投稿 ICIS2018，Workshop：Statistical Leaning for Intelligent Information Processing
欢迎各位一阅并提供宝贵批评建议。--作者

从贝叶斯推理到逻辑贝叶斯推理

——一个新的数学框架用于语义通信和机器学习

鲁晨光
辽宁工程技术大学智能工程和数学学院 阜新 123000)
lcguang@foxmail.com

摘要 贝叶斯推理（BI）使用贝叶斯后验而逻辑贝叶斯推理（LBI）使用真值函数或隶属函数作为推理工具。提出 LBI 主要是因为贝叶斯推理不兼容传统的贝叶斯预测，也没使用能反映语义的逻辑概率。在新的数学框架中，统计概率和逻辑概率被严格区分并被同时使用；通过新发现的第三种贝叶斯定理，两种概率可以相互转换；我们可以从 Shannon 信道直接导出一组真值函数或语义信道。语义通信模型用作机器学习模型分两部分: 接收者的标签学习（也就是语义信道匹配 Shannon 信道）；发送者的标签选择或分类（也就是 Shannon 信道匹配语义信道）。最大语义信息(MSI)准则等价于最大似然(ML)准则，也兼容正则化最小误差平方(RLS)准则。两种信道相互匹配就能方便地实现多标签分类——同时考虑到类别不平衡和实例先验概率分布变化，而不需要考虑二元关联。使用信道匹配(CM)迭代算法，就能得到不可见实例分类和最大似然估计(属于半监督学习)。求解混合模型(属于无监督学习)则需要另一种匹配——最小化 Shannon 互信息和语义互信息之差。两种迭代的收敛都可以通过 $R(G)$函数证明——$R(G)$函数是信息率失真函数 $R(D)$的改进，G 是语义互信息，可被理解为负的正则化的失真。文中提供了一些应用例子,包括有监督学习、半监督学习和无监督学习的例子。理论分析和算法实践显示:逻辑贝叶斯推理比贝叶斯推理在机器学习的大多数方面有更好表现。结尾也讨论了该框架的局限性。

关键词 贝叶斯定理，贝叶斯推理，逻辑概率，Shannon 信道，语义信道，机器学习，多标签分类，最大似然估计，混合模型。

附图摘要——

问：为什么需要逻辑贝叶斯推理？

答：我们需要求概念外延，用它产生新预测！

例 2 x 表示不同年龄，y_1=“成年人”，我们从一个群体得到的先验分布 $P(x)$和后验分布 $P(x|y_1)$. 现在换一个群体，y_1 使用规则不变，先验分布变为 $P'(x)$.

1)假设“成年人”表示年满 x^*岁(未知)，标签都是真的，求“成年人”的外延和适应不同 $P'(x)$的预测模型。

2)求新的后验分布 $P'(x|y_1$ 是真的)。

这个例子对于似然方法和贝叶斯推理同样是难题。

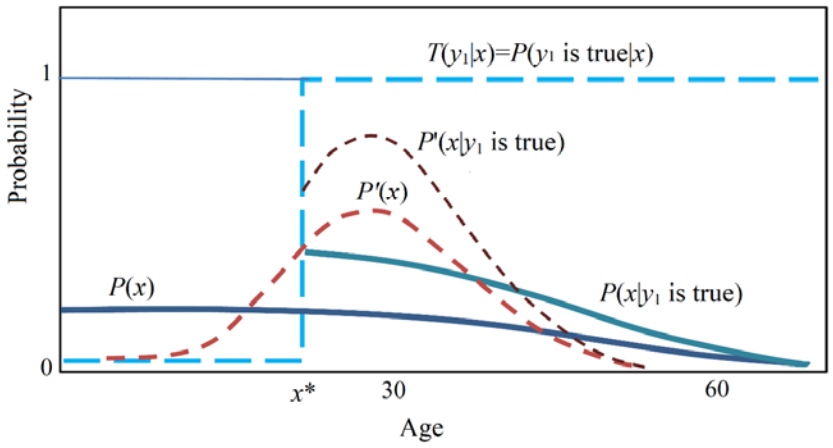


图 1 求 y_1=“成年人”的外延和后验分布 $P'(x|y_1$ is true).

1 引言

贝叶斯推理即贝叶斯主义推理 (Bayesian Inference, 后面缩写为 BI)【1,2】，有人把它翻译为“贝叶斯推断”，或许更准确。为了中文顺口，也为了和大多数翻译一致，我们还是用“贝叶斯推理”或 BI。BI 并不是贝叶斯提出的，而是贝叶斯学派提出的。贝叶斯本人只提出联合概率和条件概率之间的关系【3】，我们常见的贝叶斯公式是拉普拉斯整理的【4】，至于贝叶斯学派也是后来在和频率主义争论中形成的【5,6】。贝叶斯学派和频率学派的主要差别是对概率的认识。频率学派认为概率——包括统计概率和预测的概率(似然度)是事件发生的频率或频率的极限. 贝叶斯主义又分为主观贝叶斯主义和逻辑贝叶斯主义，主观贝叶斯主义者（比如使用 BI 者）认为概率取决于主观信任度，而逻辑贝叶斯主义者认为概率是逻辑真值或逻辑的推广【6】。大多数逻辑贝叶斯主义者(比如 Keynes 和 Carnap)使用真值函数作为推理工具， 但是没有用到样本检验。本文主要工作就是使用样本检验和语义信息方法发展逻辑贝叶斯主义的推理方法，并且反过来用于语义通信和统计学习。

也不是前面有“贝叶斯”字样的方法就属于贝叶斯主义，因为“Bayes’”也翻译成定语“贝叶斯”，比如“Bayes’ theorem”翻译成“贝叶斯定理”。很多贝叶斯方法其实是经典方法或频率主义的方法，比如 Fisher 的似然方法【7】和 Shannnon 信息论【8】中用的贝叶斯方

法。即使用了主观概率，比如贝叶斯信念网，方法也可能还是经典的贝叶斯方法。

Fisher 开创[1]的似然方法用参数 θ 构造的 x(数据、实例、或证据点)的后验概率分布，即似然函数 $P(x|\theta)$，作为假设检验的工具。BI 的核心思想是：假设存在参数的先验概率分布——又叫贝叶斯先验 $P(\theta)$；通过贝叶斯定理，从 $P(\theta)$和预测模型(即一组似然函数)可以推导出贝叶斯后验 $P(\theta|x)$。在假设检验中用 $P(x|\theta)$还是用 $P(\theta|x)$就成了区别频率主义和贝叶斯主义的主要标志。然而，据考察，$P(\theta|x)$最早也是 Fisher 提出的【1】。其实在 Bayes 的文章【3】中就有概率的两种解释：用打赌做例子时使用频率作为概率；但是讲到两个概率之和为 1 时，又使用了逻辑互补概念。可见，两派同源。

本文提出逻辑贝叶斯推理(Logical Bayesian Inference，缩写为 LBI)。它强调逻辑概率，主要为了表达概念的外延或语义. LBI 表面上看是更极端的贝叶斯主义方法，但是它更加兼容频率主义，甚至用频率主义观点解释真值函数(或隶属函数)和逻辑概率。所以 LBI 是激进的贝叶斯主义和激进的频率主义的结合。也许这种结合正是 Bayes 和 Fisher 希望看到的。

Shannon 采用统计概率建立经典信息论，Fisher 等人采用预测概率建立用于假设检验的似然方法。两者都是频率派的成就。然而，似然方法用参数预测客观事件 x 发生的后验概率 $P(x|\theta)$时不便利用先验知识；比如先验知识 $P(x)$改变以后，以前的预测模型即似然函数 $P(x|\theta)$就会失效。为了利用先验知识，也为了强调主观概率，贝叶斯学派利用和似然函数相对称的函数 $P(\theta|x)$或 $P(\theta|\mathbf{X})$(x 是一个实例，$\mathbf{X}$ 是一组实例)——贝叶斯后验——作为推理工具，得到和似然方法不同贝叶斯推理方法，最大似然估计(Maximum Likelihood Estimation， 即 MLE)被改造成最大一个后验估计(Maximum A Posterior, 即 MAP)【9】。虽然 BI 在机器学习领域较之似然方法有自己的优点，特别在样本较小时；然而贝叶斯推理至少存在下面几个问题：

1) BI 声称使用逻辑概率和主观概率【5,6】，但是它实际上没有使用逻辑概率，因而也不能表达语义。
2) 和经典的贝叶斯预测不兼容。
3) 先验知识 $P(\theta)$是主观任意的；而更多的情况下，我们更需要使用客观的先验知识 $P(x)$。
4) 最大后验估计和 Shannon 【8】和 Popper【10】强调的信息准则不兼容，也和 Fisher 的似然准则及流行的正则化最小误差平方(RLS)准则【11】不兼容。

关于这四个问题下节详细解释。

我们提出 LBI 以及基于这一推理的新的数学框架，是为了弥补似然方法和 BI 的缺陷，改进语义通信和机器学习，特别是改进其优化——希望按统一的语义信息准则即对数标准(normalized)似然度准则优化。它是已有数学方法的补充和修正而不是替代。这个框架的所有部件， 比如逻辑概率， 真值函数， 统计概率， 似然函数法….都是已有的；其中大多数公式要么是已有的， 要么是在已有公式基础上做了不多改动的。比如，在此框架中，真值函数被带进贝叶斯公式和信息公式，从而有语义信息公式。这个框架有下面特点：

1) **严格区分并同时使用统计概率和逻辑概率**。还使用两者之间的杂交——主观预测的概率，即似然度。

2) **用真值函数 $T(\theta_j|x)$或语义信道 $T(\theta|x)$取代贝叶斯后验 $P(\theta|x)$**或 $P(y|x;\theta)$。用第三种贝叶斯定理(前两种分别是拉普拉斯使用的和 Shannon 使用的)建立似然函数 $P(x|\theta_j)$和真值函数 $T(\theta_j|x)$之间的简单联系，以及 Shannon 信道和语义信道之间的简单联系. 使得优化的真值函数可以和转移概率函数 $P(y_j|x)$一样，能和 $P(x)$一起做贝叶斯预测, 即产生似然函数 $P(x|\theta_j)$。

3) **把真值函数和似然函数带进 Kullback-Leibler(KL)信息公式和 Shannon 互信息公式**。即用对数标准似然度定义语义信息，使得机器学习数学框架同等于语义通信数学框架，在推广 Shannon 信息理论的同时改进机器学习。改进机器学习体现在几个方面：

a)简化算法，比如简化多标签学习和分类;

[1] 至少是似然方法的继往开来者。

b)提高算法的速度和可靠性，比如用 CM 算法求解最大似然估计；

c)提高预测模型的适应性，比如先验知识 $P(x)$改变时，真值函数作为预测模型仍然适用。

4) **统一优化准则**。用语义信息作为优化机器学习的准则，它等价于最大似然准则，也等价于最小平均相对误差加 log(逻辑概率)准则，从而兼容 RLS 准则。

此外，这一框架把样本序列转换为样本分布——以便用交叉熵和交叉信息评价假设检验，使得方法更加适合样本较大场合。因为这一数学框架用真值函数(条件逻辑概率)表示概念外延即语义，LBI 也可以说是语义贝叶斯推理。这个框架并不能排斥已有的很多数学方法，相反，它试图用真值函数(反映语义)和语义信息准则把已有的很多数学方法结合在一起。这一数学框架还和已有的许多重要思想兼容，比如和贝叶斯，Fisher，Popper, Shannon，Zadeh，Wittgenstein，Tarski，and Davidson 等人的基本思想兼容。

本研究基于作者 20 年前的语义信息论研究【12-14】和最近几年将语义信息论应用到机器学习的研究【15-18】。交叉熵是其中重要工具。虽然交叉熵方法是最近 20 年才流行的【19】， 但是作者在 20 年前就提出交叉熵(当时称之为广义熵)，还提出交互交叉熵(用来定义语义互信息)【20】，并推广 Shannon 信息率失真函数 $R(D)$【8】到 $R(G)$函数(R 是 Shannon 互信息，G 是语义互信息或平均 log(标准似然度)【12-14】)。$R(G)$函数就是两种信息或两种信道的相互匹配函数。让语义信道和 Shannon 信道相互匹配，可以得到一种新的迭代算法，即信道的匹配(CM)算法， 它可以用于最大似然估计【16】(包括不可见实例分类)和混合模型【15】。 通过 $R(G)$函数可以严格证明迭代收敛。CM 算法和流行的 EM 算法【21，22】和 VBEM 算法【23】相比，收敛更快，收敛证明更严格【24】。

本文后面部分组织如下：第二节讨论从贝叶斯定理到似然方法(LM)到贝 BI 的转变，以及为什么需要 LBI 。第三节介绍数学基础，包括逻辑概率和统计概率区别和联系、第三种贝叶斯定理、以及从 Shannon 信道到语义信道的转换。第四节介绍语义信息方法，逻辑贝叶斯推理(即标签学习)、语义信息方法和贝叶斯推理比较以及和 RLS 准则的兼容性。第五节讨论多标签分类(难题之一)和信源 $P(x)$可变时的二分类、以及假设(或标签)的置信水平和确证度。第 6 节介绍 CM 迭代算法用于 Shannon 信道可变时的最大互信息和最大似然估计——包括不可见实例分类(难题之二)和混合模型(难题之三)，包括用 $R(G)$函数严格证明两种迭代收敛. 最后是总结。

2 似然方法→贝叶斯推理→逻辑贝叶斯推理

2.1 转移概率函数、似然方法及先验问题

我们先定义：

- x: 一个数据点或实例，可能的取值是 $x_1, x_2, \ldots, x_m \in U$。大写 X 表示取值为 x 的随机变量。
- y: 一个假设或标签，可能的取值是 $y_1, y_2, \ldots, y_n \in V$. 大写 Y 表示随机变量。
- θ: 预测数据点分布的模型或模型参数。假设它是离散的，相应每个 y_j，θ 取值 θ_j。
- $\mathbf{X}$: 一组实例 $x(1), x(2), \ldots, x(N) \in U$, 构成一个同标签样本。假设这些实例来自 N 个独立同分布随机变量(即 IID 假设)。
- $\mathbf{D}$：一组样例构成的一个带有不同标签的样本$\{(x(t), y(t))\}$，t=1 to N。由 $\mathbf{D}$ 可以得到 N 个不同的条件下的样本 $\mathbf{X}_j$；如果样本很大，从 **D 可以**得到联合概率分布

$P(x,y)$；从 $\mathbf{X}_j$ 可以得到条件样本分布 $P(x|y_j)$或 $P(x|j)$(标签未定时). 本方法严格区分给定和没给定的标签和模型，所以使用下表较多。

我们希望通过过去的样本和新的标签得到新的概率预测，以便决策。Shannon 信道就是很好的概率预测工具。Shannon 信道是一个转移概率矩阵 $P(y|x)$， 其中 $y=y_j$ 的一行 $P(y_j|x)$被称为转移概率函数。根据 x 的先验 $P(x)$和 $P(y_j|x)$，通过贝叶斯公式，可以得到 x 的后验概率分布。即使 $P(x)$变为 $P'(x)$, 转移概率函数还是可以用来做概率预测得到 $P'(x|y_j)$:

$$P'(x \mid y_j) = P(y_j \mid x)P'(x) / \sum_i P(y_j \mid x_i)P'(x_i) \tag{2.1}$$

然而，如果样本不很大，我们得不到连续的转移概率函数，也得不到连续的后验分布. 于是就有了似然方法(Likelihood Method, 简写为 LM)。

Fisher 用 $L(\theta|x)$表示似然函数，也许他本想得到的就是用参数构造的转移概率函数， 但是实际上，他和后来人用的似然函数都是 $P(x|\theta)$或 $P(x|\theta_j)$；用 $L(\theta|x)$显得多此一举。

因为和 BI 比，LM 中模型参数是一个值，而不是一个分布，为了有所区别，我们用 θ_j 表示似然函数中的模型参数：$P(x|\theta_j)$. 给定一个样本 $\mathbf{X}_j$，且在 IID 假设下，θ_j 的似然度为：

$$P(\mathbf{X}_j \mid \theta_j) = P(x(1), x(2), ..., x(N) | \theta_j) = \prod_{t=1}^{N} P(x(t) \mid \theta_j) \tag{2.2}$$

我们把样本写成概率分布(或频率)的形式，则 $P(x_i|y_j)=N_i/N_j$ 表示样本 $\boldsymbol{X_j}$ 有 N_i 个 x_i。那么 $\log P(\boldsymbol{X_j}|\theta_j)$就可以用负的交叉熵表示：

$$\log P(\mathbf{X}_j \mid \theta_j) = \log \prod_i P(x_i \mid \theta_j)^{N_i} = N_j \sum_i P(x_i \mid y_j) \log P(x_i|\theta_j) = -N_j H(x \mid \theta_j) \tag{2.3}$$

假设有 N 个不同条件下的样本 $\mathbf{X}_j$, 其条件概率分布是 $P(x|j)$(标签未定), $j=1,2,\ldots,N$。对于每个条件样本，都可以找到一个优化的模型参数得到最大似然估计(MLE)：

$$\theta_j^* = \arg\max_{\theta_j} P(\mathbf{X}_j \mid \theta_j) = \arg\max_{\theta} \sum_i P(x_i \mid j) \log P(x_i \mid \theta_j) \tag{2.4}$$

它等价于最小相对熵估计【25, 26】:

$$\theta_j^* = \arg\min_{\theta_j} H(P_{x|j} \,\|\, P_{x|\theta_j}) = \arg\min_{\theta_j} \sum_i P(x_i \mid j) \log \frac{P(x_i \mid j)}{P(x_i \mid \theta_j)} \tag{2.5}$$

当 $P(x|\theta_j^*)=P(x|j)$时，相对熵等于 0，模型最优。如相应 θ_j^*的标签是 y_j, 则 $P(x|y_j)=P(x|j)$。

上面求 θ_j^* 就是训练预测模型. 使用模型就是发送信息。

1) 根据似然函数提供概率预测，比如得到消息 $y=y_j$ 时，预测 x 在一定范围内的可能性。比如，我们提供预测：“x 以 x_0 为中心误差不超过 d 的可能性是 90%”。如果 $P(x|\theta_j^*)$是有偏分布，我们可能要提供不同方向上的误差上限。比如，金融风险控制的 VaR(Value at Risk)指标就是根据资本变化的下限提供预测，比如:“最近一周在亏损不超过 100 万(或 5%)的可能性是 95%”
2) 另一种方式是给定 $x=x_i$ 或 $P(x|j)$时，我们选择一个假设或标签 y_j 发送出去。比如，给定 x_i 时, 我们看 $P(x_i|\theta_1), P(x_i|\theta_2),\ldots$哪个大，第 j 个大就选择 y_j 发送. 如果给定一个分布 $P(x|j)$, 我们看哪个 θ_j 导致的相对熵 $H(P_{x|j}||P_{x|\theta j})$小, 选择导致相对熵最小的 y_j 发送。

更复杂的应用是 x 不可见时，我们根据观察条件划分样本。我们不但要优化似然函数，还要优化观察条件空间的划分。后面讨论 Shannon 信道可变的最大似然估计时谈及。

频率学派还提供了置信区间和置信水平分析。上面给定区间的概率预测基于似然函数 $P(x|\theta_j)$，而置信区间和置信水平反映 Shannon 信道特性(噪声少则置信水平高)，和信源无

关。详见 5.3 节。

似然方法的主要缺陷是不能利用先验知识 $P(x)$, $P(y)$或 $P(\theta)$，不适合 $P(x)$变化的场合，不能像转移概率函数那样，在 $P(x)$变化后还能做贝叶斯预测。

比如对于医学检验，从普通人检验数据得到的似然函数可以用来预测有病的概率，但是应用到高危人群就会失效。GPS 类似。我们在大操场上测试车载 GPS 的误差，指示位置是 y_j 而实际位置是 x. 这样可以得到正态分布似然函数 $P(x|\theta_j)$. 用这样的似然函数预测小车的位置，在附近操场上或草原上是没有问题的，但是小车开到高速公路上时，实际位置 x 相对指示的位置 y 就不再是正态分布了(参看后面图 2)。这是因为小车的先验概率分布 $P(x)$和道路有关。而似然方法没有考虑到路况或先验知识 $P(x)$的变化.

贝叶斯推理声称考虑先验知识， 但是它考虑的不是 x 而是 θ 的先验分布 $P(\theta)$。

2.2 贝叶斯推理及其问题

BI 把模型或参数 θ 的先验分布 $P(\theta)$带进贝叶斯公式得到【2】

$$P(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta)P(\theta)}{P_\theta(\mathbf{X})},\ P_\theta(\mathbf{X}) = \sum_j P(\mathbf{X} \mid \theta_j)P(\theta_j) \tag{2.6}$$

这里 $P_\theta(\boldsymbol{X})$不是 $\boldsymbol{X}$ 的先验知识，而是归一化常数，和 $P(\theta)$相关。假设一个先验分布 $P(\theta)$，加上一组似然函数就可以求出 θ 的后验分布。优化预测模型就是, 求使 $P(\theta|\boldsymbol{X})$达最大的 $\theta=\theta_j$*, 这时最大似然估计变为最大后验估计(MAP)：

$$\theta_j^* = \arg\max_{\theta_j} P(\theta \mid \mathbf{X}_j) = \arg\max_{\theta_j}[\sum_i P(x_i \mid j)\log P(x_i \mid \theta_j) + \log P(\theta_j)] \tag{2.7}$$

其中或略了 $P_\theta(\mathbf{X})$，这时因为 $P_\theta(\mathbf{X})$对不同的 θ_j 是相同的。容易看出:

1) 如果忽略 $P(\theta)$或者 $P(\theta)$是等概率分布的时候，它等价于 MLE;

2) 当样本继续增加的时候，$\log P(\theta)$占的比例越来越小，MAP 就越来越接近 MLE.

最大似然方法提供 y_j 的置信水平(正例比例)和置信区间，比如 y_j 在不超过 90%的情况下来自某些 x，这些 x 相对 x_j 误差不超过多少。类似地，BI 可以提供 θ 的后验的置信区间及置信水平【27】，不过这个置信区间和置信水平依赖主观的 $P(\theta)$(后面再议)。

另外，BI 用下面公式预测给定任意样本 **X** 时的 x 的概率分布【2】:

$$P_\theta(x \mid \mathbf{X}) = \sum_j P(x|\theta_j)P(\theta_j \mid \mathbf{X}) \tag{2.8}$$

和预测的 x 的先验概率分布

$$P_\theta(x) = \sum_j P(x|\theta_j)P(\theta_j) \tag{2.9}$$

这里的 $P_\theta(x|\mathbf{X})$和 $P_\theta(x)$都取决于 $P(\theta)$，带有很大主观性，和真实的 x 的先验 $P(x)$或 $P(x|\mathbf{X})$，比如来自大样本的 $P(x)$或 $P(x|\boldsymbol{D})$，可能相差很远。有人还使用超参数 α，即用 $P(\theta|\alpha)$代替 $P(\theta)$【2】。但是，每个地方都加 α 和每个地方都不加是一样的。我们也可以假定，$P(\theta)$已经包含了 α 的信息，如果没有任何限制，$P(\theta)$就是等概率的。

还有一种贝叶斯后验是 Y 的贝叶斯后验：

$$P(Y \mid \mathbf{X}, \theta) = \frac{P(\mathbf{X} \mid \theta)P(Y)}{P_\theta(\mathbf{X})},\ P_\theta(\mathbf{X}) = \sum_j P(\mathbf{X} \mid \theta_j)P(y_j) \tag{2.10}$$

这两种后验通常不被区分，但是实际上是不同的。主要是因为 $P(Y)$是标签空间上的概率分布，可能是客观的。如果可选标签有 n 种，那么概率 $P(y_1), P(y_2)\cdots$就有 n 个。而 $P(\theta)$

是参数空间上的概率分布，参数空间的点可能有 n^2 个甚至更多。机器学习比如分类中，Y 的贝叶斯后验更加常见。哲学界在使用 BI 时，常用 P(H)做贝叶斯先验，即本文的 $P(Y)$，而不用 $P(\theta)$.

Y的贝叶斯后验容易理解， 因为如果 x 可变，它就是转移概率函数 $P(y_j|x)$或 Shannon 信道的带参数形式。而 θ 的贝叶斯后验只有在相应参数空间每个点 θ 有一个相应 Y的时候才好理解；否则不好理解。

下面我们分析贝叶斯推理存在的问题。

1） 关于逻辑概率

用不用逻辑概率或真值函数关系到能不能表达概念的外延或语义。贝叶斯主义声称贝叶斯概率是逻辑概率，然而实际上 BI 中并没有使用逻辑概率。理由是：BI 使用的 $P(\theta)$和 $P(\theta|X)$都是归一化的，而逻辑概率不是归一化的。考虑天气预报“明天无雨”，“明天有雨”，“明天有小雨”， “明天有小到中雨”，...的逻辑概率，它们之和大于 1 而不是等于 1。条件逻辑概率就是真值函数或隶属函数，取值于[0,1]，它反映假设或标签的语义，更不是归一化的。而 Bayesian Inference 并没有使用真值函数，因而也不能解决语义问题。模糊数学使用的隶属函数可以当作真值函数，从而反映语义。所以，我们需要一种推理，能从样本分布得到隶属函数或真值函数。

看来 BI 比频率主义更关注频率问题。从介绍的 BI 的书籍所举例子看，所有过硬的例子中，先验概率分布 $P(\theta)$都是关于频率发生器的频率，比如不同骰子的可能性分布，机器编号上限的可能性分布【27】。所以，BI 擅长的是频率问题，而不是逻辑分类问题或预测问题. 模糊预测也可以看作提供模糊类别，也是逻辑分类问题。用 BI 解决分类问题也要求不同类别相互排斥，而事实上大多数情况下，类别有很多，不同类别之间存在蕴含关系或交集。这时候用 BI 解决问题就非常困难。这也是为什么用贝叶斯推理解决二类别分类比较成功，而解决多类别分类比较困难。

2） 关于贝叶斯预测

首先，从 BI 得到的 $P_\theta(x)$依赖于 θ。而当样本 **D** 很大时，预测的 x 的分布应该等于从大样本统计得到的概率分布 $P(x)$，而 $P_\theta(x)$不能保证。第二，用经典贝叶斯方法，当先验分布 $P(x)$变为 $P'(x)$时，用经典贝叶斯方法，我们仍然可以从转移概率函数 $P(y_j|x)$得到后验分布 $P'(x|y_j)$. 但是使用 BI，我们并不能得到类似于转移概率函数的函数，也得不到接近 $P'(x|y_j)$的概率预测 $P'(x|\theta_j)$。

所以，为了兼容传统的贝叶斯预测，我们需要一种带参数的转移概率函数做贝叶斯预测，使得预测结果接近经典贝叶斯方法得到的结果.

3） 关于先验知识

在 BI 中，先验分布 $P(\theta)$是主观的。而人类使用的先验知识通常是 $P(x)$，它是比较客观的，和参数无关。BI 使用了 θ 的先验 $P(\theta)$和似然函数 $P(x|\theta)$, 得到的 $P_\theta(x)$也是主观的。 而我们需要客观的先验知识。比如，HIV 检验中，要预测阳性者带有 HIV 的概率，我需要知道基础概率 $P(x)$, 以及对于不同人群基础概率的变化。使用 GPS 时，为了能根据 GPS 箭头预测小车实际位置，我们需要知道标注了各种道路和建筑物的地图，其中包含了先验知识 $P(x)$。

4） 关于优化准则

在 BI 中，先验概率 $P(\theta_j)$大的 y_j 容易被选择，这不符合信息准则。因为按照信息准则，先验逻辑概率越小，如果经得起检验，信息量越大，如 Popper 所说【10】；按 Shannon 信息论类似，y_j 先验概率越小, 对信息贡献越大。所以我们需要考虑先验逻辑概率大和信息量小之间的矛盾。

后面我们还会讲到 BI 推理的其他问题。

和似然方法比， BI 也有一些优点， 比如：

1)它考虑先验知识，如果 $P(\theta)$是根据 $P(x)$产生的(比如对于医学检验和 GPS)，那么 $P(\theta)$也包含关于 x 的信息，在样本较小时有优势。

2)它便于把现在的后验 $P(\theta|\mathbf{X})$变成下一步的先验 $P(\theta)$，使得上一步的样本信息积累在下一步的先验中。

3)$P(\theta|\mathbf{X})$在 θ 空间的分布范围随样本尺寸增大而缩小，当样本很大时，它就逐渐收缩到一点，这一点就是由 MAP 或 MLE 得到的 θ_j*. 所以，BI 将模型学习、模型优化和判决融为一体了, 比较直观。后面讨论分类问题时将说明，这是优点也是缺点。

2.3 从两个简单例子看为什么需要逻辑贝叶斯推理

例 1 对于 HIV 病毒检验，y_1=阳性+，$y0$=阴性-；x_1=有 HIV，$x0$=没有 HIV。对于普通人群，先验概率 $P(x_1)$=0.002. 检验呈阳性的人，后验概率是 $P(x_1|y_1)$=0.5. 现在测试人群变为很多不同人群, 其中之一是男性同性恋者人群，$P'(x_1)$=0.1。请问：1)$P'(x_1)$=0.1 时，显示阳性的人带 HIV 的概率是多少？2)能不能得到一个预测模型用于不同人群？

用似然方法，由普通人群检验数据，我们得到 $P(x_1|\theta_1)$=0.5. 但是换成高危人群，先验分布 $P(x)$变了，以前得到的似然函数 $P(x_1|\theta_1)$=0.5 就失效了。所以似然方法不能利用新的先验知识，用来解决这个问题无效。.

如果用 BI，也看不出有什么办法。即使假设 $P(y_1)$=$P'(x_1)$=0.1, 还是不行；需要补充 $P(x_1|\theta_0)$才能得到贝叶斯后验。

然而，用 LBI 可以很容易解决这个问题。用经典的贝叶斯公式也能解决，但是要经过一番推导(不赘)。推导结果和 LBI 的结果相同(见 5.4 节)。

例 2 x 表示不同年龄，y_1=“成年人”，我们从一个群体得到的先验分布 $P(x)$和后验分布 $P(x|y_1)$. 现在换一个群体，y_1 使用规则不变，先验分布变为 $P'(x)$.

1)假设“成年人”表示年满 x*岁(未知)，标签都是真的，求“成年人”的外延和适应不同 $P'(x)$的预测模型。

2)求新的后验分布 $P'(x|y_1$ 是真的)。

这个例子对于似然方法和贝叶斯推理同样是难题。

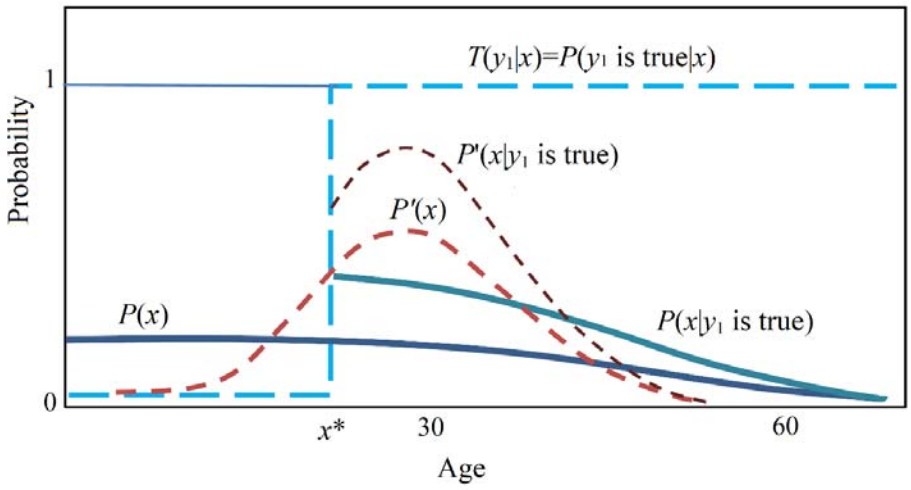


图 1 求 y_1=“成年人”的外延和后验分布 $P'(x|y_1$ is true).

Figure 1 Solving the denotation of y_1="x is adult" and $P'(x|y_1$ is true).

如果画出 $P(x|y_1)$的分布图，人眼看过后，人脑想一下就知道，可以首先求出“成年人”外延(即 y_1 的真值函数)，外延就是适合不同 $P'(x)$的预测模型。然后把外延内的 $P'(x)$归一化就得到：

$$P(x|y_1 \text{ is true}) = \begin{cases} P'(x) / \sum_{x \geq x^*} P'(x), & x \geq x^*; \\ 0, & x < x^*. \end{cases} \tag{2.11}$$

然而， 到目前为止，已有的数学方法中，还没有简单的求概念外延的方法。如果“成年人”不是按法定年龄说的，而是来自自然语言，即成年人的集合是模糊的，又怎么解决呢？这就是 LBI 首先要解决的问题。

3 新框架基础：两种概率和三种贝叶斯定理

3.1 区分并同时使用统计概率和逻辑概率

首先澄清下面几种概率和相关概念：

- 统计概率——即通常说的概率，是频率的极限，是客观的。
- 逻辑概率——假设为真的概率；给定实例的条件逻辑概率就是假设的真值。
- 可能性，似然度——主观预测的概率。
- 信任度(degree of belief)——贝叶斯主义者称贝叶斯概率就是信任度；虽然有的贝叶斯主义者(比如 Jaynese)认为它是逻辑概率【3】，但是 BI 实际用的 $P(\theta)$是主观预测的频率。信任度还有一种用法和下面确证度相关。
- 确证度(confirmation measure)或可信度——归纳支持程度【17,28】。因为归纳支持程度可能是负的(证据支持其否定)，确证度在-1 和 1 之间。所以这种用法已经超出概率范围了。
- 置信水平(confidence level)【29】——就是一个假设的正例所占所有例子(正例加反例)的比例， 也在 0 和 1 之间变化。它和确证度类似，不具有概率性质。

综上所述， 概率包括统计概率、逻辑概率和主观预测的概率。笔者认为这三种概率同时存在，不需要各执一端。而主观预测的概率可以看成统计概率和逻辑概率的杂交， 所以基本的概率是两种：统计概率和逻辑概率。

一个假设或标签有两种概率，一是它被选择的概率——统计概率，二是它被判定为真的概率——逻辑概率。两者通常是不同的。 比如， 考虑 “明天有小雨”和“明天有小到大雨”… 后者逻辑概率较大， 而被选择概率较小。考虑“他是老年人”和“他不是老年人”，后者逻辑概率大于前者，但是它被选择的概率小于前者。一个永真句“他可能是也可能不是老年人”，其逻辑概率更大，是 1， 而选择概率接近 0.

标签被选择的概率是归一化的，即所有标签被选择的概率相加等于 1，**而标签的逻辑概率不是归一化的**。比如前面提到的天气预报语句。再比如，考虑描述人的年龄的标签：“小孩”、“年轻人”、“中年人”、“成年人”、“老年人”… 它们的被选择的概率之和为 1，但是逻辑概率之和远大于 1. 原因在于，一个年龄可能使几个标签为真。 比如, 年龄是 25 岁，标签“年轻人”、“成年人”、“非老年人”…都是真的。

时钟指针可以看作是时间的标签，共有 S=12*60*60 种标签，一种标签出现的统计概率是 1/S。 设指针偏差不超过 1 分钟算是正确的，那么一个指针位置的逻辑概率就是 2*60/S, 是相应的统计概率的 120 倍。GPS 指针，温度表，秤…的指针或读数是类似的，

逻辑概率远大于统计概率。如果指针的真假在 0 和 1 之间，结论类似。后面谈及。

数学家和哲学家都定义过概率的公理系统【30,31】，这些系统都采用归一化的概率，也没有同时兼用两种概率，于是导致频率学派和贝叶斯学派之争。Zadeh 定义的模糊集合的隶属函数【32】可以用作真值函数，表达自然语言的语义；他定义的模糊事件的概率【33】可以用做逻辑概率。但是所定义的隶属函数和模糊事件的概率却没有和统计概率有机地结合在一起。下面我们定义并存且相关的两种概率， 并通过第三种贝叶斯定理把统计概率和隶属函数(反映概念外延)联系起来，使得转移概率函数和隶属函数可以相互转换。

定义 2.1 设论域 U 中有元素 $x_1, x_2, \ldots, x_m$; X 取值于 U 中某个元素 x 的随机变量(按照信息论习惯，随机变量用大写字母)，即 $X\in U=\{x_1, x_2, \ldots, x_m\}$. 再设论域 V 中有元素 $y_1, y_2, \ldots, y_n$; Y 是取值于 V 中某个元素 y 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 2.2 我们用等号“=”表示的随机事件的统计概率(简称概率)——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 $P(x)$或 $P(X)$表示 $P(X=x)$. 同样地，我们用 $P(y)$或 $P(Y)$表示 $P(Y=y)$. **我们再用属于符号“∈”表示随机事件的逻辑概率**。比如 $P(X\in A_j)$是逻辑概率。

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【34】，$P(X\in A_j)=P$("$X\in A_j$"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 $P(y_j)$，即 yj 被选择的概率，和逻辑概率 $P(y_j$ 是真的)**. 为了更清楚区分两者，后面我们用 $T(A_j)$或 $T(yj)$表示 y_j 的逻辑概率，即

$$T(A_j)=T(yj)= P(y_j \text{ 是真的})= P(X\in A_j) \tag{3.1}$$

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

$$T(A_j)=\sum_i P(x_i)T(A_j \mid x_i) \tag{3.2}$$

根据 Davidson 的真值条件语义学【35】，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

$$P(y_1)+P(y_2)+\ldots+P(y_n)=1，\ P(y_1|x_i)+P(y_2|x_i)+\ldots+P(y_n|x_i)=1 \tag{3.3}$$

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，

$$T(A_1)+T(A_2)+\ldots+T(A_n)\geqslant 1 \tag{3.4}$$

只有在$\{A_1, A_2, \ldots, A_n\}$是 U 的划分并且 y 总是被正确地使用的情况下，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)是构成 Shannon 信道的转移概率函数，可用作贝叶斯预测，产生 x 的后验概率分布 $P(x|yj)$. 它也不是归一化的，即一般情况下

$$P(y_j|x_1)+P(y_j|x_2)+\ldots+P(y_j|x_m)\neq 1 \tag{3.5}$$

后面用 θ_j 表示相应 y_j 的模糊集合(包括清晰集合)，则相应的逻辑概率是 $T(\theta_j)$. 逻辑概率分布——即后面的 $T(\theta)$——虽然很像流行的贝叶斯推理中的 $P(\theta)$， 但是，$T(\theta)$既不是横向归一化的也不是纵向归一化的，它本身就是归一化系数。真值函数或隶属函数 $T(\theta_j|x)$的最大值是 1。可以说它是纵向归一化的. 即：

$$\max(T(\theta_j|x_1), T(\theta_j|x_2), \ldots, T(\theta_j|x_m))=\max T(\theta_j|X)=1 \tag{3.6}$$

这一重要性质将给求解真值函数带来方便.

3.2 三种贝叶斯定理

第一种贝叶斯定理是关于两个逻辑概率和条件逻辑概率之间关系的定理。相应公式形式是拉普拉斯根据贝叶斯思想整理出来的【4】。

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

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)=T(A|B)T(B)+T(A|B')T(B') \tag{3.7}$$

类似地， 也可对称地求出 $T(A|B)$。如果$\{B_1, B_2, \ldots, B_K\}$构成 U 的一个划分，则

$$T(A)=\sum_{k=1}^{K} T(A\mid B_k)T(B_k) \tag{3.8}$$

第二种贝叶斯定理是 Shannon 使用的关于两个统计概率之间关系的定理。

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

$$P(x\mid y_j)=P(x)P(y_j\mid x)/P(y_j),\ P(y_j)=\sum_i P(x_i)P(y_j\mid x_i) \tag{3.9}$$

类似地，也可以对称地求出 $P(y_j|x)$.

之所以说上面两个定理是不同定理，是因为贝叶斯定理 1 中随机变量和论域只有一个，而贝叶斯定理 2 中随机变量和论域都是两个。贝叶斯定理 3 是笔者提出的，是关于一个统计概率和一个逻辑概率之间关系的定理； 随机变量是一个， 概率有三种：统计概率、 逻辑和两者的杂交——预测的概率或似然度.

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

$$P(x\mid A_j)=P(x)T(A_j\mid x)/T(A_j),\ T(A_j)=\sum_i P(x_i)T(A_j\mid x_i) \tag{3.10}$$

$$T(A_j\mid x)=T(A_j)P(x\mid A_j)/P(x),\ T(A_j)=1/\max[P(x\mid A_j)/P(x)] \tag{3.11}$$

这个定理包含的两个公式是不对称的， 所以两个都要写出来。解释：(3.10)中 $P(x|A_j)$是似然函数；$T(A_j)$是 $P(x|A_j)$的横向归一化系数. 而在(3.11)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。max[ ]表示其中函数最大值。 (3.11)就是求概念外延公式(参看 2.3 节中例 2 和图 1)， 不过它还考虑到模糊概念.

如果仿照**贝叶斯定理 1 和 2**， (3.11)中应有

$$P(x)=\sum_j T(A_j)P(x\mid A_j) \tag{3.12}$$

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. **所以上式是不成立的！** 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 $P(X=x，x\in A_j)$，则

$$P(X=x, x\in A_j)=P(X=x|\, x\in A_j)P(x\in A_j)=P(x|A_j)T(A_j)$$
$$P(X=x, x\in A_j)=P(x\in A_j\mid X=x)P(X=x)=T(A_j|x)P(x)$$

于是有

$$P(x\mid A_j)=P(x)T(A_j\mid x)/T(A_j),\ T(A_j|\mathrm{x})=T(A_j)P(x\mid A_j)/P(x)$$

因为 $P(x|A_j)$是横向归一化的，所以 $T(A_j)=\Sigma_i P(x_i)\ T(A_j|x_i)$. 因为 $T(A_j|x)$是纵向归一化的，把(3.6)代入上式，可以得到

$$1=\max[T(A_j)P(x|A_j)/P(x)]=T(A_j)\max[P(x|A_j)/P(x)]$$

所以 $T(A_j)=1/\max[P(x|A_j)/P(x)]$。**证毕.**

贝叶斯定理 3 的第二个公式也可以直接写成：

$$T(A_j\mid x)=[P(x\mid A_j)/P(x)]/\max[P(x\mid A_j)/P(x)] \tag{3.13}$$

这就是通过 x 的先验和后验求概念外延、真值函数或隶属函数的公式。贝叶斯定理 3 就可以用来求解 2.3 节中的两个例子。

信息论中习惯用大写表示随机变量，下面的概率和熵中，我们用 X 取代 x, 用 Y 取代 y. 图 2 直观显示了 $T(A_j|X)$，$P(X|A_j)$和 $P(X)$三者之间的关系.

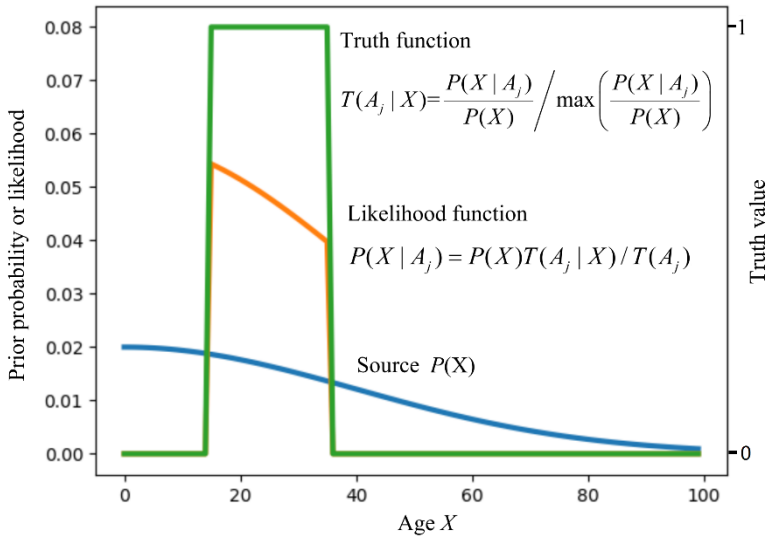


图 2 贝叶斯定理 3 中 $T(A_j|X)$，$P(X|A_j)$和 $P(X)$之间的关系

Figure 2 Relationships between $T(A_j|X)$，$P(X|A_j)$ and $P(X)$ in Bayes' Theorem 3

我们可以这样产生概率分布等于 $P(X|A_j)$的样本：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.

我们假设一个模糊集合 θ_j 的隶属函数 $T(\theta_j|X)$是用参数构造的，也像 $P(y_j|X)$一样，可以用于概率预测，于是 θ_j 同时也是一个预测模型或模型参数(矢量). 我们再用 θ 表示 θ_1 , θ_2, …, θ_n 中任何一个。我们也可以认为 θ 是一个预测模型，θ_1 , θ_2, …, θ_n 是相应的子模型。我们再令 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，这里的 $P(X|\theta_j)$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 $\mathbf{X}_j$，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.

贝叶斯定理 3 推广到集合模糊的场合就是

$$P(X \mid \theta_j)=P(X)T(\theta_j \mid X)/T(\theta_j),\quad T(\theta_j)=\sum_i P(x_i)T(\theta_j \mid x_i) \tag{3.14}$$

$$T(\theta_j \mid X)=[P(X \mid \theta_j)/P(X)]/\max[P(X \mid \theta_j)/P(X)] \tag{3.15}$$

证明同样，不赘。

这里 $T(\theta_j|X)$类似于带参数的 Y 的后验 $P(y_j|X;\theta)$，差别是 $T(\theta_j|X)$是纵向归一化的。

3.3 从 Shannon 信道到语义信道——大样本非参数逻辑贝叶斯推理

逻辑贝叶斯推理(LBI)就是从样本或统计概率分布得到真值函数，或者说从 Shannon 信道得到语义信道。本节讨论大样本时的 LBI，第 3 节讨论样本不大时，用参数构造真值函数的 LBI。

Shannon 信息论【8】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：

$$P(Y|X)\Leftrightarrow\begin{bmatrix}P(y_1|x_1) & P(y_1|x_2) & ... & P(y_1|x_m)\\ P(y_2|x_1) & P(y_2|x_2) & ... & P(y_2|x_m)\\ ... & ... & ... & ...\\ P(y_n|x_1) & P(y_n|x_2) & ... & P(y_n|x_m)\end{bmatrix}\Leftrightarrow\begin{bmatrix}P(y_j|X)\\ P(y_j|X)\\ ...\\ P(y_n|X)\end{bmatrix} \tag{3.16}$$

其中双向箭头表示等价. Shannon 称其中一行 $P(y_j|X)$为转移概率函数(y_j不变X变). 所以，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，我们可以用它和 $P'(X)$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

$$\frac{P'(X)kP(y_j|X)}{\sum_i P'(x_i)kP(y_j|x_i)}=\frac{P'(X)P(y_j|X)}{\sum_i P'(x_i)P(y_j|x_i)}=P'(X|y_j) \tag{3.17}$$

现在我们定义语义信道。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：

$$T(\theta|X)\Leftrightarrow\begin{bmatrix}T(\theta_1|x_1) & T(\theta_1|x_2) & ... & T(\theta_1|x_m)\\ T(\theta_2|x_1) & T(\theta_2|x_2) & ... & T(\theta_2|x_m)\\ ... & ... & ... & ...\\ T(\theta_n|x_1) & T(\theta_n|x_2) & ... & T(\theta_n|x_m)\end{bmatrix}\Leftrightarrow\begin{bmatrix}T(\theta_1|X)\\ T(\theta_2|X)\\ ...\\ T(\theta_n|X)\end{bmatrix} \tag{3.18}$$

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. 温度表，秤，GPS…也提供语义信道，它们的读数的真值函数可以用没有系数的高斯分布表示，其中标准差反映精度。而它们的 Shannon 信道是有噪声信道，噪声大小取决于设备制造的技术水平和使用环境。

由式(3.17)可知，当真值函数和转移概率函数成正比时，语义贝叶斯预测和经典贝叶斯预测——用 $P(y_j|X)$和 $P(X)$产生 $P(X|y_j)$——等价. 所以我们可以从 Shannon 信道得到与之等价的语义信道：$T(\theta_j|X)\propto P(y_j|X)$($j$=1，2，…)。令 $T(\theta_j|X)$的最大值是 1， 于是有

$$T^*(\theta_j|X)=P(y_j|X)/\max[P(y_j|X)],\ j=1,\ 2,\ \ldots,\ \mathrm{n} \tag{3.19}$$

再根据**贝叶斯定理 2**，我们得到优化的真值函数的数值解

$$T^*(\theta_j|X)=[P(X|y_j)/P(X)]/\max[P(X|y_j)/P(X)] \tag{3.20}$$

式(3.20)比(3.19)更加实用，因为求 $P(y_j|X)$需要 $P(y_j)$, 后者通常很难得到。上面两个公式适合大样本场合。当样本不大时，我们需要用参数构造真值函数，用语义信息准则得到 $T(\theta_j|X)$的参数解。后面谈及。

汪培庄教授用随机集落影(即集值统计)定义隶属函数【36】，下面证明，用(3.19)得到的真值函数 $T(\theta_j|X)$作为隶属函数，和集值统计结果一致。

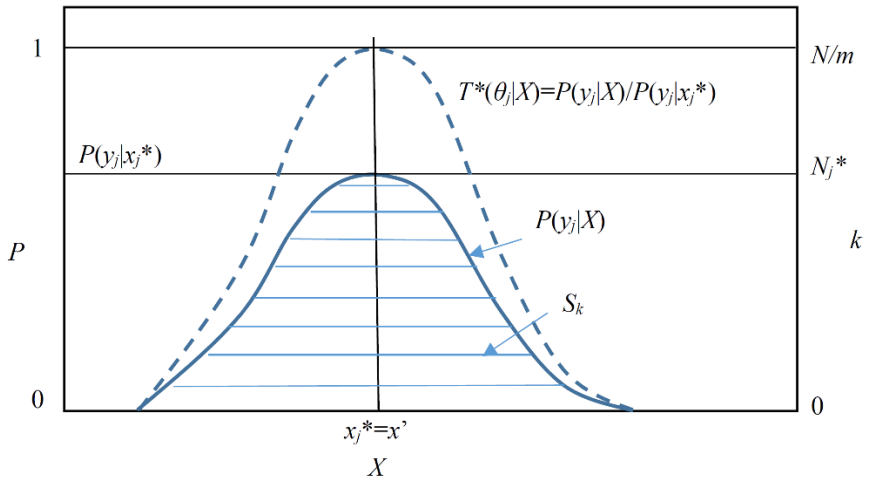


图 3 贝叶斯定理 3 和模糊集合的随机集落影理论兼容

Fig.3 The Bayes' theorem 3 is compatible with the random sets falling shadow theory about fuzzy sets

假设一个 Shannon 信道由一个尺寸为 $N\to\infty$的大样本 **D** 产生，其中 X 是等概率的，包含每个 x_i 的样例(x_i; Y)有 N/m 个. 我们选出所有包含 y_j 的样例，设这些样例中含有实例 x'的样例最多，我们记 x_j*=x'，它有 N_j*个。那么，我们可以通过样例合并，合并成 N_j*个多实例单标签样例——比如 S_k=(x_{k1}, x_{k2},…; y_j)， k=1,2,…,N_j*，每个样例包含 x_j*。S_k 就可以看做随机集合的一个取值，即一个集值。设第 k 个集值的特征函数是 $F_k(X)$, 则根据随机集落影理论，我们可以得到隶属函数：

$$m_{\theta_j}(X)=\frac{1}{N_j*}\sum_{k=1}^{N_j*}F_k(X) \tag{3.21}$$

根据经典统计我们也可以得到转移概率函数

$$P(y_j\mid X)=\frac{1}{(N/m)}\sum_{k=1}^{N_j*}F_k(X) \tag{3.22}$$

比较两者， 可得

$$m_{\theta_j}(X)=P(y_j\mid X)/[N_j*/(N/m)]=P(y_j\mid X)/\max[P(y_j\mid X)]=T*(\theta_j\mid X) \tag{3.23}$$

如果 X 不是等概率的，我们可以筛选出等概率的样例。因为转移概率函数独立于信源 $P(X)$，信源是否等概率不会影响这样求出的隶属函数。

3.4 理解 GPS 定位——提供似然函数还是真值函数？

考虑全球定位系统(GPS). GPS 精度通常用 RMS(Root Mean Square)即标准差表示。根据 RMS 的定义，GPS 指示 $Y=\hat{X}$ 和实际位置 $X=x_i$ 之间的关系由条件概率函数表示：

$$P(Y\mid x_i)=K\exp[-|\hat{X}-x_i|^2)/(2d^2)],\quad j=1,2,\ldots,n \tag{3.24}$$

其中 d=RMS，K 是归一化系数， $|\hat{X}-x_i|$是实际位置和预测位置之间的距离。因为误差是

对称的，所以也有转移概率函数和相应的 Shannon 信道——标准差为 d 的有噪声高斯信道：

$$P(y_j \mid X)=K\exp[-|\hat{x}_j - X|^2)/(2d^2)], \quad j=1, 2, \ldots, n \tag{3.25}$$

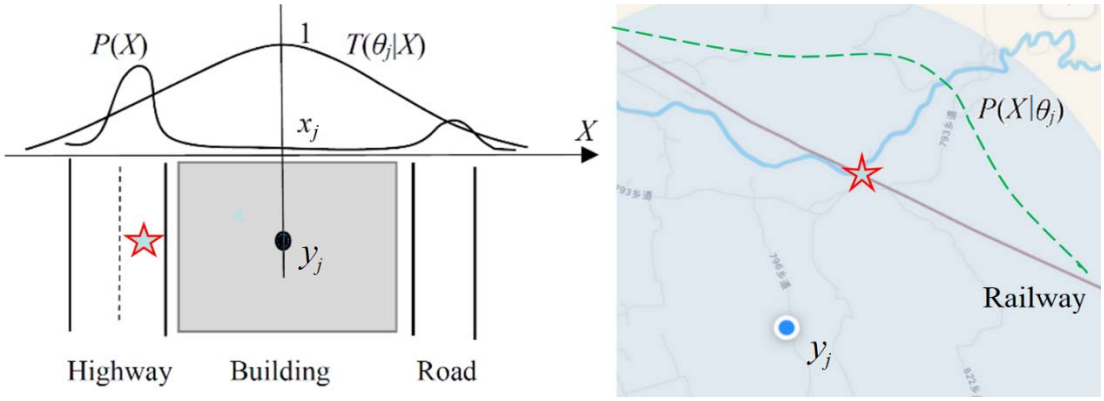
现在考虑 GPS 显示屏上的定位(箭头或小圆圈)的含义。由定位 y_j 和 d 我们能得到上述转移概率函数吗？回答是否定的。因为我们不知道 $P(y_j|X)$的最大值。但是认为定位提供真值函数就不需要知道 $P(y_j|X)$的最大值。我们可以把定位理解为 y_j="$X\approx x_j$"，则有真值函数

$$T(\theta_j|X)=\exp[-|X-x_j|^2/(2d^2)], \quad j=1, 2, \ldots, n \tag{3.26}$$

我们称之为高斯分布真值函数。这样一组真值函数(对于不同的 y_j)就构成一个语义信道 $T(\theta|X)$. 一个时钟、一个秤、一个温度表，它们的指针和读数含义类似. 具有这样含义的 y_j 可谓无偏估计，都可以用上式表示。 真值函数也可以理解为 X 和 x_j 之间的相似度函数，相似度和距离有关。根据真值函数预测的 X 的后验分布是

$$P(X \mid \theta_j)=\frac{P(X)\exp[-|X-x_j|^2/(2d^2)]}{\sum_i P(X)\exp[-|X-x_j|^2/(2d^2)]} \tag{3.27}$$

或许有人认为 GPS 定位 y_j 提供了似然函数 $P(X|\theta_j)$或贝叶斯后验 $P(\theta|x_j)$. 下面我们用两个例子说明这是不对的。考虑 GPS 定位的特殊环境如图 4(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 4 GPS 定位图解. 圆点是指针位置，五角星是小车或人的最为可能的位置. (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 4 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.

如果认为 y_j 提供了似然函数 $P(X|\theta_j)$或贝叶斯后验 $P(\theta|x_j)$, 那么我们会预测小车在楼顶上的概率最大。然而，常识告诉我们这是不对的. 根据(3.27)，小车在高速公路上(如五角星所示)的概率最大——这是符合常识的.

再看图 4(b), 人在高铁列车上，$P(X)$在铁路线上是等概率的。如果认为 y_j 提供了似然函数 $P(X|\theta_j)$或贝叶斯后验 $P(\theta|x_j)$, 则人在圆点处概率最大。根据贝叶斯定理 3，似然函数如虚线所示，五角星表示最为可能位置. 上述结论也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 4 还显示，我们可以利用先验概率分布 $P(X)$纠正 y_j 的偏差。后面说明通过优化语义信道也可以系统地纠正 Y 的偏差.

4 结合 Shannon 和 Fisher 理论的语义信息方法

4.1 把似然函数和真值函数带进信息公式

已有不少文献关于语义信息研究【37-40】，但是把隶属函数和由它产生的似然函数带进语义信息公式是笔者的创见。笔者早在 1990 的文章【20】中就提出广义熵和广义互信息(即语义互信息)。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. 后来其他作者也提出广义熵，并称之为交叉熵【19】. 所以下面语义互信息也可以称之为交叉互信息. 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.

y_j 提供关于 x_i 的语义信息(量)被定义为对数标准(normaized)似然度【12,13】:

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{4.1}$$

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 5 所示.

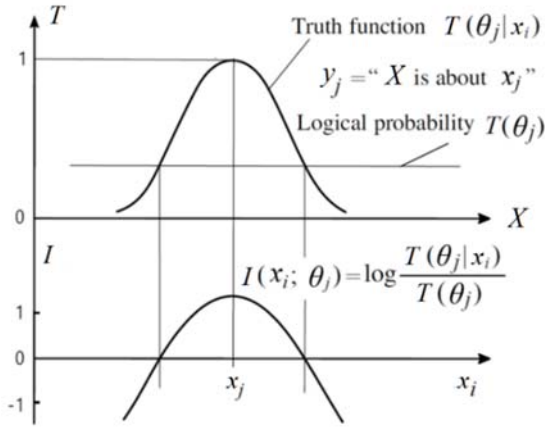


图 5 语义信息量图解. 偏差越大，信息越少；逻辑概率越小，信息量越大；错误预测提供负的信息.

Figure 5 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【10】: (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.

把高斯分布真值函数代入式(4.1), 就得到

$$I(x_i;\theta_j)=\log[1/T(\theta_j)]-\left|X-x_j\right|^2/\left(2d_j^{\,2}\right) \tag{4.2}$$

其中 $\log[1/T(\theta_j)]$ 就是 Bar-Hillel 和 Carnap 定义的语义信息测度【37】. 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. 从图 5 容易理解语义信息准则类似于 Regularized Least Square (RLS)准则【41】。标准偏差 d_j 增大了, $T(\theta_j)$也会增大，信息量就会减少。 但是标准偏差太小，误差项又太大。4.6 节详谈。

对 $I(x_i; \theta_j)$求平均，就得到广义 Kullback-Leibler (KL)信息【42】:

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{4.3}$$

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，就得到广义互信息或语义互信息公式：

$$\begin{aligned}I(X;\theta)&=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\\ &=H(X)-H(X\mid\theta)=H(\theta)-H(\theta\mid X)\end{aligned} \tag{4.4}$$

其中 H(X)是 Shannon 熵， 其他三种熵是交叉熵：

$$\begin{aligned}&H(X)=-\sum_i P(x_i)\log P(x_i),\ H(X\mid\theta)=-\sum_j\sum_i P(x_i,y_j)\log P(x_i\mid\theta_j)\\ &H(\theta)=-\sum_j P(y_j)\log T(\theta_j),\ H(\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i)\end{aligned} \tag{4.5}$$

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，或真值函数正比于转移概率函数时，即 $T(\theta_j|X)\propto P(y_j|X)$(对于所有 j)时，上述广义 KL 信息达到其上限——KL 信息；语义互信息也达到其上限——Shannon 互信息.

Akaike 揭示了似然度和 KL 信息之间的联系【25】. 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 $P(x_i|y_j)= N_{ij}/N_j$. 因此就有 log(标准似然度)和广义 KL 信息之间关系：

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_j I(X;\theta_j) \tag{4.6}$$

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：

$$\sum_j\frac{N_j}{N}\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=I(X;\theta) \tag{4.7}$$

因为优化模型 θ_j 时 $P(X)$不变，所以最大语义互信息准则等价于最大似然准则.

4.2 语义通信模型

通信系统中分信号发送者和信号接受者。从语义通信的角度看机器学习，也分标签发送者和标签接收者。接受者通过样本学习得到语义，即真值函数或逻辑分类函数——一个实例可能隶属于多个类别。而发送者划分实例空间给实例分类，每个实例只属于一个类别——可能带有多个标签或一个复合标签。详见图 6 通信模型。

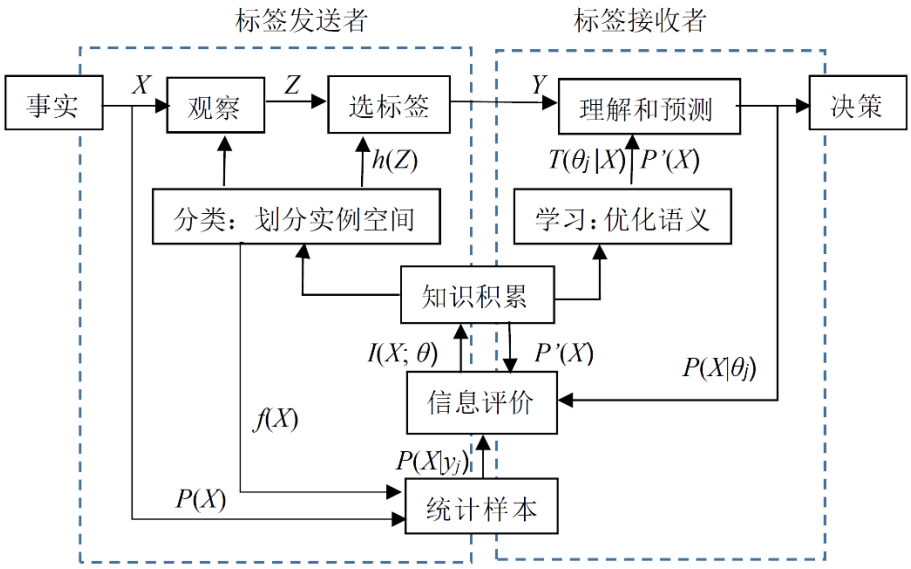


图 6 用于机器学习的语义通信模型。在检验、估计、预测和分类时，信息来自样本分布 $P(X|y_j)$对似然函数 $P(X|\theta_j)$的检验，收信人理解和发送人分类在语义信息准则要求下不断优化。

Figure 6. Semantic communication model for machine learning. In tests, estimations, and predictions, information comes from that the sampling distribution examins the likelihood function; the receiver's undstanding and the sender's classification are continuously optimized with maximum semantic information criterion.

其中假设发送者根据观察条件 $Z\in C$ 发送标签 $Y=h(Z)$。实际情况也可能 $Z=X$, 划分函数是 $Y=f(X)$， 这时就是有监督学习。当 $Z\neq X$ 且 C 的划分 $h(Z)$不确定时，学习就是半监督学习。如果先有 $P(X)$, 需要通过学习产生 $P(y_j|X)$和 $T(\theta_j|X)$, 学习就是无监督学习。其中 $P'(X)$是主观预测的 X 的先验分布。一般情况下可以假设 $P'(X)=P(X)$。

在 BI 中，学习和分类是都由贝叶斯后验 $P(\theta|X)$决定的. 在 LBI 中，学习和分类是分开的，学习得到真值函数 $T(\theta_j|X)$或语义信道 $T(\theta|X)$, 它作为预测模型，在信源 $P(X)$变化后仍然适用。信源变化不改变收信人的逻辑分类，但是改变发信人的选择分类。从第 5 节可以看到分开的好处——适合不同的 $P(X)$。

4.3 语义信道匹配 Shannon 信道——逻辑贝叶斯推理

语义信道匹配 Shannon 信道就是逻辑贝叶斯推理(LBI)。Shannon 信道匹配语义信道将在讨论机器学习实际例子时介绍。优化一个语义信道等价于优化一组真值函数或模型参数 $\theta_1, \theta_2, \ldots, \theta_n$. 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 $T(\theta_j|X)$. 于是有带参数的 LBI 是

$$T^*(\theta_j \mid X)=\arg\max_{T(\theta_j|X)} I(X;\theta_j)=\arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{4.8}$$

$I(X;\theta_j)$可以写成两个 KL 距离的差，

$$I(X;\theta_j)=\sum_i P(x_i \mid y_j)\log\frac{P(x_i|y_j)}{P(x_i)}-\sum_i P(x_i \mid y_j)\log\frac{P(x_i|y_j)}{P(x_i \mid \theta_j)} \tag{4.9}$$

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 和 $P(X|\theta_j)=P(X|y_j)$可以得到

$$T^*(\theta_j|X)=P(y_j|X)/\max[P(y_j|X)]= [P(X|y_j)/P(X)]/\max[P(X| y_j)/P(X)] \tag{4.10}$$

上式适合有大样本时的非参数估计，而式(4.8)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转移概率函数 $P(y_j|X)$做贝叶斯预测，结果相同. 所以和 BI 相比，LBI 更加兼容贝叶斯定理 2.

当我们不知道信源 $P(X)$而只知道 $P(y_j|X)$时，我们不妨假设 X 是等概率分布的，于是 $P(X|y_j)$和 $P(y_j|X)$成正比。然后我们得到：

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i \frac{P(y_j|x_i)}{\sum_k P(y_j|x_k)} \log \frac{T(\theta_j \mid x_i)}{\sum_k T(\theta_j|x_k)} \tag{4.11}$$

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，$T(\theta_j|X)\le T(\theta_k|X)$时，$y_j$ 蕴含 y_k.

在所有标签中， 可能有些标签是逻辑互补的。 假设 y_j’ 是 y_j 的否定，则我们可以用两个条件样本分布 $P(X|y_j)$和 $P(X|y_j$’)训练一个真值函数：

$$\begin{aligned} T^*(\theta_j \mid X) &= \arg\max_{T(\theta_j|X)} I(X;\theta_j) \\ &= \arg\max_{T(\theta_j|X)} \sum_i [P(x_i, y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} + P(x_i, y_j') \log \frac{1\text{-}T(\theta_j \mid x_i)}{1\text{-}T(\theta_j)}] \end{aligned} \tag{4.12}$$

流行的分类方法中，需要考虑二元相关性【42】，即对于每个标签，要考虑所有实例是属于互补的两类中的一类。这对样本 **D** 要求很高，使得多类别分类非常困难。但是上面公式也适合多类别多标签分类，不需要有完整的二元关联。因为上式把所有实例分为三类：正例、反例和不确定实例。上面只是接收者学习用的公式，而不是标签发送者用的公式。因为优化的真值函数 $T^*(\theta_j|X)$只和转移概率函数 $P(y_j|X)$和 $P(y_j'|X)$有关，而和 $P(X)$无关， 所以上式实际上把所有不确定实例排除在外了。只有发送者做划分的时候，才考虑所有实例，包括 $P(X)$。那时候用最大似然准则或最大语义信息准则划分就容易了。

上面假设 Shannon 信道是确定的， 但是在它不确定的情况下，比如, 对于实例不可见场合，我们只能根据观察数据 $Z\in C$ 预测 X 的概率分布 $P(X|Z)$时，我们需要划分函数 $Y=h(Z)$。在 $h(Z)$不确定的情况下，我们需要一个最优划分，使得不同标签的语义信息达最大。这时候需要用最大语义互信息公式求最优 Shannon 信道和语义信道：

$$T^*(\theta \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta \mid h) = \arg\max_{T(\theta_j|X)} \sum_j T(C_j) \sum_i P(x_i \mid C_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{4.13}$$

其中 $T(C_j)=P(Z\in C_j)$, $h(Z)$和 $P(Z|X)$一起确定了一个 Shannon 信道。这是一个半监督学习问题，要用到迭代算法——信道匹配(CM)算法，第 6 节详谈。

4.4 GPS 重定位和根据先验 $P(X)$翻译

如果 GPS 根据小车到三个卫星的距离定位存在误差，参看图 4， 我们可以根据 Shannon 信道纠正系统误差。假设转移概率函数或 Shannon 信道是

$$P(y_j \mid X) = K \exp[-|X - x_j - \Delta x|^2 /(2d^2)], \quad j=1, 2, \ldots, n \tag{4.14}$$

其中 x_j 是所指位置，即 y_j="X=x_j"中的 x_j；K 是系数，Δx 是系统偏差，d 是标准偏差， 那么根据逻辑贝叶斯推理得到的纠正的语义信道是（y_j 变为 y_k）：

$$T^*(\theta_k \mid X) = \exp[-|X - x_k|^2/(2d^2)], \quad k=1,\ 2,\ \ldots,\ n \tag{4.15}$$

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，都可以用客观的 Shannon 信道矫正语义信道。

假设 GPS 装置厂家提供的均方误差(RMS)较大，参看图 4，比如说是 200 米。在 GPS 装置和司机之间有个翻译，他要根据地形即 $P(X)$提供“小车在某点附近不超过 20 米处”或“小车在这条小路上”这样的精度更高的翻译或预测。那么他可以用 $P(X|\theta_j)$取代样本分布 $P(X|y_j)$，用下面公式优化翻译：

$$y_k^* = \arg\max_k I(X;\theta_k|y_j) = \arg\max_k \sum_i P(x_i \mid \theta_j)\log\frac{T(\theta_k \mid x_i)}{T(\theta_k)} \tag{4.16}$$

其中 $P(x_i|\theta_j)$是根据 GPS 得到的语义贝叶斯预测. $T(\theta_k|x_i)$是翻译 y_k 的真值函数。

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

4.5 逻辑贝叶斯推理和贝叶斯推理比较

贝叶斯推理(BI)和逻辑贝叶斯推理(LBI)使用的贝叶斯公式分别是：

$$\text{BI：}\ P(\theta \mid \mathbf{X}) = \frac{P(\theta)P(\mathbf{X} \mid \theta)}{P_\theta(\mathbf{X})}\ ,\ P_\theta(\mathbf{X})=\sum_j P(\theta_j)P(\mathbf{X} \mid \theta_j) \tag{4.17}$$

$$\text{LBI：}\ T(\theta_j \mid X) = \frac{T(\theta_j)P(X \mid \theta_j)}{P(X)}\ ,\ T(\theta_j)=1\Big/\max\left[\frac{P(X \mid \theta_j)}{P(X)}\right] \tag{4.18}$$

两者都强调利用先验知识。但是两者的区别是：

1） BI 使用的先验知识是概率分布 $P(\theta)$； $P_\theta(\mathbf{X})$是横向归一化系数，和模型相关。而 LBI 使用的先验知识是概率分布 $P(X)$；$T(\theta_j)$是纵向归一化系数，和模型相关。

2） BI 求 θ 的后验概率分布 $P(\theta|\mathbf{X})$，而 LBI 求使标签 y_j 或模型 θ_j 为真的真值函数 $T(\theta_j|X)$。BI 中的 $P(\theta|\mathbf{X})$是条件概率函数；θ 是变化的；$\mathbf{X}$ 是矢量空间中的一点，是不变的。而 LBI 中，$T(\theta_j|X)$类似于转移概率函数；θ_j 不变 X 变。

3） BI 需要很多似然函数 $P(X|\theta_j)$和很多条件样本 $\mathbf{X}_j$(对所有 j)，而 LBI 只要一个似然函数 $P(X|\theta_j)$，一个 $\mathbf{X}_j$ 或 $P(X|j)$.

笔者认为 X 的先验知识通常更加重要，是因为我们决策的依据的是预测的 X 的概率分布 $P(X|\theta_j)$，这也是最大似然方法的价值所在。笔者认为真值函数更加重要，是因为有了真值函数，我们不但能求出 $P(X|\theta_j)$, 还能在 $P(X)$变为 $P'(X)$时，求出相应的后验概率分布 $P'(X|\theta_j)$。这正是迁移学习需要的。

我们之所以求真值函数而不求带参数的转移概率函数 $P(y_j|X;\theta)=P(y_j)P(X|\theta_j)/P(X)$有下面几个原因：

1） 求解 $P(y_j|X;\theta)$比较困难，因为通常不知道 $P(y_j)$; 而求解 $T(\theta_j|X)$比较容易，并且 $T(\theta_j|X)$能和转移概率函数一样，和 $P(X)$一起能产生 X 的后验概率分布

$P(X|\theta_j)=P(X|y_j)$。

2） $T(\theta_j|X)$的最大值是 1，能反映概念外延，便于记忆，也便于用自然语言表达，比如 y_j 的高斯真值函数可以描述为“X 大约是 x_j”。

因为用真值函数可以取代转移概率函数，语义信息方法同样能提供一定置信区间内的置信水平 CF(Confidential Level)，在 0 和 1 之间变化。另外，语义信息方法还提供一个全称假设的确证度 CM(Confirmation Measure)，在-1 和 1 之间变化, 它反映归纳支持度（详见 5.4 节）。

使用最大语义信息准则优化模型参数可谓最大语义信息估计, 简记为 MSI。它和最大似然估计 MLE 即最大后验估计 MAP 比较如下(θ_j*中加下标 $_j$ 是为了区别混合模型参数)：

$$\text{MLE: } \theta_j^*=\arg\max_{\theta_j} P(\mathbf{X}_j\mid\theta_j)=\arg\max_{\theta_j}\sum_i P(x_i\mid j)\log P(x_i\mid\theta_j) \tag{4.19}$$

$$\text{MAP: } \begin{aligned}\theta_j^*&=\arg\max_{\theta_j} P(\theta_j\mid\mathbf{X}_j)=\arg\max_{\theta} P(\theta_j)P(\mathbf{X}_j\mid\theta_j)\\&=\arg\max_{\theta_j}[\sum_i P(x_i\mid j)\log P(x_i\mid\theta_j)+\log P(\theta_j)]\end{aligned} \tag{4.20}$$

$$\text{MSI: } \begin{aligned}\theta_j^*&=\arg\max_{\theta_j}\frac{P(\mathbf{X}_j\mid\theta_j)}{P(\mathbf{X})}=\arg\max_{\theta_j}\sum_i P(x_i\mid j)\log\frac{P(x_i|\theta_j)}{P(x_i)}\\&=\arg\max_{\theta_j}\sum_i P(x_i\mid j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\end{aligned} \tag{4.21}$$

MSI 和 MLE 是等价的，但是两者的区别是：

1)对最大似然度的理解不同。MSI 认为需要最大化的是标准(normalized)似然度. 因为 $P(X)$不依赖于参数，所以 MSI 和 MLE 等价。

2)MSI 用真值函数作为预测模型(少数情况下，比如求混合模型时，也用似然函数作为预测模型)。真值函数只反映信道特性，而似然函数包括信道和信源信息。MSI 用真值函数产生似然函数。在涉及语义、数值预测、定位…的场合，以及信源可变场合(需要迁移学习)，即需要正则化的误差准则的场合(详见下节)，真值函数作为预测模型更加适用。

MSI 和 MAP 的主要区别是：

1)预测模型不同。MSI 求真值函数中的参数而不是似然函数中的参数。

2)MAP 包含似然函数中参数估计。样本不大时，该估计和 MLE 差别显著，因而也和 MSI 不兼容。样本很大时，三种估计得到的似然函数 $P(X|\theta_j)$相同, 都等于样本分布 $P(X|j)$。

4.6 最大语义信息准则如何兼容正则化最小误差平方(RLS)准则

下面我们说明 MSI 准则是一个特殊的也许是更合理的 RLS 准则【42】。

把高斯分布真值函数带进广义 KL 公式(4.3)得到

$$I(X;\theta_j)=-\log T(\theta)-\sum_i P(x_i\mid y_j)(x_i-x_j)^2/(2d_j^2) \tag{4.22}$$

右边第一项是 Bar-Hillel-Carnap 信息，第二项是平均相对误差平方。把高斯分布真值函数带进语义互信息公式(4.4)得到

$$I(X;\theta) = H(\theta) - \sum_j \sum_i P(x_i, y_j)(x_i - x_j)^2 / (2d_j^2) \tag{4.23}$$

右边第一项 $H(\theta)$是 Y 的交叉熵，第二项是平均相对误差平方。可见，最大语义信息准则可理解位一个特殊的 RLS 准则。$H(\theta|X)$反映相对误差，类似于 RLS 准则中的误差项；$H(\theta)$就类似于正则化项。$I(X; \theta)$就是负的损失函数. 最小误差准则好像是“无过便是德”的准则，而最大语义信息准则好像是“功大于过便是德”的准则。因为最大语义信息准则等价于最大似然准则，所以可以说它是兼容最大似然准则的 RLS 准则。

流行的 RLS 准则中的正则化项有多种，惩罚项惩罚所有参数【42】。解释是避免过度拟合，提高泛化能力。但是，从上面分析我们看到，正则化项也是为了解决类别不平衡问题，并使优化准则兼容最大似然准则。$H(\theta)$作为正则化项，表示最大潜在信息(Barhillel-Carnap 信息的平均【13】)，具有容错编码意义【14】，含义明确。每个逻辑概率 θ_j 越小则潜在信息量 $H(\theta)$越大。但是这就要求每个 d_j 小。然而，每个 d_j 小了，相对误差就大了。所以相对误差项是惩罚函数。反过来也可以这样理解，要想相对误差小，就要求每个 d_j 大。但是每个 d_j 大了，潜在信息量 $H(\theta)$就小了。所以要用 $H(\theta)$限制每个 d_j. 但是 $H(\theta)$并不限制真值函数的期望值 x_j。笔者相信，把期望值排除在正则化项之外，用标准差 d_j 或影响相似函数分布范围的参数构造惩罚项，应能改善机器学习。

另外，$T(\theta_j)$还和 $P(X)$有关。如果 $P(X)$在某个区间的分布密度很小，尽管相应的 d_j 较大，$T(\theta_j)$还是很小。对于这样的预测 y_j, 信息准则允许有较大的相对误差。比如我们猜某人寿命可能 100 岁(先验概率较小)，实际 90 岁，那也还是不错的猜测； 如果猜测寿命可能 70 岁(先验概率较大)，实际 60 岁，这预测就差多了。语义信息准则用于分类时，$T(\theta_j)$就表示第 j 个类别中实例的相对数。所以语义信息准则也考虑了类别不平衡问题【45】。和误差准则相比，它注重减少小概率事件的漏报——比如减少艾滋病、地震、价格暴涨暴跌，百岁老人…的漏报；同时允许小概率事件有较多的误报。

有人认为从 BI 可以推导出 RLS 准则， 认为 $\log P(\theta|\mathbf{X})=\log P(\theta)+\log[P(\mathbf{X}|\theta)/P(\mathbf{X})]$, $\log P(\theta)$就是惩罚函数【44】。笔者认为不对。因为在贝叶斯推理中，似然函数中的很多参数，并不出现在先验分布 $P(\theta)$中, $P(\theta)$中参数也不会在优化模型的时候改变，所以 $P(\theta)$不能作为惩罚项。因为 θ 的后验分布和先验分布成正比，$P(\theta)$的作用是局部增强，而不能惩罚似然函数中需要惩罚的参数。

RLS 流行使得好像最大似然准则过时了。其实不然！当我们优化 X 的似然函数 $P(X|\theta_j)$时候，应该不需要加正则化项。只有我们在用 Y 的后验 $P(Y|X, \theta)$作为似然函数的时候，我们才需要正则化项。从语义信息论的角度看，最大似然准则只合适优化 X 的后验分布 $P(X|\theta_j)$，而不合适优化 Y 的后验分布 $P(Y|X, \theta)$。而最大标准似然准则或语义信息准则才适合两者。所以，如果我们用预测误差函数表示 $P(Y|X, \theta)$，我们只能用样本检验标准似然函数 $P(Y|X, \theta)/P(Y)$ 或 $T(\theta_j|X)/T(\theta_j)$。而平均 $\log P(Y)$或 $\log T(\theta_j)$就是正则化项。因为 $P(Y)$不容易求出，所以用 $T(\theta_j|X)/T(\theta_j)$是最好选择。

总之，用逻辑概率 $T(\theta)$或交叉熵 $H(\theta)$既可以解决正则化问题，也可以解决类别不平衡问题【45】。从语义信息论的角度看，这两个问题涉及同一个问题——是否兼容最大标准似然准则问题。

5. Shannon 信道匹配语义信道——发信人选择分类

5.1 发信人多标签分类

当前的多标签分类大多使用 Y 的贝叶斯后验，在学习阶段就考虑划分。一个流行的方法是把多类别多标签分类转换为多个二分类，要求我们为每个实例对每个标签(即后面的原子标签)提供“*Y*es”或“*N*o”【46,43】。本文方法与之类似，但是有所不同。不同之处：

1)标签学习只优化真值函数——通常是模糊的，不划分实例空间，标签发送才划分实例空间；

2)每个标签可以单独学习， 或者和其否定标签同时学习(参看(4.12))， 两种都可以；

3) “*Y*es”或“*N*o”能提供多少是多少，缺少也没关系；

4)多标签在这里就是多原子标签或由它们构成的复合标签。允许一个标签(复合标签)是多个原子标签的布尔函数。设有 q 个原子标签，用流行的二元关联方法(Binary Relevance)【43】，复合标签最多有 2^q 种。是用本文方法，可能的复合标签是 2^{2^q} 种。不过实际用到的标签可以随样本或需要而定。

下面考虑多标签选择分类。假设 Y 是一个可选择标签(原子标签或复合标签), 其真值函数已经给定。关于如何把多标签学习简化为多原子标签学习，下一节讨论。

为了得到联合概率分布 $P(X,Y)$，我们可以仿照已经采用的分拆方法【47】，首先要把单实例多标签样例，比如$(x_1; y_1,y_2)$，和多实例单标签样例，比如$(x_1,x_2; y_1)$, 分解成单实例单标签样例，然后得到 $P(X,Y)$。

在实例可见时，使用最大语义信息准则，选择分类的分类函数是

$$y_j^*=f(X)=\arg\max_{y_j}\log\frac{T(\theta_j\mid X)}{T(\theta_j)} \tag{5.1}$$

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：

$$y_j^*=f(X)=\arg\min_{y_j \text{ with } T(\theta_j|X)=1} T(\theta_j) \tag{5.2}$$

也就是最丰富内涵准则。可以说，标签学习得到标签的(模糊)外延，而标签选择依据标签的信息——可以理解为内涵。

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【45】，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或划分观察数据空间， 即

$$y_j^*=h(Z)=\arg\max_{y_j}\sum_i P(x_i\mid Z\in C_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{5.3}$$

在预测不准时，模糊集合外延较大可以减小误差带来的信息损失， 所以用上式选出来的 y_j^*, 其逻辑概率 $T(\theta_j)$未必最小. 这也是为什么天气预报经常报“小到中雨”。

对于不可见实例分类，划分函数 $h(Z)$(确定每个 C_j)会改变 Shannon 信道和与之相匹配的语义信道，所以又要求重新划分… 因此。我们要用到迭代方法。后面谈及。

因为每个标签学习是单独的，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——因为不用考虑逻辑分类函数相加是否等于 1. 其否定标签可以一同学习，如(4.12)，也可以单独学习， 也可以没有。比如：X 表示年龄， 标签可以有：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⩾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， 无论有两种标签，还是 n 种标签，这些隶属函数的参数形式都可以不变. 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.

5.2 通过原子标签学习简化多标签学习和分类

设 Y 是一个可选择标签(原子标签或复合标签)， $Y\in V=\{y_1, y_2, \ldots, y_n\}$. 再设原子标签是 $a\in W=\{a_1, a_2, \ldots, a_q\}$；一个复合标签是若干原子标签的布尔函数。那么 q 个原子标签可以构成 2^q 个用 and 链接的互斥复合标签，它们是集合 2^W 中的元素。q 个原子标签的布尔函数等于 2^q 个互斥复合标签中部分标签相加， 最多可能有 2^2^q 种。太多的可能标签会带来太大的学习工作量。好在信道匹配算法允许我们选择其中很少的标签来学习。但是，我们仍然可以用原子标签学习取代复合标签学习，从而减小学习工作量。

首先，我们选择所有可选择标签需要的原子标签，筛选相关的样例构成学习样本——包括从多标签样例中分拆出单标签样例。没有筛选干净或不筛选也没有关系，因为每个原子标签的学习是独立的。发送者分类时， 先用这些原子标签的隶属函数产生复合标签的隶属函数——可能需要模糊逻辑运算，再用 5.1 节中的方法得到分类函数。比如，如果两个标签不相关，那么 $T(\theta_1\cap\theta_2|X)=T(\theta_1|X)T(\theta_2|X)$. 如果相关性很大，则可用模糊逻辑运算，比如：$T(\theta_1\cap\theta_2|X)=\min[T(\theta_1|X),T(\theta_2|X)]$.

为了使标签逻辑运算构成布尔代数，隶属函数的逻辑运算也要与之兼容。我们可以使用一种与 Zadeh 模糊逻辑【32】有点不同的模糊逻辑【20】，其中有

$$T(\theta_1\cap\theta_2{}^c|X)=\max(0，T(\theta_1|X)-T(\theta_2|X)) \tag{5.4}$$

这是笔者建立色觉机制数学模型——译码模型——时使用的模糊逻辑【48】，它用 UX[0,1]空间上的布尔代数定义 U 上的模糊逻辑运算(参看图 7)。

假设年龄集合上有模糊集合 θ_1={成年人}，θ_2={年轻人}， 那么 4 个合取标签的真值函数就如图 7 中两条曲线划出的四个区域所示，标签 $a_1\wedge a_2$’的真值函数 $T(\theta_1\cap\theta_2{}^c|X)$ 就如阴影部分所示。

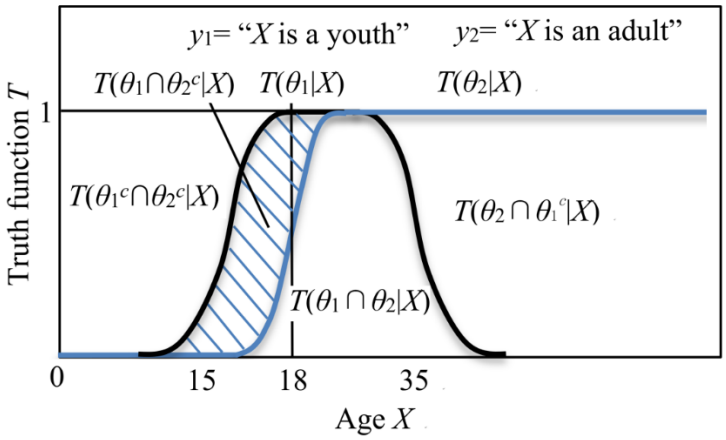


图 7 两个原子标签产生四个合取标签，隶属函数运算服从布尔代数，

Fig.7 Logical operations of truth functions make the fuzzy set algebra be a Boolean algebra

但是对于相关性不明了的原子标签，要得到合理的复合标签的真值函数，还要采用更细致方法，可能要结合已有方法【46,43】；要么直接用样本训练复合标签。

5.3 可见实例二分类及其随 *P*(X)的变化

我们以二分类为例说明标签发送者分类如何随先验分布 $P(X)$变化，以及如何解决类别不平衡和正则化问题。

设有 y_0 和 y_1 两种标签。实例空间是一维的——为了理论上讨论方便，后面都假设实例空间是一维的。两个真值函数 $T(\theta_1|X)$和 $T(\theta_0|X)$可以是分别用两组样本训练的，见(4.8)；也可以是相关的，见(4.11)，比如是一对 Logistic 函数表示的。流行的方法中通常用带参数的转移概率函数 $P(y_j|X;\theta)$表示分类函数，归一化要求是 $P(y_1|X;\theta)+P(y_1|X;\theta)=1$。但是用信道匹配算法，我们并不要求 $T(\theta_1|X)+T(\theta_0|X)=1$. 如果两个真值函数是一起优化的，则 $T(\theta_1|X)+T(\theta_0|X)=1$ 成立。下面先看采用信道匹配算法如何使用 Logistic 函数做二分类。设

$$T(\theta_1 \mid X)=\frac{1}{1+e^{-a(X-b)}},\quad T(\theta_0 \mid X)=\frac{e^{-a(X-b)}}{1+e^{-a(X-b)}} \tag{5.5}$$

语义互信息是

$$\begin{aligned}
&I(X;\theta)=\sum_i[P(x_i,y_1)\log\frac{T(\theta_1\mid x_i)}{T(\theta_1)}+P(x_i,y_0)\log\frac{1\text{-}T(\theta_1\mid x_i)}{1\text{-}T(\theta_1)}]\\
&=H(\theta)-H(\theta\mid X)\\
&H(\theta\mid X)=\sum_i[P(x_i,y_1)\log\frac{1}{1+e^{-a(x_i-b)}}+P(x_i,y_0)\log\frac{e^{-a(x_i-b)}}{1+e^{-a(x_i-b)}}]\\
&H(\theta)=-P(y_1)\log T(\theta_1)-P(y_0)\log[1-T(\theta_1)]\\
&=-P(y_1)\log\sum_i\frac{P(x_i)}{1+e^{-a(x_i-b)}}-P(y_0)\log\sum_i\frac{P(x_i)e^{-a(x_i-b)}}{1+e^{-a(x_i-b)}}
\end{aligned} \tag{5.6}$$

用上面公式就可以求出使 $H(X;\theta)$达最大的参数 a 和 b. 上式中误差项 $H(\theta|X)$和流行的方法相同，但是正则化项 $H(\theta)$复杂些。其好处是：

1） 可以解决类别不平衡问题；

2） 和最大似然准则兼容；

3） 信源 $P(X)$变化后， 模型仍然适用。

因为这种方法不要求两个真值函数相加恒等于 1，我们或许可以找到两个更加便于计算的真值函数模型。对此， 需要进一步研究。

我们以标签“老年人”为例说明类别划分随实例的先验分布 $P(X)$变化。

设 X 表示年龄，标签 y_1=“老年人”。假设

$$T(\theta_1 \mid X)=\frac{1}{1+e^{-0.2(X-75)}},\quad P(X)=1\text{-}\frac{1}{1+e^{-0.15(X-c)}}$$

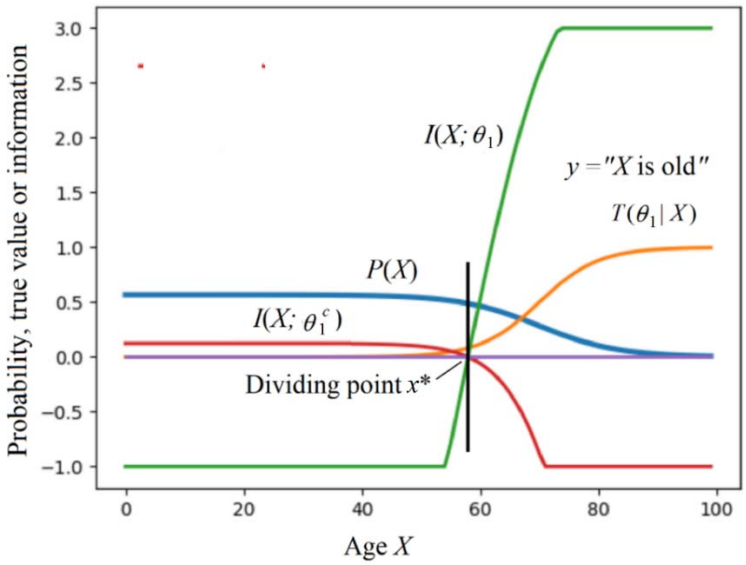


图 8 老年人分类划分点 x^*. 它在老年人人口比例增加时右移。

Fig. 8 The optimized dividing point x^* for label "Old person". The x^* moves rightward when old population increases.

图 8 显示了有关函数和优化的划分点 x^*——它随 X 的先验分布 $P(X)$中参数 c 的变化而变化，参看表 1。

表 1 "老年人" 分类器 $f(X)$ 随人口密度分布 $P(X)$中的参数 c 的变化

Table 1 The classifier $f(X)$ for y_j="Old person" changes with $P(X)$

c $P(X)$中参数	Population density decreasing area 人口密度下降区域	Dividing point x^* 划分点($y_1=f(X\geqslant x^*)$
50	40-60	49
60	50-70	55
70	60-80	58

注意：这时候逻辑分类函数并没有变化。划分点右移会导致新的本中样例的概率分布 $P(x_i, y_1)$以及转移概率函数 $P(y_1|X)$变化，从而导致标签接受者理解的真值函数变化。这些变化也会反过来影响划分函数 $f(X)$。"老年人" 的语义和使用规则就是这样进化的。没有人规定老年人的严格分界年龄在哪里，语言交流过程中会自动形成一个模糊分界，它随人口的年龄分布变化而变化。老年人多的群体，可能 70 岁才算老年人，50 岁算年轻人。以前人的寿命短，50 岁就算老年人了。"大雨" 类似，雨水多的地区可能日降雨量 50mm 才算大雨；而雨水少的地区可能日降雨量 20mm 就算大雨了。

5.4 不可见实例二分类的置信水平和确证度——兼评贝叶斯推理的置信分析

我们以医学检验为例说明检验和不可见实例二分类的 Shannon 信道和语义信道，置信水平和确证度。

医学检验可以看成一个 2X2 有噪声 Shannon 信道(见图 9)。相应实例集合是 $U=\{x_0, x_1\}$，其中 x_0 表示未感染者，x_1 表示感染者；相应标签集合是 $V=\{y_0, y_1\}$，y_0 表示阴性，y_1 表示

阳性. 图中 $Z\in C$ 表示观察数据， Z'是划分函数 $Y=h(Z)$的划分点。对于不可见实例分类，x_0 和 x_1 表示真标签，y_0 和 y_1 是根据 Z 选择的标签。

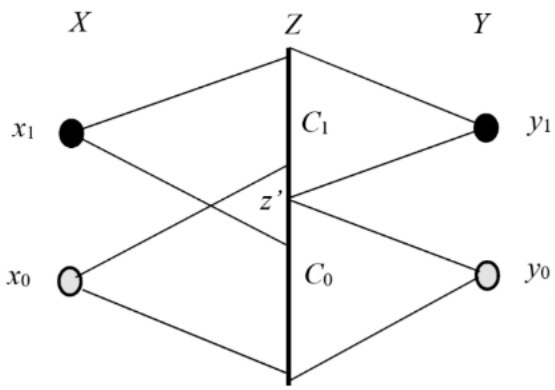


图 9 医学检验图解。检验可以被抽象为 2X2 有噪声 Shannon 信道。

Figure 9. Illustrating the medical test. The test can be abstracted as a 2×2 Shannon nos*y* channel. The Shannon mutual information changes with dividing point *z*'.

医学检验中，条件概率 $P(y_1|x_1)$叫做敏感性(sensitivity)，$P(y_0|x_0)$叫做特异性(specificity)【48】。敏感性和特异性构成 Shannon 信道, 如表 2 所示。

表 2 医学检验的敏感性和特异性构成 Shannon 信道 $P(Y|X)$

Table 2 The sensitivity and Specificity of Medical Tests Form a Shannon's Channel $P(Y|X)$

Y	**感染者 Infected x_1**	**未感染者 Uninfected x_0**
Positive y_1	$P(y_1\|x_1)$=sensitivity	$P(y_1\|x_0)$=1-specificity
negative y_0	$P(y_0\|x_1)$=1-sensitivity	$P(y_0\|x_0)$=specificity

如果检验绝对可信，应有真值 $T(y_1|x_1)=T(y_0|x_0)=1, T(y_1|e_0)=T(y_0|x_1)=0$. 如果可能有错，优化的真值函数应在 0 和 1 之间。设一个假设 y_j 分为可信部分和不可信部分，可信比例是 b_j，则其真值函数是

$$T(\theta_j|X)= b_j' +b_jT(y_j|X) \tag{5.7}$$

$b_j'=1-|b_j'|$是 y_j 中含有永真句(无意义)的比例，也是反例的真值，可谓不信度。于是相应的语义信道如表 3 所示。

表 3 两种不信度构成医学检验的语义信道

Table 3 Two no-confidence Levels of a Medical Test Form a Semantic Channel $T(\theta|X)$

Y	**感染着 Infected x_1**	**未感染者 Uninfected x_0**
Positive y_1	$T(\theta_1\|x_1)=1$	$T(\theta_1\|x_0)=b_1'$
*N*egative y_0	$T(\theta_0\|x_1)=b_0'$	$T(\theta_0\|x_0)=1$

根据(4.10)，两个优化的不信度为

$$b_1'^*= P(y_1|x_0)/P(y_1|x_1); \quad b_0'^*=P(y_0|x_1)/P(y_0|x_0) \tag{5.8}$$

我们把优化的不信度叫做否证度，相应的信任度叫做确证度或可信度。医学界用似然比(Likelihood Ratio)表示一个检验有多好 【50,51】. 上式与之兼容， 因为

$$\mathrm{LR}^+=P(y_1|x_1)/P(y_1|x_0)=1/b_1'^*; \quad \mathrm{LR}^-=P(y_0|x_0)/P(y_0|x_1)=1/b_0'^* \tag{5.9}$$

说明似然比大和否证度小或确证度大是一一对应的。

上面方法得到的确证度和已有的各种确证度不同【52】,反例的比例对确证度影响更大。比如阳性的可信度主要不取决于敏感性的大小， 而是取决于特异性的大小。比如：一个医生，把有病的人的一半诊断为有病，从来不把无病的人诊断有病。那么，这个医生诊断“有病”就完全可信；相反，诊断“无病”反而不太可信。这样的确证度和 Popper 的证伪思想(重视反例的存在)兼容。

医学检验中用检验阳性(预测有病)来自正例的相对比例叫做阳性的置信水平(Confidence Leval)【29】, 即

$$CL_1=P(y_1|x_1)\ /[P(y_1|x_0)+P(y_1|x_1)] \tag{5.10}$$

容易证明，一个标签的置信水平 CL 和确证度 b^*之间的关系是【29，17】:

$$b^*=\begin{cases}1-CL'/CL, \text{ if } CL>0.5\\ CL'/CL-1, \text{ if } CL\le 0.5\end{cases} \tag{5.12}$$

其中 CL'=1-CL。上式中 CL 在 0 和 1 之间变化， 而 b^*在-1 和 1 之间变化。当 CL=0.5 时，b^*=0；可解释为：一半可信时，归纳支持度为 0。CL=0 时，b^*=-1；可解释为，全部不可信时，证据支持相反假设。

考虑到检验的两种结果，有一种更加重要(价值或损失的原因)，我们需要限制其 CL 下限(可能要牺牲另一种结果的置信水平或导致放弃检验)，于是就有显著性水平 α 要求, 即要求 CL'=1-CL< α. 比如在医学检验中，为了防止漏报，我们可能要求阴性的 CL' 或 b'^*不大于 0.1。在垃圾邮件分类时，误报重要，我们可能要求阳性(预测“是”)的 CL' 或 b'^*不大于 0.05.

另外用 b'^*做贝叶斯预测比用转移概率函数还要方便。

2.3 节中例 1 解： $P(x_1)$=0.002, $P(x0)$=1-0.002=0.998; $P(x_1|y_1)=P(x0|y_1)$=0.5. 根据(4.10)，可求出 y_1 的真值函数：$T(\theta_1|x_1)$=1 和

$T(\theta_1|x0)=b_1'=[P(x_0|y_j)/P(x_0)]/\max[P(X|y_j)/P(X)]=0.5/(1-0.002)/(0.5/0.002)=0.002$

两者构成一个预测模型，对于任意 $P'(x)$, 可以用贝叶斯定理 3 求出 x 的后验。$P(x_1)$变为 $P'(x_1)$=0.1 时，用第三种贝叶斯定理得到

$$P'(x_1|\theta_1)=P'(x_1)/[P'(x_1)+b_1'P'(x0)]=0.1/(0.1+0.002*0.9)=0.98$$

解毕。

上面置信水平和确证度只适合类别明确的估计；否则，要添加置信区间。比如：说“温度表读数的置信水平是 0.9”不行， 还要加上置信区间。比如说“温度表读数的误差在正负 0.5 度之间的置信水平是 0.9”。GPS 的置信水平类似。置信水平 0.9 转换成确证度是 1-1/9=0.89.

综上所述，优化的语义信道方法和医学检验方法完全兼容，而且能提供更好确证度，更方便做概率预测。不可见实例分类(需要划分观察数据空间 C，而不是实例空间 U)在本质上和估计(包括检验)类似。比如垃圾邮件分类就和医学检验在本质上相同。西瓜好差分类有些区别，主要是因为西瓜从好到差中间没有明确分界，客观的真类别不是两种。但是为了方便起见，我们也不妨按照医学检验方法处理。为了更加合理，我们可以也不妨添加中间类别。

置信水平和确证度在本质上是对 Shannon 信道噪声的评估，而和信源无关。分类也要加进这样的考虑才能使分类适合信源可变情况下的迁移学习。

上述置信水平分析方法是频率主义方法。BI 提供一种类似但是与之不同的方法，那就是用贝叶斯后验 $P(\theta|\mathbf{X})$, 提供参数 θ 的后验分布区间和相应的置信水平【27】。前面讲过预测 x 的后验分布 $P(x|\theta_j)$落入某个范围的概率，和先验分布 $P(x)$有关，那是范围预测，类似于于 VaR 分析，并不是用于 Y 的置信分析。置信水平和确证度类似，反映正例和反例的比例，间接或直接表明归纳支持程度。现在 BI 预测 θ 的范围，并称范围内的 $P(\theta)$是置信水平， 这

已经是偷换“置信水平”概念了。反例在哪里？没有反例比例，如何得到置信水平？

有一种误解，认为频率主义的置信水平反映的是 X 的后验分布特性，而贝叶斯推理的置信水平反映 θ 或 Y 的后验分布特性。其实，频率主义的置信水平反映的是转移概率或 Shannon 信道的特性，而不是反映 X 的后验 $P(X|y_j)$的特性。BI 提供置信水平分析没有用到 Shannon 信道，也没有考虑反例。这种分析在分析频率发生器的频率的时候才有概率预测意义，类似于 Var 分析(这时 θ 可以看做是和 x 一样的客观事实)，但是存在概念错误。这种分析用于分类(用 $P(Y|\mathbf{X};\theta)$)则会更加令人费解；在我看来也是不必要的。频率主义的置信分析已经足够了。

6 *R(G)*函数和两种信道相互匹配的迭代算法

6.1 把 *R(D)*函数改进为 *R(G)*函数

“信道匹配”英文是“Channels’ matching”，简写为“CM”。这种算法用在 Shannon 信道确定时是非迭代算法，不确定时是迭代算法——可用于最大似然估计(包括不可见实例分类)和混合模型。为了说明这种算法有效性和迭代收敛可靠性，本节介绍语义互信息 G 和 Shannon 互信息 R 之间的相互匹配函数 $R(G)$。这一函数是经典的 $R(D)$函数的改进版本(D 是平均失真上限 D，G 是语义互信息下限)，它最早被用来讨论主观信息和客观信息的匹配关系，包括如何根据视觉分辨率量化或压缩像素等级【12-14】。现在笔者发现它可以用来改进机器学习。用 G 代替 D 可以理解为用兼容 ML 准则的 RLS 准则代替失真准则。

Shannon 在开创性文章中提出保真度信息率理论【8】，并提出信息率失真函数 $R(D)$, Shannon 于 1959 年进一步讨论了 $R(D)$函数【53】, 并提供了一个二元信道的 $R(D)$函数的表达式。因为最大保证度是用最小失真表示的，所以从 Berger 开始称之为信息率失真理论【54】。周炯槃先生的《信息理论基础》【55】有最详细的讨论，特别是关于 $R(D)$函数参数形式的迭代求解，在英文文献中都很难找到。

给定信源 $P(X)$和平均失真下限 D 时，带有参数 s 的信息率失真函数是

$$\begin{aligned} D(s) &= \sum_i \sum_j d_{ij} P(x_i) P(y_j) \exp(s d_{ij}) / \lambda_i \\ R(s) &= sD(s) - \sum_i P(x_i) \ln \lambda_i \end{aligned} \tag{5.1}$$

其中 d_{ij} 是用 y_j 表示 x_i 的失真量。$R(G)$ 函数用语义信息量 $I_{ij}=I(x_i;\theta_j)$表示保真度，替代失真量 d_{ij}； 用语义互信息的下限 G 替代平均失真上限 D。在这些限制下求 Shannon 互信息的极小值 R, 于是得到 $R(G)$函数【13】:

$$\begin{aligned} G(s) &= \sum_i \sum_j I_{ij} P(x_i) P(y_j) 2^{sI_{ij}}) / \lambda_i = \sum_i \sum_j I_{ij} P(x_i) P(y_j) m_{ij}^s / \lambda_i \\ R(s) &= sG(s) - \sum_i P(x_i) \log \lambda_i \\ m_{ij} &= T(\theta_j \mid x_i) / T(\theta_j),\ \lambda_i = \sum_j P(y_j) m_{ij}^s \end{aligned} \tag{5.2}$$

其中 m_{ij} 是标准似然度，$P(y_j)\ m_{ij}^s/\lambda_i$ 就是 $P(y_j|x_i)$, 反映匹配的 Shannon 信道 $P(Y|X)$. 任何一个 $R(G)$函数的形状都是碗状的(可能不对称)。一个对称的二元信源 $R(G)$函数【13，16】如图 10

所示。

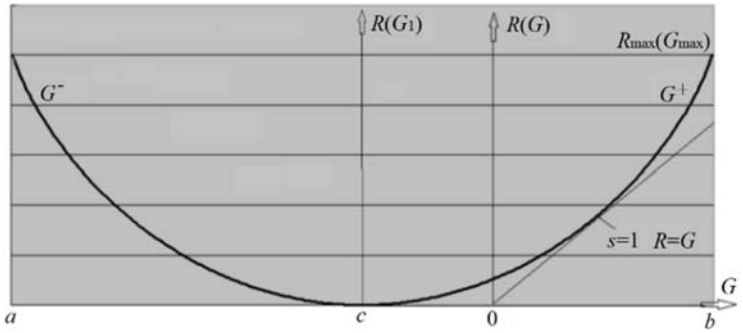


Figure 10. The $R(G)$ function of a binary source. As s=1, R=G, which implies that the semantic channel matches the Shannon channel.

图 10 二元信源的 $R(G)$函数。给定 R 改变语义信道使得 s=1，即 G=R，就是语义信道匹配 Shannon 信道。

Shannon 信道匹配语义信道有两种方式：1)目的是 R 和 G 同时达到最大，如右上角 $R_{max}(G_{max})$ 所示，求最大似然估计需要这种匹配; 2)目的是最小化 R-G 使 R=G, 求混合模型需要这种匹配。

对于 $R(D)$函数，给定 R 只有一个 D；而对于 $R(G)$函数，给定 R 存在一个极大 G 即 G^+ 和一个极小 G 即 G^-。负的 G^-意味着谎言或错误预测会带来负的信息，或者说为了给敌人带来一定的信息损失，$R(G^-)>0$ 意味着一定量的 R 是必要的。R=0 时 G 是负的， 意味着听信他人随机乱说会减少我们已有的信息。

当 s=1 时, λ_i=1, R=G。这表示两种信道匹配正好，语义互信息等于 Shannon 互信息。

参数 s 的绝对值$|s|$反映预测精度或噪声，在 $R(D)$函数中，$\mathrm{d}R/\mathrm{d}D=s$ ($s\le 0$). 容易证明也有 $\mathrm{d}R/\mathrm{d}G=s$, s 可能小于 0 也可能大于 0. s 由正变负时，$R(-s_1)=R(s_1)$ 且 G 从 G^+ 变为 G^- (见图 10). 据此可以证明最大似然检验等价于最大似然比检验。

已有的检验的似然比表达都不直观【50】。给定 C 的划分$\{C_0,C_1\}$时，检验的似然比的具体表达式是

$$r_L=\left[\prod_{i=0}^{1}\left(\frac{P(x_i\mid\theta_1)}{P(x_i\mid\theta_0)}\right)^{P(x_i|C_1)}\right]^{NP(C_1)}\left[\prod_{i=0}^{1}\left(\frac{P(x_i\mid\theta_0)}{P(x_i\mid\theta_1)}\right)^{P(x_i|C_0)}\right]^{NP(C_0)} \tag{5.3}$$

根据式(4.7), $\max(\log r_L)=\max(-NH(X|\theta))-\min(-NH(X|\theta)=N(G^+-G^-)$ (参看图 10). 在 R 和 G^+ 确定后，s 就确定了, 所以 $G^-(s)$也确定了。因此，最大似然度和最大似然比是同时发生的。通过两者分别得到的最优 Shannon 信道是相同的。所以两个准则是等价的。已有文献也肯定两者等价， 但是没看到这一等价的证明。

6.2 CM 算法用于检验、估计和不可见实例分类

前面优化语义信道用的是 y_j 的平均语义信息公式——求 $I(X;\theta_j)$， 下面要用到语义互信息公式——求 $I(X;\theta)$。

对于检验、估计和不可见实例分类，设观察数据集合 C 的一个划分是 $S=\{C_1, C_2, \ldots\}$。当 $Z\in C_j$ 时，我们就选择 y_j, 即 $y_j=h(Z|Z\in C_j)$。由此就得到一个 Shannon 信道。当 C 的划分可变时，在标签接受者优化语义信道得到与之匹配的语义信道后，标签发送者需要重新划分

Shannon 信道，使之匹配语义信道。实践表明，如此循环，就可以得到 Shannon 互信息达最大的划分。这也是最大似然检验和估计要达到的目的。下面介绍这种迭代算法及其收敛证明。

给定 S 或 $Y=h(Z)$，平均语义互信息是：

$$I(X;\theta|S)=\sum_j P(C_j)\sum_i P(x_i\mid C_j)I(x_i;\theta_j)=\sum_j\sum_i P(C_j)P(x_i\mid C_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \quad (5.4)$$

设样本中的一个样例是$\{x_i, z_k\}$，如果样本足够大，我们可以从样本得到 $P(X,Z)$. 如果样本不够大，我们可以用参数构造 $P(Z|X)$(使用最大似然准则或语义信息准则, 不赘)。然后我们可以计算构成 Shannon 信道的转移概率函数：

$$P(y_j\mid x_i)=\sum_{z_k\in C_j} P(z_k\mid x_i),\ i=1, 2, \ldots, m \quad (5.5)$$

有了 $P(X)$和 $P(Y|X)$, 可以得到 $P(X|Z)$和下面广义 KL 信息分布函数：

$$I(X;\theta_j|Z)=\sum_i P(x_i\mid Z)I_{ij},\ \text{其中}\ I_{ij}=\log[T(\theta_j\mid x_i)/T(\theta_j)] \quad (5.6)$$

最大互信息划分迭代算法：

Matching I：语义信道匹配 Shannon 信道。用(4.10)得到优化的非参数语义信道。

Matching II：Shannon 信道匹配语义信道。用下面公式：

$$P(y_j|x_i)=\lim_{s\to\infty} P(y_j)m_{ij}^s/\lambda_i,\ i=1, 2, \ldots \quad (5.7)$$

其中 m_{ij} 是标准似然度 $T(\theta_j|x_i)/T(\theta_j)$. 参看(5.2)。若划分不能改进， 迭代终止， 否则转到 Matching I。

分析上式。当 $s\to\infty$, 转移概率函数 $P(y_j|X)$变成集合 C_j 的特征函数，Shannon 信道就是无噪声的，使 R 和 G 达最大(位于一个 $R(G)$函数曲线的右上角)。这种方法等价于使用划分函数

$$Y=h(Z)=\arg\max_{y_j} I(X;\theta_j|Z) \quad (5.8)$$

重复上面两种匹配就可以找到使 Shannon 互信息和语义互信息达最大的 C 的划分。流行的最大似然估计方法是用梯度法或牛顿法。因为上面迭代过程中没有搜索只有赋值，运算也更简单，速度也更快。

下面用 $R(G)$函数证明 R 收敛到 $R_{max}=G_{max}$(参看图 11)。

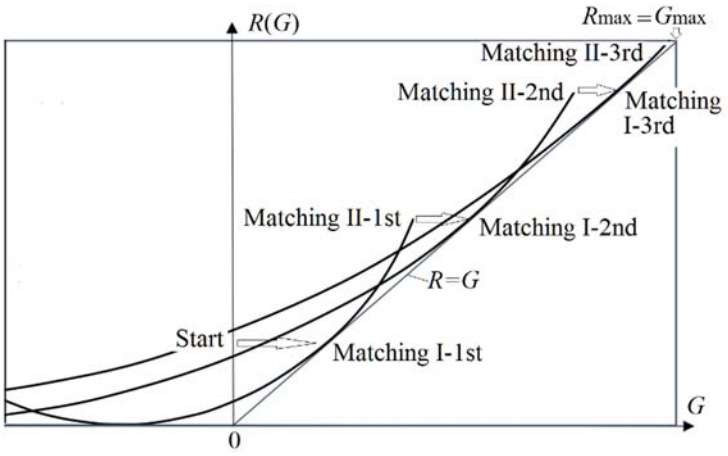


图 11. 图解检验和估计的迭代收敛. 每个语义信道决定一条 $R(G)$ 函数曲线. 匹配 I 使得 $G=R$，并产生一个新的语义信道；匹配 II 增大 R 到 $R(G)$ 函数曲线的顶部. 重复匹配 I 和匹配 II 可以使得 R 和 G 达到最大值 R_{max} 和 G_{max}.

Figure 11. Illustrating the iterative convergence for classifing unseen instances. 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}.

我们可以把语义信道看成是 Shannon 信道往上爬的阶梯。爬到顶上再产生一个更高的阶梯。

最大互信息划分收敛证明：优化步骤是(参看图 11)：初始化 S 和 Shannon 信道—>Matching I-1st——>Matching II-1st——>Matching I-2nd ——>Matching II-2nd … 直至 $R=R_{max}$ 且 $G_{max}=R_{max}$。因为上面每一步都最大化 R 和 G，所以迭代收敛到 $G_{max}=R_{max}$。**证毕**。

是否收敛的划分不是全局最优？这还需要进一步研究。笔者测试过几个例子，说明迭代快速且收敛可靠。其中一个例子如下：

假定 $P(Z|x_i)$高斯分布(K_i 是归一化系数)：

$$P(Z|x_i)=K_i\exp[-(Z-c_i)^2/(2d_i^2)], \quad i=1,2,\ldots \tag{5.9}$$

实验条件：给定 $P(x_0)=0.5, P(x_1)=0.35, P(x_2)=0.15; c_0=20, c_1=50, c_2=80; d_0=15, d_1=10, d_2=10$. 用两组不同起点: (1) z_1'=50 和 z_2'=60(较好的一对)；(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 12 所示. 可见在非常极端的情况下，迭代收敛也是可靠的[2].

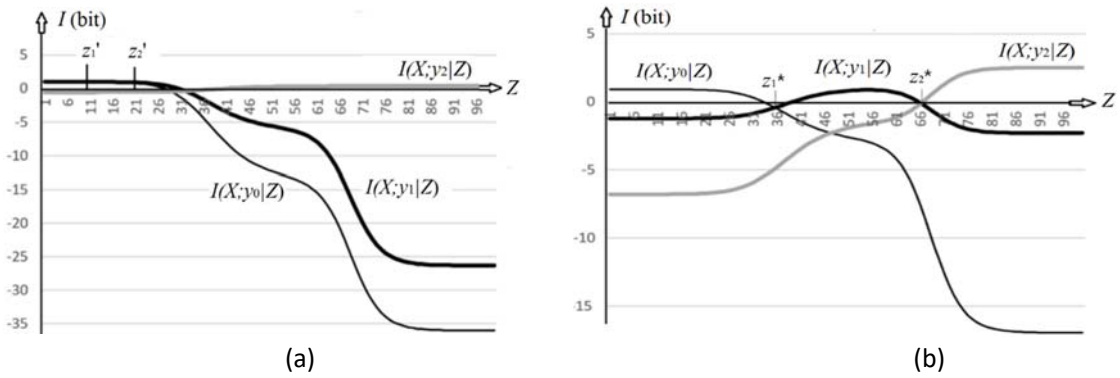


图 12 分界起点很差时的迭代. (a)显示迭代开始时三条信息曲线正的部分很小；(b)显示了迭代收敛时三条信息曲线正的部分较大.

Figure 12 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.

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.

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

Wittgenstein 有一个著名观点：“一个词的含义在于其应用”【56】。上面结论支持这种观点。

前面关于“老年人”例子中语义随实例先验分布 $P(X)$改变而改变，而上图显示的是：$P(X)$不变，但是实例不可见，语言理解和使用相互匹配导致进化。两中进化有所不同，前一种进化(两种信道随 $P(X)$改变)包含后一种进化。

6.3 CM 算法用于混合模型——严格收敛证明

笔者在【15】中介绍了 CM 算法用于混合模型，其中收敛证明虽然比和 EM 算法【21，22】即 VBEM 算法【23】的收敛证明好点，但是还不严格。下面提供更严格的证明。

假设样本分布 $P(X)$是某种条件概率分布(比如高斯分布)$P_{\theta^*}(X|Y)$按比例 $P_{\theta^*}(Y)$混合产生的. 我们只知道模型构件是 n 个. 要求的是 $P_{\theta^*}(Y)$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 $P_\theta(X)$和 $P(X)$尽可能接近(似然度尽可能大)，相对熵即 KL 距离 $H(P||P_\theta)$尽可能小.

CM 算法的基本思想是不断最小化 R-G，直至 R-G 接近 0， 这时候 $H(P||P_\theta)$也接近 0.

我们考虑高斯分布函数的混合. 迭代之前初始化 $P(Y)$(比如假设等概率)，初始化高斯分布函数

$$P(X|\theta_j)= K_j \exp[-(X-c_j)^2/(2d_j)^2], \quad j=1, 2, \ldots \tag{5.10}$$

中的参数 c_j 和 d_j. 其中 K_j 是归一化系数. 然后开始迭代运算.

混合模型的 CM 迭代算法(兼和 EM 算法比较)：

Left-step-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【21】，即

$$P(y_j \mid X) = P(y_j)P(X \mid \theta_j) / P_\theta(X),\ P_\theta(X) = \sum_k P(y_k)P(X \mid \theta_k),\ j=1, 2, \ldots \tag{5.11}$$

Left-step-b：调整 $P(Y)$(EM 算法中没有这一步，VBEM 这一步不同), 即轮流用下面两个公式做局部迭代：

$$\begin{aligned} &P(y_j) = \sum_i P(x_i)P(y_j \mid x_i) = \sum_i P(x_i)P(y_j \mid x_i),\ j = 1,2,... \\ &P(y_j \mid x_i) = P(x_i \mid \theta_j)P(y_j) / \sum_k P(y_k)P(x_i \mid \theta_k),\ i = 1,2,...; j = 1,2,... \end{aligned} \tag{5.12}$$

直至 $P(Y)$不变， 记为 $P^{+1}(Y)$. 这样做的原因是: 1)由(5.11) 产生的 Shannon 信道 $P(Y|X)$同 $P(X)$和 $P(X|\theta_j)$一般不相匹配, 即 $\sum_i P(x_i)P(Y|x_i) \neq P(Y)$. 重复上式也会减小相对熵 $H(P||P_\theta)$(后面证明); 2)这一步也能减小相对熵.

如果相对熵 $H(P||P\theta)$小于一个极小值， 比如 0.0001 比特, 则迭代结束. 否则继续.

Right-step：通过改变参数最大化语义互信息:

$$G = I(X;\theta) = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)} \tag{5.13}$$

然后转到**匹配 II-a**. 在 EM 算法中，这一步是最大化负的联合熵 Q=-NH(Y,$X|\theta$) 【21,22】，在 VBEM 中，这一步是最大 $Q+NP(Y) \approx NG$ 【23】.

CM 算法混合模型收敛证明：证明 $P_\theta(X)$ 收敛到 $P(X)$也就是证明 $H(P||P_\theta)$收敛到 0. 首先我们求 $H(P||P_\theta)$的表达式(EM 算法收敛证明没有提供)。为此先定义

$$R'' = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} \tag{5.14}$$

容易证明 $R''-G=H(P||P_\theta)$. 然后得到

$$\begin{aligned} R &= \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \left[\frac{P(x_i \mid \theta_j) P(y_j)}{P_\theta(x_i) P^{+1}(y_j)} \right] \\ &= R'' - H(Y^{+1}||Y) \\ H(Y^{+1}||Y) &= \sum_j P^{+1}(y_j) \log[P^{+1}(y_j) / P(y_j)] \end{aligned} \tag{5.15}$$

$$H(P \,||\, P_\theta) = R'' - G = R - G + H(Y^{+1} \,||\, Y) \tag{5.16}$$

CM 算法中的三步似乎正好分别改进 R, $H(Y^{+1}||Y)$和 G. 然而，收敛证明难在：当我们最小化 R 或 $H(Y^{+1}||Y)$时，其他两项也会改变。

因为在每一个 Left-step-b 之后, R-G= $H(P||P_\theta)$. 所以不断减小 R-G 就能使 G 接近 R 且 $H(P||P_\theta)$接近 0. 因为 Right-step 在最大化 G 的时候不改变 R，剩下的问题就是证明 Left-step a 和 Left-step b 如何最小化 R-G：

$$R - G = I(X;Y) - I(X;\theta) = \sum_i \sum_j P(x_i) P(y_j \mid x_i) [\log \frac{P(y_j \mid x_i)}{P(y_i)} - \log \frac{P(x_i \mid \theta_j)}{P(x_i)}] \tag{5.17}$$

我们仿照求解 Shannon 信息率失真函数 $R(D)$用到的变分方法和迭代方法(参看[ ], p. 316)。现在, 失真量 d_{ij} 变成语义信息量 $I(x_i;\theta_j)$= $\log[P(x_i|\theta_j)/P(x_i)]$；$R(D)$函数的参数 s 变成 1。因为 $P(Y|X)$ 和 $P(Y)$相互依赖, 为了最小化 $I(X;Y)$-$I(X;\theta)$, 我们用拉格朗日乘子法固定一个优化另一个。优化 $P(Y|X)$的限制条件是

$$\sum_j P(y_j \mid x_i) = 1, \; i = 1, 2, ..., n \tag{5.18}$$

优化 $P(Y)$的限制条件是

$$P(y_j) = \sum_i P(x_j) P(y_j \mid x_i), j = 1, 2 \tag{5.19}$$

所以拉格朗日函数是

$$F = I(X;Y) - I(X;\theta) - \mu_i \sum_j P(y_j \mid x_i) - \alpha \sum_j P(y_j) \tag{5.20}$$

为了优化 $P(Y|X)$，令 $\partial F / \partial P(y_j \mid x_i) = 0$．于是得到优化的 $P(Y|X)$(具体推导从略):

$$P^*(y_j \mid x_i) = P(y_j) P(x_i \mid \theta_j) / \sum_k P(y_k) P(x_i \mid \theta_k), \; i=1,2,\ldots,n; \quad j=1,2 \tag{5.21}$$

这正是 EM 算法中 E-step 和 CM 算法中 Left-step a 用到的公式.

为了优化 $P(Y)$，我们固定 F 中的 $P(y_j|x_i)$并且令 $\partial F / \partial P(y_j) = 0$．于是得到优化的 $P(Y)$:

$$P^*(y_j) = \sum_i P(x_j) P(y_j \mid x_i), \; j=1,2,\ldots,n \tag{5.22}$$

这正是 Left-step-b 用到的公式。它和式(5.21)一起在 Left-step b 中不仅最小化 R-G, 也最小化 $H(Y^{+1}||Y)$使之为 0。

容易证明，二阶偏导都大于 0， 所以 Left-step-a 和 Left-step-b 都最小化 R-G。从而减小

R''和相对熵 $H(P||P_\theta)$. **证毕**.

和 EM 算法的收敛证明比，CM 算法的收敛证明更加清晰，严格；迭代收敛也明显较快[15]. 下面是一个混合模型例子。这个例子是对 EM 算法收敛证明的的挑战。流行的 EM 算法收敛证明【22】认为不断增大 Q 可以达到目的，并且认为 Q 在 E-step 是不减的. 然而，下面这个例子中，Q 在 E-step 中可能会下降；正是因为 Q 在第一个 Right-step-a 之后下降，Q 才能接近收敛时的 Q. 具体数据和结果参看表 4 和图 13.

表 4 **真实和猜测的参数及迭代结果** ($R<R^*$)

Table 4 Real and guessed model parameters and iterative results for Example 2 ($R>R^*$)

	Real parameters			Starting parameters $H(P\|\|P_\theta)$=0.680 bit			Parameters after 5 Right-steps $H(P\|\|P_\theta)$=0.00092 bit		
	c	d	$P^*(Y)$	c	d	$P(Y)$	c	d	$P(Y)$
y_1	35	8	0.1	30	8	0.5	38	9.3	0.134
y_2	65	12	0.9	70	8	0.5	65.8	11.5	0.866

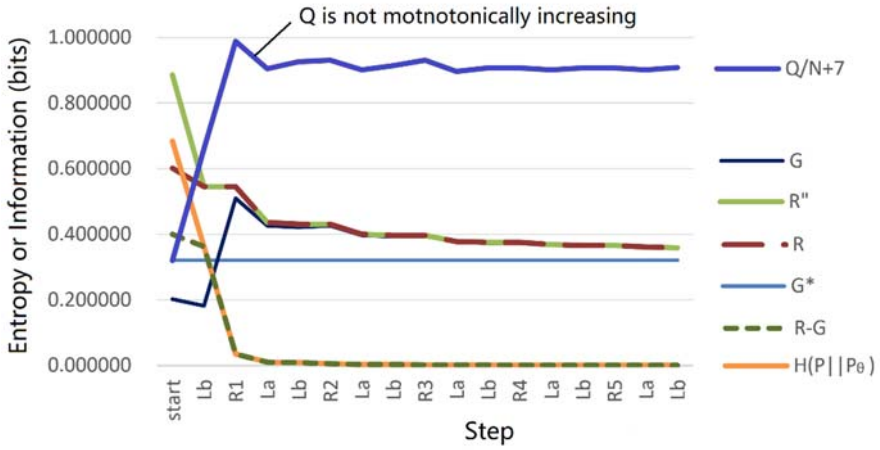


Fig. 13 The iterative process as $R>R^*$. $H(P||P_\theta)=R''-G$ decreases in all steps. R is monotonically decreasing. G increases in all Right-steps and decreases in all Left-steps. G and R gradually approach $G^*=R^*$ so that $H(P||P_\theta)=R''-G$ is close to 0. Five iterations are needed for convergence.

图 13 $R>R^*$时的迭代过程. $H(P||P_\theta)=R''-G$ 在每一步减小. R 单调减小， 而 G 在所有 Right-step 增大，在所有 Left-step 减小. G 和 R 逐渐接近 $G^*=R^*$，使得 $H(P||P_\theta)=R''-G$ 接近 0. 迭代 5 次后收敛.

实际迭代例子表明 CM 算法在大多数情况下收敛很快， 包括子模型分布重叠的情况. 马近文, 徐雷和 Jordnan【57】证明了 EM 算法在子模型分布不太重合的情况下有较好的渐进收敛率, 而在重合较多的情况下不然. 这个结论也是是可信的, 因为重合少时真模型的后验联合熵 $H^*(X,Y)$就小, 相应的联合似然度 Q^*就大, 不断增大 Q 就可以收敛.

混合模型求出以后，使用模型就是根据 X 选择 y_j. 可以采用式(5.7)。另外，我们也可以用真值函数构造混合模型，使得模型适合信源可变场合。对此需要进一步研究。

6 总结

本文讨论了 BI 的缺点，提出用新的数学框架优化语义通信和机器学习， 用真值函数或

语义信道代替贝叶斯后验。这一框架通过第三种贝叶斯定理，把已有的模糊集合论和概率论结合在一起， 进而用真值函数产生似然函数， 把两者带进 Shannon 信息公式得到语义信息公式， 再用语义信息公式优化机器学习。这样做的好处是：语义信息准则不仅等价于最大似然准则，也兼容正则化的最小误差平方(RLS)准则。机器学习被抽象为两个信道相互匹配。文中提出几种信道匹配(CM)算法，用以解决机器学习中三个比较困难的三个任务：多标签分类，最大似然估计（包括和不可见实例分类），和混合模型。文中提供了一些计算实例。理论分析和计算实践都表明，这些算法有能大大减少机器学习的工作量并提高学习的可靠性。

文中使用了 $R(G)$函数——它是信息率失真函数 $R(D)$的改进版本，G 取代 D 可理解为用正则化最小误差平方(RLS)取代失真。通过 $R(G)$函数，迭代的收敛证明严格而且直观， 特别是 CM 算法用于混合模型的收敛证明，明显优于 EM 算法的收敛证明。

新的数学框架主要用于估计、预测或分类的优化，但是只能支持或改进并不能取代已有的很多数学方法，比如深度学习、神经网络、概率图模型、梯度下降等方法。它也有局限性：1)它要求有较大样本， 以便把样本序列转换为较为均匀的样本分布；2)对于缺少 X 的先验 $P(X)$的场合，它可能不如 BI；3)要解决频率发生器(比如骰子)的频率问题，它也不如 BI。

新的数学框架中，不同类型的学习(包括标签学习和选标签分类)是：

- 求语义信道学习： 给定 $P(X)$，$P(Y|Y)$，求 $T(\theta|X)$和 $Y=f(X)$, 包括 $P(X)$迁移到 $P'(X)$的学习。对应无监督学习。
- 通过观察条件求两种信道学习：给定 $P(X)$，$P(X,Z)$，求最大互信息 Shannon 信道(即最优 $Y=h(Z)$)和语义信道。对应半监督学习。
- 通过信源 $P(X)$求两种信道(包括似然函数)学习：1)给定 $P(X)$和似然函数类型(比如类型是高斯分布)，求 $P(Y|X)$和 $P(X|\theta_j)=P(X)\,P(y_j|X)/P(y_j)$ (all j)。 对应无监督学习。

还有其他学习，有的能划到这三种学习，有的不能。比如很少标签的半监督学习，本文方法没有涉及。一个设想是把一个带标签实例当做一个假设，用假设的似然函数做新的样本分布。对此需要结合已有方法做进一步研究。

作为一个新的数学框架， 它一定有很多不完善地方，和其他数学方法的衔接也需要研究。这些都需要在实践中不断改进。欢迎更多研究者一起探索。

致谢 感谢汪培庄教授鼓励和支持(汪培庄教授是作者 1990 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,促使我三年前开始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 算法的问题和出路**

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 Stephen E. Fienberg, When Did Bayesian Inference Become “Bayesian"? Bayesian Analysis (2006) 1, Number 1, pp. 1-40.

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

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

4 Anon, Bayes’ Theorem, Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Bayes%27_theorem, Edited on 24 May 2018.

5 Anon, Bayesian Probability, Wikipedia: the Free Encyclopedia. http://en.wikipedia.org/wiki/Bayesian_probability, Edited on 7 May 2018.

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

7 R. A. Fisher, On the mathematical foundations of theoretical statistics. *Philo. Trans. Roy. Soc.,* A222, 1922:309-368.

8 C. E. Shannon, A mathematical theory of communication. *Bell System Technical Journal*. 1948, 27(3):379–429 and 623–656.

9 Anon, Maximum a posteriori estimation, Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation, edited on 16 May 2018.

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

11 Ian Goodfellow and Yoshua Bengio， Deep Learning, The MIP Press, Cambridge, 2016.

12 鲁晨光，广义信息论，中国科学技术大学出版社，合肥， 1993

13 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

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

15 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

16 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

17 Lu, C.: Semantic Information Measure with Two Types of Probability for Falsification and Confirmation, https://arxiv.org/abs/1609.07827 (Submitted on 26 Sep. 2016)

18 Lu, C.: Semantic Channel and Shannon Channel Mutually Match for Multi-label Classification, https://arxiv.org/abs/1805.01288 (Submitted on 2 May 2018).

19 Anon, Cross entropy, Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Cross_entropy, edited on 13 January 2018.

20 鲁晨光: B-模糊准布尔代数和广义交互熵公式，模糊系统和数学， 5(1),76-80 (1991).

21 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

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

23 Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. in: Michael I. Jordan (ed.) *Learning in Graphical Models, pp 355–368. MIT Press, Cambridge, MA (1999).*

24 鲁晨光，EM 算法的问题和出路，http://survivor99.com/lcg/CM/Recent.html ，Edited on May 1，2018

25 H. Akaike, A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 1974,

19:716–723.

26 T. M. Cover, and J. A. Thomas, *Elements of Information Theory, 2nd Edition*, New York: John Wiley & Sons, 2006.

27 Allen B. Downey, 贝叶斯思维：统计建模的 Python 学习法（许杨译），北京：人民邮电出版社，2015。

28 Hawthorne, J.: "Inductive Logic." In: Edward N. Zalta (ed.) Stanford Encyclopedia of Philosophy http://plato.stanford.edu/entries/logic-inductive. Edited on Mar 19, 2018.

29 Anon ， Confidence Interval ， Wikipedia: the Free Encyclopedia. ， https://en.wikipedia.org/wiki/Confidence_interval， Edited on 20 May 2018

30 Hájek, Alan, "Interpretations of Probability", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/entries/probability-interpret/#KolProCal，*First published Mon Oct 21, 2002; substantive revision Mon Dec 19, 2011*

31 Rudolf Carnap，Logical Foundations of Probability，Chicago：University of Chicago Press，1950。

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

33 L. A. Zadeh, Probability measures of fuzzy events, J. of Mathematical, Analysis and Applications, 23(2)(1986)421-427.

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

35 Davidson, D.: Truth and meaning. Synthese 17(1): 304-323 (1967).

36 汪培庄，模糊集合随机集落影，北京：北京师范大学出版社，1985.

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

38 Floridi L, Semantic conceptions of information, in Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/information-semantic/, last accessed 2015/1/7.

39 D'Alfonso, Simon, On Quantifying Semantic Information. Information, 2, 61-101 (2011)

40 Zhong, Y. X.: A theory of semantic information , China Communications, 14(1), 1-17 (2017).

41 *Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, J.: Generative Adversarial Networks.* arXiv:1406.2661 *[cs.LG] (2014).*

42 S. Kullback, R.A. Leibler, On information and sufficiency, Annals of Mathematical Statistics , 22 (1)(1951)79–86.

43 Min-Ling Zhang，Yu-Kun Li，Xu-Ying Liu，Xin Geng，Binary relevance for multi-label learning: an overview，Frontiers of Computer Science， April 2018, Volume 12, Issue 2, pp 191–202

44 Charlie Frogner, Bayesian Interpretations of Regularization, Bayesian Analysis (2006) 1, Number 1, pp. 1-40

45 周志华，机器学习，清华大学出版社，北京， 2016

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

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

48 鲁晨光, 色觉的译码模型及其验证, 光学学报，9(2)(1989), 158-163.

49 Anon ， Sensitivity and specificity, Wikipedia: the Free Encyclopedia, https://en.wikipedia.org/wiki/Sensitivity_and_specificity, Edited on 25 May 2018.

50 Thornbury, J.R., Fryback, D. G., Edwards, W.: Likelihood Ratios as a Measure of the Diagnostic Usefulness of Excretory Urogram Information. Radiology 114(3), 561–565 (1975).

51 Anon，Likelihood-ratio test ，Wikipedia: the Free Encyclopedia，https://en.wikipedia.org/wiki/Likelihood-ratio_test，edited on 15 May 2018.

52 Tentori K, Crupi V, Bonini N, Osherson D. Comparison of confirmation measures[J]. Cognition, 2007,103:107–119

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

54 T. Berger, Rate Distortion Theory, Enklewood Cliffs, N.J. Prentice-Hall, 1971.

55 周炯槃，信息理论基础，中国邮电出版社，北京，1983。

56 Wittgenstein, L.: Philosophical Investigations, Basil Blackwell Ltd, Oxford (1958).

57 Yu J, Chaomu C, Yang M S. On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures[J]. Pattern Recognition, 77(5): 188-203

From Bayesian Inference to Logically Bayesian Inference: A New Mathematical Framework for Semantic Communication and Machine Learning

LU，Chenguang

College of Intelligence Engineering and Mathematics, Liaoning Engineering and Technology University,
Fuxin, Liaoning, 123000, China
lcguang@foxmail.com

Abstract: Bayesian Inference (BI) uses Bayes' posterior whereas Logical Bayesian Inference (LBI) uses truth function or membership function as the inference tool. The LBI is proposed mainly because the BI is not compatible with classical Bayes' prediction and has not used logical probability in fact and hence cannot express semantic meaning. In the new mathematical framework, statistical probability and logical probability are strictly distinguished, used at the same time, and linked by the third kind of Bayes' theorem newly found so that the likelihood function and the truth function can be converted from one to another; a semantic channel that consists of a group of truth functions can be directly derived from a Shannon channel. The semantic communication model is used as the machine learning model. It has two parts: receivers' label learning which lets semantic channel match Shannon's channel and senders' label selection or classification which lets Shannon's channel match semantic channel. The Maximum Semantic Information (MSI) criterion is equivalent to the Maximum Likelihood (ML) criterion, and compatible with the Regularized Least Square (RLS) criterion. The two channels' mutual matching can conveniently achieve multi-label classifications with the considerations of the class imbalance and the changed prior distribution of instances, without the need to consider the binary relevance. Using the Channels' Matching (CM) algorithm, an iterative method, we can achieve unseen instance classifications and maximum likelihood estimations，which belongs to semi-supervised learning. Differently, the CM algorithm for mixture models, which belongs to unsupervised learning, repeatedly minimizes the difference between Shannon's mutual information and semantic mutual information. The convergences of two kinds of iterations can be strictly proved by $R(G)$ function, wich is the improved rate distortion function $R(D)$, where G is semantic mutual information and can be understood as negative regularized distortion. Some examples of applications to supervised, semi-supervised, and unsupervised learning are provided. Theoretical analyses and actual computations shows that in most cases of machine

learning, the LBI can perform better than the BI. Finally, the LBI's limitations are discussed.
Keywords： Bayes' theorem; Logical probability; Truth function; Semantic communication; Machine learning; Multi-label classification； Maximum likelihood estimation; Mixture models