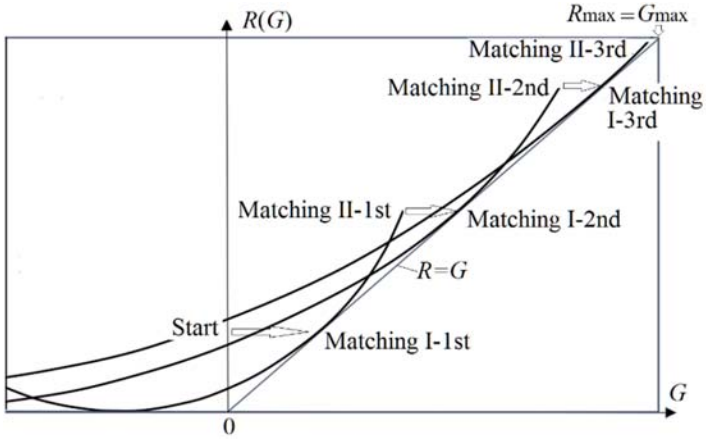


图 6. 图解检验和估计的迭代收敛. 每个语义信道决定一条 $R(G)$ 函数曲线. 匹配 I 使得 $G=R$，并产生一个新的语义信道；匹配 II 增大 R 到 $R(G)$ 函数曲线的顶部. 重复匹配 I 和匹配 II 可以使得 R 和 G 达到最大值 R_{max} 和 G_{max}.

Figure 6. Illustrating the iterative convergence for tests and estimations. Each semantic channel ascertains a $R(G)$ function curve. The matching I is for $G=R$ and a new semantic channel; the matching II is to increase R to the top of a $R(G)$ function. Repeating the matching I and matching II can maximize R and G to obtain R_{max} and G_{max}.

语义信道匹配 Shannon 方法如式(3.16)所示. 参考信息率失真函数的参数形式推导过程[27]，我们得到 Shannon 匹配语义信道方法：令

$$P(y_j|Z)=\lim_{s\to\infty}\frac{P(y_j)[\exp(I(X;\theta_j|Z))]^s}{\sum_{j'}P(y_{j'})[\exp(I(X;\theta_{j'}|Z))]^s},\quad j=1,2,\ldots,n \tag{4.8}$$

当 $s\to\infty$, $P(y_j|Z)$变成集合 C_j 的特征函数，取值于{0,1}. 上面公式的直观解释就是，给定 Z，哪个 y_j 使 $I(X;\theta_j|Z)$达最大，我们就把 Z 划到相应的 C_j, 选择 $y_j=f(Z|Z\in C_j)$.

例 4.3.1 对于图 4 所示检验，设 $Z\in C=\{1,2,\ldots,100\}$，给定 x_1 和 x_0 时 $P(Z|X)$是高斯分布

$$P(Z|x_1)=K_1\exp[-(Z-c_1)^2/(2d_1^2)],\quad P(Z|x_0)=K_0\exp[-(Z-c_0)^2/(2d_0^2)]$$

其中 K_1 和 K_0 是归一化常数. $P(x_0)$=0.8; c_0=30, c_1=70; d_0=15, d_1=10. 求使 Shannon 互信息最大的划分点 z*.

迭代求解：从 $P(X)$和 $P(Z|X)$算出条件概率分布 $P(X|Z)$ (具体数值省略). 假定开始的划分点是 z', 比方说 z'=50, 做下面迭代运算：

匹配 I：依次计算下面各项(具体数值省略):

构成 Shannon 信道的 4 个转移概率，两个不信度 b_1'* 和 b_0'*和两个逻辑概率 $T(\theta_1)$和 $T(\theta_2)$;

四个信息量 $I_{ij}=I(x_i; \theta_j)$ (根据式(3.8))，i=0, 1; j=0, 1;

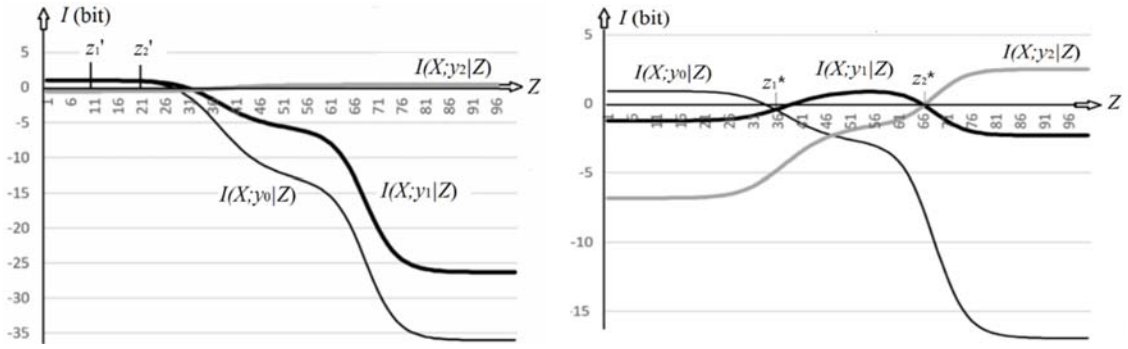
给定不同 Z 时的平均语义信息 $I(X; \theta_1|Z)$ 和 $I(X; \theta_0|Z)$ (显示为两条曲线).

匹配 II：利用式(4.8)改进 C 的划分. 如果划分点和上个 z'相同，则令最优分界点 z*=z', 迭代结束；否则转到匹配 I.

迭代过程：一次迭代后 z'=53; 两次迭代后 z'=54; 三次迭代后 z'不变, 收敛点 z*=54. **解毕**.

例 4.3.2 一个 3X3 Shannon 信道用作简化的估计(和检验不同，迭代方法同理)，$P(x_0)$=0.5, $P(x_1)$=0.35，$P(x_2)$=0.15; c_0=20, c_1=50，c_2=80; d_0=15, d_1=10, d_2=10. 求解两个最优划分点 z_1*和 z_2*. 用两组不同起点: (1) z_1'=50 和 z_2'=60；(2) z_1'=9 和 z_2'=20(很差的一对).

迭代结果：对于(1), 迭代 4 次收敛，得到 z_1*=35 和 z_2*=66. 对于(2), 迭代 11 次收敛，同样得到 z_1*=35 和 z_2*=66. 对于(2), 迭代前后的三条信息曲线如图 7 所示. 可见迭代收敛可靠[1].



(a) (b)

图 7 分界起点很差时的迭代. (a)显示迭代开始时三条信息曲线正的部分很小；(b)显示了迭代收敛时三条信息曲线正的部分较大.

Figure 7 The iteration with bad start points. (a) shows that at the beginning of the iteration, three information curves cover very small positive areas; (b) shows that at the end of the iteration, three information curves cover much larger positive areas.

由上面两个例子可见，用 CM 算法求解信道不确定时的最大似然检验和估计，快速且可靠。而且收敛可以通过 $R(G)$函数得到直观证明. 流行的求解最大互信息和最大似然度方法是牛顿法[29]，

[1] 检验、估计和混合模型的迭代过程见 excel 文件: http://survivor99.com/lcg/CMiteration.zip

梯度法[30]和最小最大法[31]. 我们比较过 CM 算法和流行算法求解混合模型的速度，CM 算法明显较快[12]. CM 算法用于检验和估计的速度没有和流行方法做具体比较. 但是 CM 算法用于检验和估计比用于混合模型更加简单，其速度比流行方法明显较快的可能性很大. 更详细分析见[13].

我们可以把 CM 算法用到一般预测， 比如天气预报. 这时候就可以用 CM 算法解释语义进化. Shannon 信道反映语言用法，而语义信道反映听众理解方式. 语义信道匹配 Shannon 信道(匹配 I)就是理解匹配用法；Shannon 信道匹配语义信道(匹配 II)就是用法匹配理解. 语义信道和 Shannon 信道相互匹配和迭代，就是预报者用法和听众理解相互匹配，相互促进. 自然语言应该就是这样进化的.

4.4 CM 算法用于混合模型

CM 算法和 EM 算法[32]类似，也可用于混合模型(一种聚类). 假设样本分布 $P(X)$是某种条件概率分布(比如高斯分布)$P^*(X|Y)$按比例 $P^*(Y)$混合产生的. 我们只知道模型构件是 n 个. 要求的是 $P^*(Y)$和模型参数 θ^*. 和检验不同，预测不再有对错，但是要求预测的样本分布 $P(X|\theta)$——简记为 $Q(X)$——和 $P(X)$尽可能接近(似然度尽可能大)，相对熵(即 Kullback-Leibler 距离)$H(Q||P)$尽可能小.

我们以两个高斯分布函数的混合为例说明 CM 算法如何用于混合模型. 迭代之前初始化 $P(Y)$(比如假设等概率)，初始化两个高斯分布函数

$$P(X|\theta_j)= K_j \exp[-(X-c_j)^2/(2d_j)^2], \quad j=1,2 \tag{4.9}$$

中的四个参数(c_1, c_2, d_1, d_2). 其中 K_j 是归一化系数. 然后开始迭代运算. 每次迭代分为三步：

匹配 II-a： 令 $P(y_j|X)$是通过 $P(X|\theta)$和 $P(Y)$产生的, 等于 EM 算法中的 E-step[31]，即

$$P(y_j \mid X) = P(y_j)P(X \mid \theta_j) / Q(X), \qquad Q(X) = \sum_j P(y_j)(X \mid \theta_j)\text{，} \ j=1, 2, \ldots, n \tag{4.10}$$

匹配 II-b： 改变 $P(y_j)$使 $P^{+1}(y_j)=\sum_i P(x_i)P(y_j|x_i)\approx P(y_j)$. 如果 $H(Q||P)<0.001$ 则迭代结束.

匹配 I： 轮流改变四个参数最大化语义互信息(设为 G):

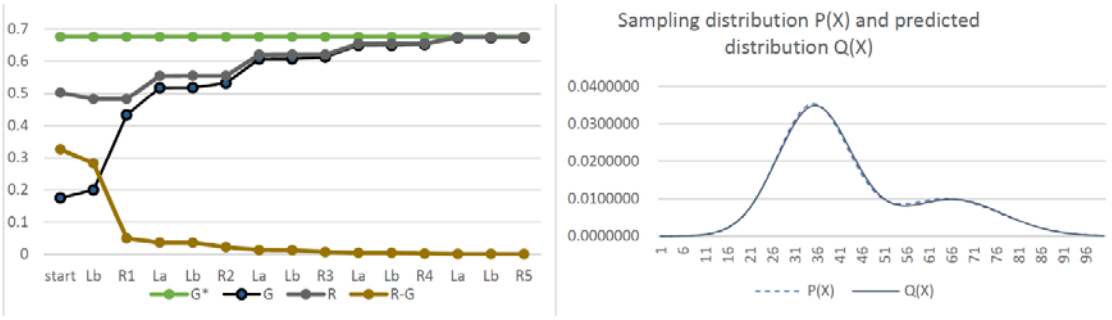
$$G = I(X;\theta) = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)} \tag{4.11}$$

然后转到**匹配 II-a**.

例 4.1.1 一个混合模型例子如表 3 所示. 设 R^*是真实的 Shannon 互信息，G^*是与之相等的语义互信息. 表中显示了真实参数和 $P^*(Y)$及迭代前后的参数和 $P(Y)$. 迭代收敛过程如图 7 所示。

表 3 $R<R^*$时的模型参数和迭代结果(迭代次数是 5)

Y	真实 $P^*(X\|Y)$和 $P^*(Y)$			初始参数 $H(Q\|\|P)$=0.410 bit			收敛后参数 $H(Q\|\|P)$=0.00088 bit		
	c	d	$P^*(Y)$	c	d	$P(Y)$	c	d	$P(Y)$
y_1	35	8	0.7	30	15	0.5	35.4	8.3	0.720
y_2	65	12	0.3	70	10	0.5	66.2	11.4	0.280



(a) (b)

图 8 （a）混合模型迭代步骤；（b）5 次迭代后 $H(Q\|\|P)$接近 0.

Figure 8 (a) The iterative process of the mixture model; (b) $H(Q\|\|P)$ is close to 0 after 5 iterations.

CM 算法收敛证明：设

$$R_Q = I_Q(X;Y) = \sum_i \sum_j P(x_i)\frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j)\log\frac{P(x_i \mid \theta_j)}{Q(x_i)} \tag{4.12}$$

容易证明，

$$R_Q - G = \sum_i P(x_i)\log Q(x_i) = H(Q\|\|P) \tag{4.13}$$

$$R_Q = R + H(Y\|\|Y^{+1}),\ H(Y\|\|Y^{+1}) = \sum_j P^{+1}(y_j)\log[P^{+1}(y_j)/P(y_j)] \tag{4.14}$$

其中 R 是 Shannon 互信息. 根据 $R(G)$函数的定义[6, 10]，$R(G)$是给定 G 的 R 的最小值. 因为**匹配 I** 增大 G，**匹配 II-b** 使 $H(Y\|\|Y^{+1})=0$，**匹配 II-a** 最小化 R (在求 $R(D)$函数参数形式的迭代过程中就用到匹配 II[28]), 所以根据(4.13)和(4.14)，$H(Q\|\|P)$在每一步减小，因此迭代收敛. 因为 $R(D)$函数是凹的(像是半个碗)[28]，$R(G)$是 $R(D)$的自然延伸[6], 所以 $R(G)$是碗状的. $R(G)$-G 也是碗状的，极小值是唯一的, 所以迭代全局收敛. **证毕**.

和 EM 算法的收敛证明[32]比，CM 算法的收敛证明要清晰得多，迭代收敛也明显较快[12]. 我们用不同真实参数和 $P^*(Y)$检验了 CM 算法. 结果表明，CM 算法达到收敛的迭代次数中 5 出现最多，大多在 4-12 次；而 EM 算法收敛大多在 17 次左右[12]. 我们还找到 R 和 G 并不是单调增加的例子[12]，这时 CM 算法的收敛依然可靠. 而这些例子对 EM 算法收敛证明[31, 32]是严峻挑战.

5. 结语

第三种贝叶斯定理是已有两种贝叶斯定理(贝叶斯提出的和 Shannon 使用的两种)的推广， 用它可以从语句的真值函数(或集合的隶属函数)和实例 X 的先验概率分布(即信源)求出似然函数，反过来也可以从似然函数(或实例的后验概率分布)和信源，求出语句的真值函数(或集合的特征函数). 当信源变化时，所求真值函数仍然可以用来做贝叶斯预测. 所用公式在集合模糊时同样适用. 在样本较大时，可以简单地通过算术运得到优化的真值函数；但是在样本不够大时，需要使用语义信

息准则优化真值函数——它也就是预测模型. 这样就能保证：1)模型能反映语义; 2)坚持使用最大似然准则(因为语义信息准则与之兼容)；3)模型适合信源可变场合；4)语义贝叶斯预测兼容传统的贝叶斯预测(即根据贝叶斯定理 2 所作的预测).

从语义通信的角度看待统计学习，分类要分收信人分类(逻辑分类)和发信人分类(选择分类). 优化预测模型就是收信人理解的语义信道匹配 Shannon 信道，得到优化的多标签逻辑分类；优化贝叶斯决策或选择分类就是发信人使用的 Shannon 信道匹配语义信道. 两种信道相互匹配和迭代推动语义通信进化. 在 Shannon 信道不确定时(即标签选择规则不确定时)，我们可以使用信道匹配算法(CM 迭代算法)方便求解具有最大互信息和最大似然度的检验和估计，和求解具有最小相对熵的混合模型. 研究显示，和流行的方法比，CM 算法速度较快，收敛理由更清晰. 可以期望，第三种贝叶斯定理和信道匹配算法会有更多应用.

致谢 感谢汪培庄教授最近几年的鼓励和支持(汪培庄教授是作者 1991 年在北师大数学系做访问学者时的指导老师). 是汪教授的鼓励，促使我继续多年前的语义信息理论及其应用研究.

补充材料:

1） 信道匹配算法用于检验，估计和混合模型实例(Excel 文件，有说明)下载：
http://survivor99.com/lcg/CMiteration.zip

2） 关于信道匹配算法更详细讨论：http://survivor99.com/lcg/CM/Recent.html 包括从 EM 算法改进到 CM 算法的通俗讲解： **EM 算法是炼金术吗？**

3） 广义信息论(包括 R(G)函数)研究：http://survivor99.com/lcg/books/GIT/

参考文献

1 Jaynes E T. Probability Theory: The logic of Science, Edited by Larry Bretthorst, Cambridge University press，New York, 2003

2 Shannon C E, 1948, A mathematical theory of communication, Bell System Technical Journal, 1948, 27: 379–429 and 623–656

3 Bar-Hillel Y, Carnap R. An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT, 1952.

4 钟义信， 信息科学原理，邮电大学出版社，北京，2002

5 Floridi L, Semantic conceptions of information, in Stanford Encyclopedia of Philosophy. First published Wed Oct 5, 2005; substantive revision Wed Jan 7, 2015. https://plato.stanford.edu/entries/information-semantic/

6 鲁晨光，广义信息论，中国科学技术大学出版社，合肥， 1993

7 Anon. Bayesian Inference，Wikipedia: the Free Encyclopedia. Edited on 14 December 2017 https://en.wikipedia.org/wiki/Bayesian_inference

8 Bayes T, Price R. An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society of London. 1763, 53(0): 370–418

9 Zadeh L A. Fuzzy sets. Information and Control, 1965, 8(3): 338–53

10 鲁晨光，广义熵和广义互信息的编码意义，通信学报，1994, 15(6): 37-44

11 Lu C. A generalization of Shannon's information theory, Int. J. of General Systems, 1999, 28 (6): 453-49

12 Lu C. Channels' matching algorithm for mixture models, in IFIP International Federation for Information Processing 2017, Shi et al. (Eds.), Springer International Publishing, Switzerland, 2017, 321–332

13 Lu C. Semantic channel and shannon channel mutually match and iterate for tests and estimations with maximum mutual information and maximum likelihood，in Proceedings of 2018 IEEE International Conference on Big Data and Smart Computing，Shanghai, Du et al (eds.), January 2018, 15-18

14 Tarski A. The semantic conception of truth: and the foundations of semantics，Philosophy and Phenomenological Research, 1944, 4(3): 341–376

15 Davidson D. Truth and meaning. Synthese, 1967, 17: 304-323

16 Zhang M L, Zhou Z H. A review on multi-label learning algorithm. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837

17 Zhang M L, Zhou Z H. ML-kNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7): 2038–2048

18 Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification. European Conference on Machine Learning, 2007, 69 (10): 406-417

19 McCallum A. Multi-label text classification with a mixture model trained by EM. in Working Notes of the AAAI'99 Workshop on Text Learning, Orlando, FL, 1999

20 Zhou Z H, Zhang M L, Huang S J, Li Y F. Multi-instance multi-label learning. Artificial Intelligence, 2012, 176(1): 2291–2320

21 Lu C. Shannon equations reform and applications. BUSEFAL, 1990, 44(4): 45—52

22 de Boer P T, Kroese D P, Mannor S, Rubinstein R Y. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134 (1): 19–67

23 Popper K. Conjectures and Refutations. Repr. Routledge, London and New York, 1963/2005

24 Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19:716–723

25 Thornbury J R, Fryback D G, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology, 1975, 114(3): 561–565

26 周志华，《机器学习》，清华大学出版社，北京， 2016

27 Shannon C E. Coding theorems for a discrete source with a fidelity criterion, IRE Nat. Conv. Rec., Part 4, 1959, 142–163

28 周炯槃，信息理论基础, 中国邮电出版社，北京，1983

29 Kok M, Dahlin J, Schon B, Wills T B, Wills A. Newton-based maximum likelihood estimation in nonlinear tate space models. IFAC-PapersOnLine, 2015, 48: 398–403

30 Anon ， Gradiant Descent ， The free encyclopedia ， edited on 16 December 2017, https://en.wikipedia.org/wiki/Gradient_descent

31 Barron A, Roos T, Watanabe K. Bayesian properties of normalized maximum likelihood and its fast computation. IEEE IT Symposium on Information theory, 2014. 1667-1671

32 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977, 39: 1–38

33 Wu C F J. On the convergence properties of the EM algorithm. Annals of Statistics, 1983, 11(1): 95–10

The Third Kind of Bayes' Theorem for Semantic Communication and Statistical Learning

College of Intelligence Engineering and Mathematics, Liaoning Engineering and Technology University, Fuxin, Liaoning, 123000, China

Email： lcguang@foxmail.com

Abstract The first kind of Bayes' theorem proposed by Bayes describes the symmetrical relationship of two logical probabilities. The second kind of Bayes' theorem used in Shannon information theory describes the symmetrical relationship between two statistical probabilities. This study proposes the third kind of Bayes' theorem which describes the asymmetrical relationship between a statistical probability and a logical probability. According this theorem, we can obtain a likelihood function from a truth function, and can also obtain a truth function from a likelihood function or a conditional sampling distribution. The semantic information measure is defined with log normalized likelihood. A group of truth functions form a semantic channel, which indicates the multiclass multi-label logical classification of a sample and may be used as a predictive model. To train a predictive model by a sample is to let the semantic channel match the Shannon channel. To select hypotheses or labels with maximum semantic information or maximum likelihood criterion is to let the Shannon channel match the semantic channel. Letting two channels mutually match and iterate, we may achieve classifications, tests, estimations, and mixture models with maximum likelihood more conveniently.

Keywords Bayes' theorem, semantic information, truth function, likelihood function, multi-label classification, tests, estimations, mixture models