P-T 概率框架用于语义通信、证伪、确证和贝叶斯推理[①]

鲁晨光

摘要：很多研究者想通过合理地定义逻辑概率或概率逻辑统一概率和逻辑；而本文试图统一统计和逻辑，以便我们可以同时使用统计概率和逻辑概率。为此，本文提出 P-T 概率框架——它是用香农(Shannon)的概率框架（用于电子通信）、柯尔莫哥洛夫(Kolmogorov)的概率公理（用于逻辑概率）和扎德(Zadeh)的隶属函数（用作真值函数）组装的。推广的贝叶斯定理可以连接两种概率，用它我们可以将似然函数和真值函数相互转换；据此，我们可以用统计中的样本分布训练逻辑中的（模糊）真值函数。这一概率框架是在作者在语义信息、统计学习和颜色视觉的长期研究中发展而成的。本文首先介绍 P-T 概率框架并且通过其应用于语义通信解释其中不同概率，然后解释我们如何将这一概率框架和语义信息方法应用于统计学习、统计力学，假设评价（包括证伪）、确证和贝叶斯推理。这些理论上的应用显示了这一概率框架的合理性和实用性。这一概率框架将有助于可解释的人工智能。为解释神经网络，我们需要进一步研究。

关键词：统计概率；逻辑概率；语义通信；信息率失真；Boltzmann 分布；证伪；逼真度；确证测度；乌鸦悖论；贝叶斯推理。

1 引言

1.1 有不同概率存在还是一种概率有不同解释？

我们可以发现概率有几种不同解释：频率主义的、逻辑的、主观的和倾向性的解释[1,2]。差不多每个学派都相信它的解释是正确的。然而，我们可以看到不同学派有不同优点。比如，香农（Shannon）信息论[3]是客观的频率主义概率的成功例子。最大似然方法[4]是统计学习不可缺少的工具，它使用客观概率（样本分布）也使用主观概率（似然函数）。使用贝叶斯推断（Bayesian Inference）[5]的贝叶斯主义者不仅相信概率是主观的，还相信预测模型(θ)是主观的并且有其先验分布。贝叶斯推断也有许多成功的应用。使用倾向性解释[6]，我们可以解释物理世界随机性的客观性和优化的似然函数的客观性。如 Hájek [2]指出，上面的多种解释应该是互补的；我

[①] 本文是应邀发表的英文文章 The P–T Probability Framework for Semantic Communication, Falsification, Confirmation, and Bayesian Reasoning 的中译本（略有修改）。原文提供 CC Common 许可证. 见 https://www.mdpi.com/2409-9287/5/4/25.

们需要“保留物理逻辑的(或证据的)以及主观的概率的独特概念，用一个丰富多彩的花环把它们连接在一起。”本文努力也是朝这一方向。

在这个方向遇到的第一个问题就是：只存在有一种概率——它们有几种不同解释——还是存在几种不同概率？在《科学发现的逻辑》中[7](pp.252-258)，波普尔（Popper）区分了两种概率：一个假设被判断为真的概率和一个事件发生的概率。他认为两种概率是不同的。即使我们考虑一个假设的概率，我们也不能像莱辛巴哈（Reichenbach）那样，用一个概率同时表示假设的逻辑概率和它出现的概率[7](p.252)。卡尔纳普（Carnap）[8]也区分概率 1（逻辑概率）和概率 2（统计概率）。

Galavotti [1]在回顾概率的历史时写道：

“自从概率诞生起，它就带有奇怪的双重意义的特点。如 Kacking 所描述: 概率是‘两面神雅努斯的面孔。一面是统计的，涉及偶然过程的随机性定律；另一面是认识论的，用于评估命题的合理的可信度——偏离统计背景。’”

显然，波普尔和卡尔纳普并不孤单。我同意波普尔和卡尔纳普的观点：存在两种不同的概率——统计概率和逻辑概率。我认为统计概率有频率主义的、主观的和倾向性解释。

一个假设或标签的统计概率也就是该假设被选择的概率。我们用一个例子解释一个假设有两个不同概率：被选择的概率和逻辑概率。

例 1. 假定有五个标签：y_1=“小孩”、y_2=“年轻人”、 y_3=“中年人”、y_4=“老年人”和 y_5=“成年人”。 假定有一万个有不同年龄的人通过一道门，让门卫来判断来人 x 是否是成年人或判断“x 是成年人”是否是真的。如果有 7000 人被判定为成年人，那么“成年人”的逻辑概率就是 7000/10000=0.7。如果门卫的任务改为: 从五个标签选一个并为每个人贴上一个标签。这时候可能只有 1000 个人被贴上“成年人”标签。“成年人”的统计概率是 1000/10000=0.1。为什么 y_5 被选择的概率小于其逻辑概率？原因是：其他 6000 人被贴上了“年轻人”，“中年人”或“老年人”的标签。换句话说，原因是：一个人只能使一个标签的选择概率增加 1/10000， 但是却能使两个或多个标签的逻辑概率增加 1/10000。比如，一个 20 岁的人可以同时使“年轻人”和“成年人”的逻辑概率增加 1/10000。

一个极端的例子是：一个永真句——比如“x 是成年人或不是成年人”——的逻辑概率是 1，而其统计概率差不多是 0，因为它极少被选择。

从上面例子我们能看出：

- 一个假设或标签 y_j 有两种概率：逻辑概率和被选择概率。如果我们用 $P(y_j)$ 表示 y_j 的统计概率就不能用 $P(y_j)$表示 y_j 的逻辑概率。
- 统计概率是归一化的（总和是 1），而逻辑概率不是归一化的；一般情况下，逻辑概率大于统计概率。
- 给定一个人的年龄的时候，五个标签的真值之和也大于 1。

- “成年人”的逻辑概率和人口年龄的先验概率分布 $P(x)$（它是统计概率）相关。显然，从一个学校大门和一个医院大门获得的“成年人”的逻辑概率是不同的。

用上面方法计算的逻辑概率符合赖欣巴哈的概率的频率主义定义[9]。但是赖欣巴哈并不区分逻辑概率和统计概率。

在流行的概率系统中，比如柯尔莫哥洛夫（Kolmogorov）[10]定义的概率（后面简称柯氏概率）公理系统中，所有概率都用“P”表示，使得我们不能同时使用两种不同概率，也不便确定“P”表示统计概率还是逻辑概率，比如在流行的确证测度中[11]。

波普尔抱怨柯尔莫哥洛夫的概率系统说：“他假定在 $P(a, b)$中，a 和 b 是集合，这样就排除了其他解释比如逻辑解释——据此 a 和 b 是叙述（或命题，如你愿意）。”[7](pp. 330-331）这里 $P(a, b)$是条件概率，现在应写作 $P(a|b)$。

在《猜想和反驳的新附录》(New Appendices of Conjectures and refutations [7], pp. 329-355)中，波普尔提出自己的概率公理化系统. 在他的系统中，$P(a|b)$ 中的 a 和 b 可以是集合、谓词、命题、等等。然而，波普尔也还是只用“P”表示概率， 因而他的公理系统并不同时使用统计概率和逻辑概率。 因为这一原因，波普尔的概率系统并不比柯氏概率系统更加实用。

为了区分统计概率和逻辑概率并且同时使用两者，我[12]提出用两个不同的符号“P”和“T”区别统计概率（包括似然函数）和逻辑概率（包括真值函数）。在我早期文章中[13-15]，我用“P”表示客观概率，用 Q 表示主观概率和逻辑概率。因为主观概率也是归一化的并且是用“=”而不是 ϵ（属于）定义的（见 2.1 节），我现在仍然用“P”表示主观概率。

1.2 推广逻辑遇到的问题：一个概率函数会是一个真值函数?

杰尼斯（Jaynes） ([16], pp.651-660)做了类似于波普尔的努力——改进柯氏概率系统。他是一个著名的贝叶斯主义者，但是也承认我们需要频率主义的概率，比如表示样本分布的概率。他证明了玻尔茨曼（Boltzmann）熵和香农熵之间存在等价关系，其中概率是频率主义的或客观统计的。他总结道：概率理论是逻辑的推广。在他的著作中[16], 他用了很多例子说明很多科学推理是概率的。 然而，他的概率系统缺少真值函数——它们是多值的或模糊的——作为推理工具。

一个命题有一个真值；含有相同谓词和不同主语的所有命题的真值构成一个真值函数。在经典逻辑（二值逻辑）中，一个集合的特征函数也就是一个假设的真值函数。假定有一个谓词（或命题函数）$y_j(x)$ = “x 有性质 a_j"， 且集合 A_j 包含具有性质 a_j 的所有 x。则 A_j 的特征函数就是 $y_j(x)$的真值函数. 比如，x 代表一个年龄（或一个年龄为 x 的人），年龄≥18 的人被定义为“成年人”。那么，包含所有年龄大于或等于 18 岁的人的集合是 A_j=[18, ∞). 它的特征函数也就是谓词“x 是成年人”的真值函数 (见图 1).

如果我们推广二值逻辑到连续值逻辑，真值函数应该在 0 和 1 之间变化。假定一组年龄在 60 岁的人的个数是 N_{60}. 其中有 N_{60}*个人被称为年老的，那么命题“一个 60 岁的人是年老的”的真值就是 N_{60}*/N_{60}。这个真值大概是 0.8。如果 x= 80, 则真值应该是 1.

根据概率的逻辑解释，一个概率应该是一个连续的逻辑值（真值），并且一个条件概率函数 $P(y_j|x)$ （带有可变的 x）是一个真值函数（即模糊真值函数）。香农称 $P(y_j|x)$为转概率函数(TPF,即 Transition Probability Function)[3](p.11)。下面我们用一个例子解释为什么转移概率函数并不是真值函数，解释人脑怎样使用概念外延（能用真值函数表示）思考。

例 2 （参看图 1）. 假定我们知道人口年龄先验分布 $P(x)$和后延分布 $P(x|$"成年人") and $P(x|$"老年人"), “成年人(adult)”的外延是清晰的而“老年人（elder）”的外延是模糊的。求解标签“成年人”和“老年人”的真值函数或外延。

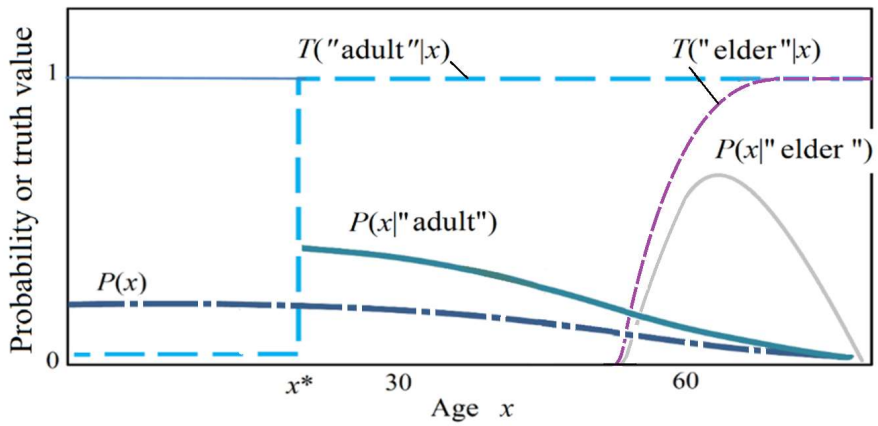


图 1. 求解标签“adult”（“成年人”）和“elder”（“老年人”）的真值函数——用先验和后延概率分布。人脑可以猜测“adult” 的真值函数 T(“adult”|x)或外延 (粗虚线)——根据 $P(x|$”adult”)和 $P(x)$，并且估计“elder” 的真值函数 T(“elder”|x) (细虚线)或外延——根据 $P(x|$”elder”)和 P(x). 等式 (12)和 (22)是计算 T(“adult”|x)和 T(“elder”|x)的新公式.

根据现存的概率理论，包括 Jaynes 的概率理论，我们不能从 $P(x)$ 和 $P(x|y_j)$获得真值函数，也不能获得转移概率函数 $P(y_j|x)$。因为根据贝叶斯定理 $P(y_j|x)=P(x|y_j)P(y_j)/P(x)$，没有 $P(y_j)$我们得不到 $P(y_j|x)$。但是人脑能估计 y_j 的外延——从 $P(x)$和 $P(x|y_j)$，虽然没有 $P(y_j)$ (见图 1). 用这样的外延，即使 $P(x)$变了，人脑仍然能预测 x 的后延概率分布并且给 x 分类。

这个例子说明：

- 现存的概率理论缺乏人脑使用的方法——1）估计一个标签的外延（归纳方法）；2）使用该外延为条件做推理或预测（演绎方法）。

- 转移概率函数和真值函数是不同的，因为一个转移概率函数和多少标签使用有关，而真值函数与此无关。

1.3 区分（模糊）真值和逻辑概率——使用扎德的模糊集合论

Dubois 和 Prade [17]指出：存在一个混淆传统——研究者经常混淆假设的可信度（逻辑概率）和命题的真值；要做逻辑运算，我们只能用真值而不能用可信度（degree of belief）或逻辑概率。我同意他们这一看法。7.2 节将讨论为什么我们不能直接用逻辑概率做逻辑运算。下面我们进一步区分逻辑概率和（连续或模糊的）真值。

语句 "x 是年老的"（x 是一群人中的一个）是一个命题函数，而"那个男人是年老的"是一个命题。前者有一个逻辑概率，后者有一个真值。因为命题函数包含一个谓词（predicate）"x 是年老的 y"和 x 的论域（x 是一群人中的一个），我们也称命题函数为谓词。因此，我们需要区分谓词的逻辑概率和命题的真值。

然而，在上面谈到的概率系统中，要么我们不能区分谓词的逻辑概率和命题的真值（比如在柯氏概率系统中），要么真值是二元的（比如在卡尔纳普[18]和赖欣巴哈 [9]的概率系统中）。

在卡尔纳普的概率系统中[18]，谓词的逻辑概率和先验概率分布 P(x)不相关。从例 1 看，两者是相关的。

赖欣巴哈[9]使用二元逻辑的统计结果定义两个命题序列（即谓词）的逻辑表达式的逻辑概率，没有使用连续的真值函数；并且没有区分逻辑概率和统计概率。

因此，上述概率系统作为逻辑的推广是不能令人满意的。有幸的是，扎德的模糊集合论[19,20]比上面的概率系统提供了更多我们需要的东西。 因为实例 x_i 在模糊集合 θ_j 上的隶属度就是命题 $y_j(x_i)$="x_i 在 θ_j 中"的真值。一个模糊集合 θ_j 的隶属函数就等价于一个谓词 $y_j(x)$的真值函数。扎德[20]提出的模糊事件的概率就是谓词的逻辑概率（可惜他仍然用"P"表示）。扎德[21]认为模糊集合理论和概率论是互补的而不是竞争的；模糊集合论可以丰富我们的概率概念。

然而，让一个机器获得真值函数或隶属函数仍然是困难的。

1.4 我们能用样本分布优化真值函数或隶属函数吗？

虽然模糊集合论在知识表达、模糊推理和模糊控制领域取得了显著成绩，但是用它做统计学习还是不行，因为隶属函数通常是专家定义的而不是从统计获得的。

一个样本包含许多样例。比如(x: 年龄, y：标签)是一个样例。一个样本中有许多样例，比如：(5, "小孩"), (20, "年轻人"), (70, "老年人") …。后验概率分布 $P(x|y_j)$表示一个样本分布或子样本分布。统计学习的核心方法是用一个样本分布 $P(x|y_j)$优化一个似然函数 $P(x|\theta_j)$（θ_j 是一个预测模型或模型参数）。我们也希望能用样本分布优化真值函数或隶属函数，以便连接统计和逻辑。

在这个方向上一个有意义的进步是汪培庄[22,23]提出的随机集落影理论。根据该理论，一个模糊集合是一个随机集合产生的；随机集合的取值是一个范围，比如

“年轻人”的随机集合取值是 15-30 或 18-28。有了许多这样的范围，我们就能计算任一 x 在这些集合上中的比例。比例的极限就是 x 在相应的模糊集合上的隶属函数。

然而，要获得真值函数或隶属函数仍然是困难的，原因是现实生活中存在很多单实例的样例，比如(70, "老年人"), (60, "老年人"), 和 (25, "年轻人"), 但是很少存在带有标签的范围——比如(15-30, "年轻人")——作为样例。

因为上面原因，模糊数学在统计学习中的表现远远不如在专家系统中的表现那么好；很多统计学家拒绝使用模糊集合。

1.5 目的、方法和本文结构

很多研究者想统一概率和逻辑。一个自然的想法是合理地定义一个逻辑概率系统，如凯恩斯(Keynes)[24]、赖欣巴哈[9]、波普尔[7]和卡尔纳普[18]所做的那样。有些研究者使用概率作为逻辑变量建立概率逻辑[25], 比如赖欣巴哈 [9] 和 Adams [26] 建立的两种概率逻辑。也有许多研究者继续扎德的努力，统一概率和模糊集合[27,28,29]。参看[25]可见很多将经典逻辑概率化的努力.

我的努力有些不同。我也继续扎德的努力将经典逻辑概率化（即模糊化），但是我想同时使用统计概率和逻辑概率并且使两者兼容——这意味着在样本很大时，逻辑推理的结果等于统计推理的结果。简言之，我想统一统计和逻辑，而不仅是概率和逻辑。

一个概率框架是一个由若干随机变量（取值于一个或多个论域）的概率构成的数学模型。比如，香农的通信数学模型是一个概率框架。P-T 概率框架包含统计概率（用“P”表示）和逻辑概率（用“T”表示）。我们可以称香农的概率框架为 P 概率框架。

本文的目的是提出 P-T 概率框架并检验这一框架——通过其应用于语义通信、统计学习、统计力学、假设评价（包括证伪）、确证和贝叶斯推理。

主要方法是：

1. 使用统计概率框架即香农使用的 P 概率框架，加上柯氏概率公理（为了逻辑概率）和扎德的隶属函数（用作真值函数），形成 P-T 概率框架；

2. 用新的贝叶斯定理，即贝叶斯定理 3，建立真值函数和似然函数之间的联系，以便我们用（统计中的）样本分布优化（逻辑中的）真值函数；

3. 使用 P-T 概率框架和语义信息公式 1）表达逼真度和检验严厉性，2）推导出两个实用的确证测度和几个新公式，从而丰富贝叶斯推理（包括模糊三段论）。

P-T 概率框架是在我先前的语义信息 G 理论（后面简称 G 理论，G 意味着香农信息论的推广）[12,14,15] 和统计学习研究中逐步形成的。我在本文首次把它命名为 P-T 概率框架，并且连同其哲学背景和应用一并介绍。

本文其余本分结构如下。第 2 节定义 P-T 概率框架。第 3 节解释这一框架——通过它在语义通信、统计学习和统计力学中的应用。第 4 节显示这一框架和语义信

息测度如何支持波普尔关于科学进步和假设检验的思想。第 5 节介绍两个实用的确证测度——分别用于医学检验和澄清乌鸦悖论。第 6 节总结各种贝叶斯推理公式并介绍一个兼容布尔代数的模糊逻辑。第 7 节讨论一些问题——包括用理论应用说明 P-T 概率框架的合理性和实用性。第 8 节是结论。

2 P-T 概率框架

2.1 香农使用的概率框架（用于电子通信）

香农使用频率主义的概率系统[30]构建 P 概率框架，它包含两个随机变量和两个论域：

定义 1 （香农用的 P 概率框架）：

- X 是一个离散随机变量，它取之 $x \in \mathcal{X}$, $\mathcal{X}$ 是一个论域 $\{x_1, x_2, ..., x_m\}$; $P(x_i) = P(X=x_i)$ 是事件 $X = x_i$ 的相对频率的极限。在下面应用中，x 表示一个实例或一个样本点.
- Y 是一个离散随机变量，它取值 $y \in \mathcal{Y} = \{y_1, y_2, ..., y_n\}$; $P(y_j) = P(Y=y_j)$. 在下面应用中，y 表示一个标签，假设或一个谓词.
- $P(y_j|x) = P(Y=y_j|X=x)$是一个转移概率函数(Transition Probability Function (TPF)，香农如此命名 [3]).

香农称 $P(X)$是信源，$P(Y)$是信宿，$P(Y|X)$是信道，见图 2（a）. 一个香农信道是一个转移概率矩阵或一组转移概率函数：

$$P(Y \mid X) \Leftrightarrow \begin{bmatrix} P(y_1 \mid x_1) & P(y_1 \mid x_2) & ... & P(y_1 \mid x_m) \\ P(y_2 \mid x_1) & P(y_2 \mid x_2) & ... & P(y_2 \mid x_m) \\ ... & ... & ... & ... \\ P(y_n \mid x_1) & P(y_n \mid x_2) & ... & P(y_n \mid x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} P(y_1 \mid x) \\ P(y_2 \mid x) \\ ... \\ P(y_n \mid x) \end{bmatrix}, \quad (1)$$

⇔意思表示等价。一个香农信道或一个转移概率函数可以被当做一个预测模型——产生 x 的后验概率分布: $P(x|y_j), j=1,2,...,n$，见等式 6。

2.2 P-T 概率框架用于语义通信

定义 2（P-T 概率框架）：

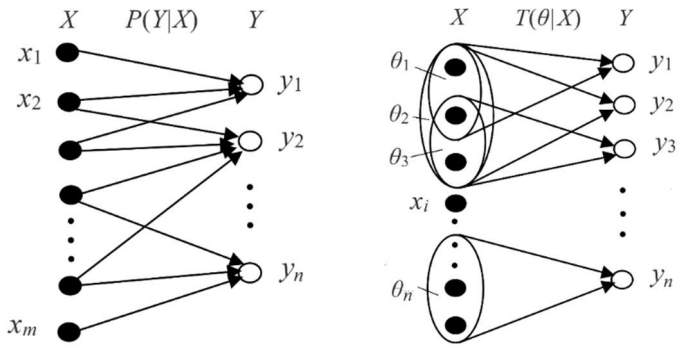
- y_j 是一个标签或假设, $y_j(x_i)$是一个命题, $y_j(x)$是一个命题函数. 我们也称 y_j 或 $y_j(x)$是一个谓词. θ_j 是论域 $\mathcal{X}$上的模糊子集, 它被用来接收命题函数 $y_j(x)$ 的语义($y_j(x)$= *"$x \in \theta_j$"*= *"x 属于 θ_j"="x 在 θ_j 中"*). θ_j 也被当做一个预测模型或一组模型参数——用于构造真值函数.
- 用"="定义的概率是统计概率, 比如 $P(y_j) = P(Y = y_j)$是统计概率；用 "$\in$"定义的概率是逻辑概率，比如 $P(X \in \theta_j)$是逻辑概率. 为了区分 $P(Y = y_j)$和 $P(X \in \theta_j)$, 我们定义 $T(y_j)=T(\theta_j) = P(X \in \theta_j)$ 为 y_j 的逻辑概率。

- $T(y_j|x) = T(\theta_j|x) = P(x \in \theta_j) = P(X \in \theta_j|X = x)$是 y_j 的真值函数，也是 θ_j 的隶属函数。它在 0 和 1 之间变化，并且最大值是 1。

根据塔斯基(Tarski)的真理论 [31], $P(X \in \theta_j)$等价于 P(“$X \in \theta$” 是真的) = $P(y_j$ 是真的). 根据 Davidson 真值条件语义学 [32], y_j 的真值函数确定了 y_j 的语义（形式语义）。 一组真值函数 $T(\theta_j|x)$ (j = 1,2…n) 构成一个语义信道:

$$T(\theta \mid X) \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid x_1) & T(\theta_1 \mid x_2) & ... & T(\theta_1 \mid x_m) \\ T(\theta_2 \mid x_1) & T(\theta_2 \mid x_2) & ... & T(\theta_2 \mid x_m) \\ ... & ... & ... & ... \\ T(\theta_n \mid x_1) & T(\theta_n \mid x_2) & ... & T(\theta_n \mid x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid x) \\ T(\theta_2 \mid x) \\ ... \\ T(\theta_n \mid x) \end{bmatrix}. \tag{2}$$

图 2 图解了香农信道和语义信道。



(a) (b)

图 2. 香农信道和语义信道。 (a)P 概率框架描述香农信道; (b)P-T 概率框架描述语义信道。一个隶属函数确定了 y_j 的语义，一个模糊集合 θ_j 可能被另一个部分覆盖或包括。

现在我们介绍逻辑概率和柯尔莫哥洛夫概率之间的联系. 在柯氏概率系统中，一个随机变量所取的值是一个集合。我们用 $2^{\mathcal{X}}$ 表示 $\mathcal{X}$ 的幂集（波乃尔集）——它包括 $\mathcal{X}$ 所有可能的子集；Θ 表示随机变量——它取值 $\theta \in \{\theta_1, \theta_2, \ldots, \theta_{|2^{\wedge}\mathcal{X}|}\}$. 那么柯尔莫哥洛夫概率 $P(\theta_j)$就是事件 $\Theta=\theta_j$ 的概率，因为 $\Theta=\theta_j$ 等价于 $X \in \theta_j$, 我们有 $P(\Theta=\theta_j)=P(X \in \theta_j)=T(\theta_j)$. 如果 $2^{\mathcal{X}}$ 中所有集合是清晰的，那么 $T(\theta_j)$ 就变成柯尔莫哥洛夫概率.

根据上面定义，我们有 y_j 的逻辑概率：

$$T(y_j) = T(\theta_j) = P(X \in \theta_j) = \sum_i P(x_i) T(\theta_j \mid x_i). \tag{3}$$

据此，逻辑概率是平均真值；真值是后延逻辑概率。注意：一般说来，一个命题只有真值而没有逻辑概率。比如，“每个 x 是白的”（其中 x 是一个天鹅）是一个谓词，它有逻辑概率；而“那个天鹅是白的”只有真值。统计概率 $P(y_j)$等于逻辑概率 $T(\theta_j)$（对于每个 j）只有在下面两个条件成立时：

集合 θ 的论域只包含 $2^{\mathcal{X}}$ 的某些子集且这些子集构成 $2^{\mathcal{X}}$ 的一个划分——意味着任何两个子集是不相交的。

- 每个 y_j 总是被正确选择。

比如，Y 的论域是 $\mathcal{Y}$={“非成年人”, “成年人”} 或 {“小孩”, ”年轻人”, ”中年人”, ”老年人”}, $\mathcal{Y}$ 确定了 $\mathcal{X}$ 的一个划分. 如果这些标签总是被正确选择，则 $P(y_j)=T(\theta_j)$（对于每个 j）。许多人并不区分统计概率和逻辑概率是因为他们假定这两个条件总是成立的。而实际上这两个条件一般是不成立的。很多人把柯氏概率系统的应用到许多场合时并没有区分统计概率和逻辑概率，这时上述两个条件是必要的。然而，后一个条件很少被提到。

卡尔纳普和 Bar-Hillel [33] 定义的逻辑概率不同于 $T(\theta_j)$. 假定有三个原子命题 a, b, c. 一个最小项，比如 $ab\bar{c}$ (a 且 b 且非 c)的逻辑概率是 1/8; $b\bar{c}=ab\bar{c}\vee\bar{a}b\bar{c}$ 的逻辑概率是 1/4. 这样的逻辑概率和 x 的先验概率分布不相关。然而，逻辑概率 $T(\theta_j)$ 和 $P(x)$是相关的。一个较小的逻辑概率不仅需要较小的外延，也需要罕见的实例。比如, “x 是 20 岁”比 “x 是年轻的” 有较小的逻辑概率，但是 “x 大于 100 岁” 有较大的外延但是也有较小的逻辑概率（和“x 是 20 岁”比）， 因为大于 100 岁的人是稀有的。

在数理逻辑中，一个全称谓词的真值等于所有含有该谓词的所有命题的逻辑乘，因此波普尔和卡尔纳普得出结论：一个全称谓词的逻辑概率是 0。 这个结论是怪异的。 但是, $T(\theta_j)$是谓词 $y_j(x)$自身（没有量词，比如没有∀x）的逻辑概率. 它是平均真值。 为什么我们这样定义逻辑概率？一个理由是：这一数学定义符合字面定义——逻辑概率是一个假设被判断为真的概率。 另一个理由是这样定义有助于推广贝叶斯定理。

2.3 三种贝叶斯定理

有三种贝叶斯定理，它们分别是：贝叶斯[34]定义的、香农使用的 [3]和我发现的 [12]。

贝叶斯定理 1: 它是贝叶斯提出的，可用柯氏概率表达。 假定有两个集合 A 和 $B \in 2^X$; A^c 和 B^c 是 A 和 B 的两个补集. 我们可用两个对称的公式表达这一定理:

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)= T(A|B)T(B)+ T(A|B^c)T(B^c), \tag{4}$$

$$T(A|B)=T(B|A)T(A)/T(B),\ T(B)= T(B|A)T(A)+ T(B|A^c)T(A^c). \tag{5}$$

注意：这里有一个随机变量，一个论域，两个逻辑概率。

贝叶斯定理 2：它可用香农使用的频率主义的概率表达。这里有两个对称的公式：

$$P(x|y_j)=P(y_j|x)P(x)/P(y_j),\ P(y_j)=\sum_i P(y_j|x_i)P(x_i), \tag{6}$$

$$P(y_j|x)=P(x|y_j)P(y_j)/P(x),\ P(x)=\sum_j P(x|y_j)P(y_j). \tag{7}$$

注意：这里有两个随机变量，两个论域，两个统计概率。

贝叶斯定理 3：它用在 P-T 概率框架中。这里有两个不对称的公式：

$$P(x|\theta_j)=T(\theta_j|x)P(x)/T(\theta_j),\ T(\theta_j)=\sum_i P(x_i)T(\theta_j|x_i), \tag{8}$$

$$T(\theta_j|x)=T(\theta_j)P(x|\theta_j)/P(x),\ T(\theta_j)=1/\max[P(x|\theta_j)/P(x)]. \tag{9}$$

两个公式是不对称的是因为这里（等式左边）存在一个统计概率和一个逻辑概率。两个公式的证明见附录 I。

等式（8）中 $T(\theta_j)$是横向归一化常数，它使 $P(x|\theta_j)$的和为 1；而等式(9)中 $T(\theta_j)$是纵向归一化常数，它使 $T(\theta_j|x)$的最大值为 1.

我们称等式（6）中 $P(x|y_j)$ 为贝叶斯预测, 称等式（8）中 $P(x|\theta_j)$是语义贝叶斯预测，称等式（8）是语义贝叶斯公式。 在文献[27]中, Dubois 和 Prade 提到过一个类似于等式（8）的公式——由 S. F. Thomas 和 M. R. Civanlar 更早提出.

等式（9）可以直接写成

$$T(\theta_j|x)=[P(x|\theta_j)/P(x)]/\max[P(x|\theta_j)/P(x)]. \tag{10}$$

2.4 统计概率和逻辑概率之间的匹配关系

怎样理解一个标签的外延？我们可以从样本学习。

设 **D** 是一个样本：$\{(x(t), y(t))|t = 1 \text{ to } N; x(t) \in \boldsymbol{X}; y(t) \in \boldsymbol{Y}\}$, 其中 $(x(t), y(t))$是一个样例；每个标签的使用差不多是合理的。**D** 中所有带有标签 y_j 的样例构成一个子集 $\mathbf{D}_j$.

如果 $\mathbf{D}_j$ 足够大，我们可以从 $\mathbf{D}_j$ 获得平滑的样本分布 $P(x|y_j)$. 根据费希尔(Fisher)的最大似然估计, 在 $P(x|y_j)=P(x|\theta_j)$时，$\mathbf{D}_j$ 和 θ_j 之间的似然度最大。因此，我们可以在 P 和 T 之间建立匹配关系，通过

$$P^*(x|\theta_j)=P(x|y_j),\ j=1,2,\ldots,n, \tag{11}$$

其中 $P^*(x|\theta_j)$ 是优化的似然函数。然后我们有优化的真值函数:

$$\begin{aligned} T^*(\theta_j|x) &= [P^*(x|\theta_j)/P(x)]/\max(P^*(x|\theta_j)/P(x)) \\ &= [P(x|y_j)/P(x)]/\max(P(x|y_j)/P(x)),\ j=1,2,\ldots,n. \end{aligned} \tag{12}$$

使用贝叶斯定理 2, 可以进一步得到

$$T^*(\theta_j|x)= P(y_j|x)/\max(P(y_j|x)),\ j=1,2,\ldots,n. \tag{13}$$

这意味着语义信道匹配香农信道。一个转移概率函数 $P(y_j|x)$表示标签 y_j 的使用规则。因此，上面公式(12)和(13)反映维特根斯坦(Wittgenstein)的思想: 意义在于用法[35] (p.80).

在我心中 “P-T”中的“-“ 意味着贝叶斯定理 3 和上面两个公式, 它们起到统计概率和逻辑概率之间桥梁的作用，因而可以作为统计和逻辑之间的桥梁。

现在我们可以解决例 2 中的问题（见图 1）。根据等式（12），我们得到两个优化的真值函数

$$T^*(\text{“adult”}|x)=[P(x|\text{”adult”})/P(x)]/\max(P(x|\text{”adult”})/P(x)),$$

$$T^*(\text{“elder”}|x)=[P(x|\text{”elder”})/P(x)]/\max(P(x|\text{”elder”})/P(x)),$$

不管两个标签的外延是否是模糊的。得到 $T^*(\theta_j|x)$后，在 $P(x)$变化以后，我们仍然可以用贝叶斯定理 3 做概率预测。

在 Dempster-Shafer 理论中, 概率(mass）≤可信度(belief)≤或然度（plausibility）[36]. 上面 $P(y_j)$就是概率, $T(\theta_j)$就像是可信度。设 $\mathscr{Y}$ 有子集 $V_j = \{y_{j1}, y_{j2},\ldots\}$, 其中每个标签都不和 y_j 矛盾。我们定义 $PL(V_j) = \sum_k P(y_{jk})$——它就像或然度. 我们也有 $P(y_j) \le T(\theta_j) \le PL(V_j)$ 和 $P(y_j|x) \le T(\theta_j|x) \le PL(V_j|x)$.

下面我们证明等式（13）和汪培庄教授提出的隶属函数的随机集统计方法兼容（参看图 3）。

设在样本 **D** 中，$P(x)$是等概率的，这意味着对于每个 x 存在 N/m 个样例。这样，曲线 $P(y_j|x)$下的面积就是 $\mathbf{D}_j$ 中样例的个数。再设(x_j^*, y_j)是所有带有标签 y_j 的样例(x, y_j)中个数最多的样例，其个数是 N_j^*。那么，我们就可以划分所有带有 y_j 的样例为 N_j^* 行，每一行可以被当做一个集合值 S_k。容易证明[37]，通过等式（13）得到的真值函数就和通过随机集统计得到的真值函数相同。

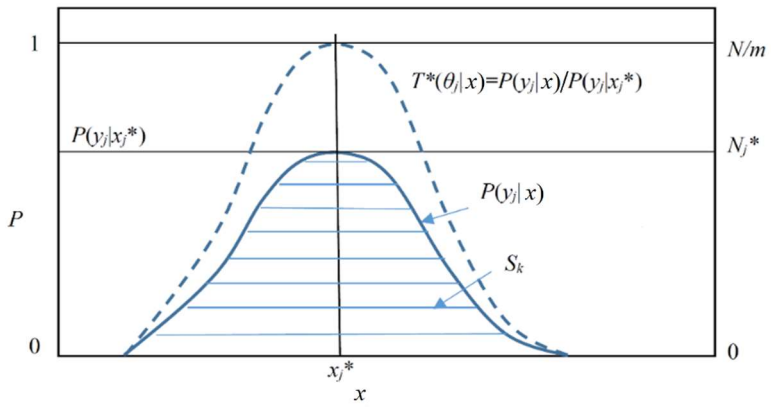


图 3. 用等式（13）得到的优化的真值函数（虚线）和用随机集统计得到的隶属函数相同[37]。S_k 是一个集合值（细线）; N_j 是集合值的数目。

如此得到的隶属函数符合赖欣巴哈关于逻辑概率的频率解释。不过，我们需要区分来自两种不同统计方法的两种不同概率。波普尔知道两种概率不同，并且认为两种概率之间应该有某种联系[7] (pp.252-258). 等式 (13)应该就是波普尔想要的。

然而，等式（12）和（13）需要 $\mathbf{D}_j$ 足够大，否则 $P(y_j|x)$不平滑，因而 $P(x|\theta_j)$没有意义. 这种情况下，我们需要用最大似然准则或最大语义信息准则优化真值函数（见 3.2 节）。

2.5 GPS指针和颜色感觉的逻辑概率和真值函数

不仅自然语言传递语义信息，时钟、秤、温度表、GPS 指针、股市指数……也传递语义信息，如 Floridi 指出[38]。我们可以用下面高斯真值函数（没有归一化系数的高斯函数）作为 y_j= “x 大约是 x_j”的真值函数：

$$T(\theta_j|x)=\exp[-|x-x_j|^2/(2\sigma^2)], \tag{14}$$

其中 x_j 是读数，x 是实际值，σ 是标准偏差。对于 GPS 设备，x_j 是 y_j所指的位置（一个矢量），x 是实际位置,σ 是误差平方的开方（RMS，即 Root Mean Square), 它表示 GPS 设备的准确性。

例 3. GPS 设备在列车上，$P(x)$均匀分布在一条线（轨道）上（见图 4），GPS 指针有偏差。请根据 y_j 找出 GPS 设备最大可能位置.

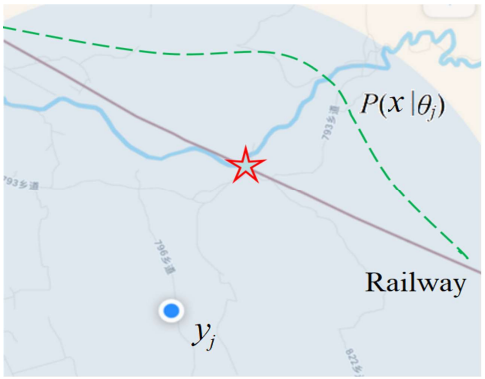


图 4. 图解 GPS 有偏差的定位。圆点是所指位置，五角星是人脑判断的最为可能位置，虚线是预测的概率分布。

使用统计概率方法，有人或许会认为 GPS 指针和圆圈告诉我们一个似然函数 $P(x|\theta_j)$或一个转移概率函数 $P(y_j|x)$. 然而，它所提供的圆圈并不表示似然函数，因为最为可能的位置不是 y_j 所指位置 x_j ；它也不是转移概率函数 $P(y_j|x)$，因为我们不知道它的最大值。合理的结论只能是：GPS 指针（和圆圈）提供了一个真值函数。

使用该真值函数, 我们可以用等式（8）做语义贝叶斯预测产生 $P(x|\theta_j)$, 据此，五角星就是最为可能的位置。大多数人没用任何数学公式也能够做同样的预测，说明人脑自动使用了类似方法——根据 y_j 的模糊外延和先验知识 $P(x)$做预测.

一种颜色感觉也可以当做一个仪器的读数，它传递语义信息[15]。在这种情况下，颜色感觉 y_j 的真值函数也就是 x 和 x_j 之间的混淆概率函数，逻辑概率 $T(\theta_j)$就是人眼把其他颜色和 x_j 相混淆的混淆概率。

我们也能用混淆概率解释一个假设 y_j 的逻辑概率。设相应每个模糊集合 θ_j 存在一个柏拉图的理念 x_j . 那么 θ_j 的隶属函数或 y_j 的真值函数也就是 x 和 x_j 之间的混淆概率函数[14].

就像 GPS 指针的语义不确定性一样，微观物理量的语义不确定性也能用真值函数而不是概率函数表示。

3 P-T 概率框架用于语义通信、统计学习和约束控制

3.1 从香农信息测度到语义信息测度

香农互信息被定义为

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}, \tag{15}$$

其中对数的底是 2. 如果 $Y=y_j$, 则 $I(X; Y)$变为 Kullback-Leibler (KL) 离散度：

$$I(X;y_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}. \tag{16}$$

如果 $X=x_i$, 则 $I(X; y_j)$ 变为

$$I(x_i;y_j)=\log\frac{P(x_i\mid y_j)}{P(x_i)}=\log\frac{P(y_j\mid x_i)}{P(y_j)}. \tag{17}$$

用似然函数 $P(x|\theta_j)$代替后验分布 $P(x|y_j)$, 我们有 y_j 关于 x_i的语义信息量：

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{18}$$

上面公式利用了贝叶斯定理 3. 其哲学意义将在第 4 节解释。如果 $y_j(x_i)$的真值总是 1，那么上面公式就变成卡尔纳普和 Bar-Hillel 的语义信息公式 [33]。

上面语义信息公式不仅可以用来度量自然语言的语义信息，也可以用来度量 GPS 指针、温度表或颜色感觉的语义信息。在后面情况下，真值函数意味着相似性函数或混淆概率函数。

对不同的 x_i 求 $I(x_i; \theta_j)$的平均，我们有平均语义信息（量）：

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}, \tag{19}$$

其中 $P(x_i|y_j)$ $(i=1,2,\ldots)$ 是样本分布。该公式可以用来优化真值函数.

对不同的 y_j 求 $I(X;\theta_j)$的平均, 我们得到语义互信息：

$$\begin{aligned} I(X;\Theta) &= \sum_j P(y_j)\sum_i P(x_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)} \\ &= \sum_i\sum_j P(x_i)P(y_j \mid x_i)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \end{aligned} \tag{20}$$

$I(X;\theta_j)$和 $I(X;\Theta)$ 都能用作分类准则。我们也称 $I(X;\theta_j)$为广义 Kullback-Leibler(KL)信息，称 $I(X;\Theta)$ 为广义互信息.

如果真值函数是高斯真值函数，$I(X;\Theta)$就等于广义熵减去平均相对平方误差:

$$I(X;\Theta) = -\sum_j P(y_j)\log T(\theta_j) - \sum_i\sum_j P(x_i,y_j)(x_i-y_j)^2/(2\sigma^2). \tag{21}$$

因此，最大语义互信息准则类似于正则化最小平方误差（RLS，即 the Regularized Least Squared Error）准则，这一准则正在机器学习中流行.

香农在其著名文章[3]中提出保真度评价函数——用于数据压缩。一个具体的保真度评价函数是信息率失真函数 $R(D)$ [39], 其中 D 是平均失真的上限；设用 y_j 表示 x_i 的失真量是 $d(x_i, y_j)$，其平均就是平均失真；R 是给定 D 时最小香农互信息。我们在第 3.3 节再联系随机事件的控制进一步介绍 $R(D)$函数。

用 $I(x_i;\theta_j)$代替 $d(x_i, y_j)$, 我发展了另一种保真度评价函数 $R(G)$ [12,15], 其中 G 是语义互信息 $I(X;\Theta)$的下限, $R(G)$是给定 G 时的最小香农互信息。$G/R(G)$表示通信效率，其最大值是 1，这时候 $P(x|\theta_j)=P(x|y_j)$（对所以 j），$R(G)$被称之为信息率逼真函数 (4.3 节进一步讨论). 它可以用于数据压缩——根据视觉分辨率[15]，也可以用于混合模型和最大互信息分类的收敛证明[12].

3.2 优化自然语言的真值函数和据此分类

统计学习的一个重要任务是优化预测模型——比如似然函数和 Logistic 函数——和分类. 因为用真值函数和先验分布能产生似然函数，真值函数也能作为预测模型，并且它有优点：在 $P(x)$改变时仍然可用。

现在我们再考虑例 2。在样本 $\mathbf{D}_j$ 不够大时，$P(x|y_j)$是不平滑的, 我们不能使用等式(12)或(13)获得一个平滑的真值函数。这时，我们可以用广义 KL 公式得到连续的真值函数。比如, 我们有优化的真值函数

$$T^*(\theta_{\text{elder}} \mid x) = \arg\max_{\theta_{\text{elder}}} \sum_i P(x_i \mid \text{"elder"}) \log\frac{T(\theta_{\text{elder}} \mid x_i)}{T(\theta_{\text{elder}})}. \tag{22}$$

其中 “arg max Σ” 意思是通过选择参数 θ_{elder} 最大化总和。在这种情况下,我们可以假设 $T(\theta_{\text{elder}})$是一个 Logistic 函数: $T(\theta_{\text{elder}})=1/[1+\exp(-u(x-v))]$, 其中 u 和 v 是要被优化的两个参数. 如果我们知道 $P(y_j|x)$而不知道 $P(x)$, 我们也可以假设 $P(x)$是常数从而获得样本分布 $P(x|y_j)$ 和逻辑概率 $T(\theta_{\text{elder}})$ [12].

逻辑贝叶斯推理[12]是

1) 从样本分布 $P(x|y_j)$和 $P(x)$ 获得优化的真值函数 $T^*(\theta_j|x)$；

2) 从 $T^*(\theta_j|x)$和 $P(x)$产生语义贝叶斯预测 $P(x|\theta_j)$.

逻辑贝叶斯推理[12]不同于贝叶斯推理[5]. 前者使用先验 $P(x)$, 而后者使用先验 $P(\theta)$.

在统计学习中，获得平滑的带参数的真值函数或隶属函数叫做标签学习；而从哲学角度看，这叫标签外延的归纳。

在流行的统计学习方法中，两个互补标签（比如“老年人”和“非老年人”）的学习是容易的，因为我们可以使用一对 Logistic 函数作为带参数的转移概率函数 $P(\theta_1|x)$和 $P(\theta_0|x)$ 或两个带参数的真值函数 $T(\theta_1|x)$和 $T(\theta_0|x)$。然而，多标签学习是困难的 [40]，因为我们不可能设计 n 个带参数的转移概率函数. 但是，使用 P-T 概率框架，多标签学习也是容易的，因为每个标签的学习是独立的[12].

使用优化的真值函数，我们可以把不同实例分类到不同类别——使用分类器 $y_j = f(x)$. 比如，我们可以把不同年龄的人贴上标签 “小孩”, “年轻人”, “成年人”, “中年人”和 “老年人”。使用最大语义信息准则， 分类器是

$$y_j^* = f(x) = \arg\max_{y_j} \log I(x;\theta_j) = \arg\max_{y_j} \log[T(\theta_j \mid x)/T(\theta_j)]. \tag{23}$$

该分类器的划分边界随人口年龄分布 $P(x)$改变而改变。随着人口寿命的延长，真值函数 $T^*(\theta_{\text{elder}}|x)$中的 v 和“老年人”的划分边界都会增大[12]。

最大语义信息准则兼容最大似然准则而不同于最大正确率准则，它鼓励减少小概率事件的漏报。比如，如果根据最大正确率准则我们会把超过 60 岁的人分类到老年人类别， 那么根据最大语义信息准则，我们会把超过 58 岁的人分类到老年人类别，因为老年人比非老年人或中年人少。要预测侧地震，如果使用最大正确率准则，我们会永远预测“下周没有地震”。这样的预测是没有意义的。如果使用最大语义信息准则，我们应在某些情况下预测“下周可能地震”，即使这样预测错误率较高。

P-T 概率框架和 G 理论也能用来改进最大互信息分类和混合模型（主要用于聚类）[12].

3.3 真值函数作为分布约束函数用于随机事件的控制

一个真值函数也是一个隶属函数，它可以被当做一个约束函数——约束一个随机事件的概率分布或一个随机粒子的密度分布。我们简单称之为分布约束函数。

KL 离散度和香农互信息也能用来度量控制随机事件的控制量。设 X 是一个随机事件， $P(x)=P(X=x)$ 是先验概率分布，X 可以是一个气体分子的能量，一个加工零件的尺寸，一个国家中的某一个人的年龄，等等。$P(x)$也可以是一个密度函数.

如果 $P(x|y_j)$是一个控制作用 y_j 发生后的后验密度分布，那么 KL 离散度 $I(X; y_j)$ 就是控制量（以比特为单位）,它反映控制复杂性。如果理想的后验分布是 $P(x|\theta_j)$, 那么有效控制量就是

$$I_c(X;\theta_j)=\sum_i P(x_i|\theta_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}. \tag{24}$$

注意：$P(x_i|\theta_j)$在 “log” 左边而不是右边. 对于广义 KL 信息 $I(X;\ \theta_j)$, 在预测 $P(x|\theta_j)$接近事实 $P(x|y_j)$时，信息量 $I(X;\ \theta_j)$接近最大；而对于有效控制量 $I_c(X;\ \theta_j)$, 在事实接近理想 $P(x|\theta_j)$时，有效控制量, $I_c(X;\ \theta_j)$ 接近最大。一个不适当的后验分布 $P(x|y_j)$可能使有效控制量 $I_c(X;\ \theta_j)$小于 0。

一个真值函数用作分布约束函数意味着：

$$1\text{-}P(x|y_j)\le 1\text{-}P(x|\theta_j)=1\text{-}P(x)T(\theta_j|x)/T(\theta_j),\ \text{for}\ T(\theta_j|x)<1. \tag{25}$$

如果 θ_j 是清晰集合，上述条件意味着 x 不能在 θ_j 之外。如果 θ_j 是模糊集合, 它意味着 x 在 θ_j 之外的比例应被限制。有很多分布 $P(x|y_j)$ 满足上面条件, 但是只有一个需要最小 KL 信息 $I(X;\ y_j)$. 比如，设 x_j 使 $T(\theta_j|x_j)=1$；如果 $x=x_j$ 时 $P(x|y_j)=1$，且 $x\ne x_j$ 时 $P(x|y_j)=0$。那么 $P(x|y_j)$ 满足上面条件，但是这一 $P(x|y_j)$需要的信息 $I(X;\ y_j)$或控制量并不是最小的.

我提出信息率容差函数 $R(\Theta)$ [41]——它是复杂性失真函数 $R(C)$ [42] 的推广。和信息率失真函数不同，复杂性失真函数的约束条件是每个失真 $d(x_i,\ y_j)=(x_i\text{-}y_j)^2$ 小于给定值 C, 这意味着约束集合大小是相同的。和 $R(C)$的约束条件不同, $R(\Theta)$的一组约束集合是模糊的，并且大小不同。我已经得到结论 [14,15]:

- 对于给定的一组约束分布 $T(\theta_j|x)$ (j=1, 2, …, n) 和 $P(x)$, 在 $P(x|y_j)=P(x|\theta_j)=P(x)T(\theta_j|x)/T(\theta_j)$时, KL 信息 $I(X;\ y_j)$和 $I(X;\ Y)$ 达到最小值，有效控制量 $I_c(X;\ y_j)$ 和 $I_c(X;\ Y)$达到最大值。如果每个集合 θ_j 是清晰的, 则 $I(X;\ y_j)=\text{-}\log T(\theta_j)$，$I(X;\ Y)=\text{-}\sum_j P(y_j)\log T(\theta_j)$.
- 一个信息论失真函数 $R(D)$ 等价于一个信息率容差函数 $R(\Theta)$, 我们能用含有真值函数或分布约束函数的语义互信息公式表达它们(详见附录 II)。 但是一个 $R(\Theta)$函数可能并不等价于一个 $R(D)$函数，因此 $R(D)$是 $R(\Theta)$的特例.

设 $d_{ij}=d(x_i,\ y_j)$ 是我们用 y_j 表示 x_i时的失真或损失，D 平均失真上限。对于给定 $P(x)$, 我们能获得香农互信息最小值即 $R(D)$。$R(D)$函数的参数形式 [43] (P. 32)包括两个公式:

$$\begin{aligned} D(s)&=\sum_i\sum_j d_{ij}P(x_i)P(y_j)\exp(sd_{ij})/\lambda_i,\\ R(s)&=sD(s)-\sum_i P(x_i)\ln\lambda_i,\ \lambda_i=\sum_j P(y_j)\exp(sd_{ij}), \end{aligned} \tag{26}$$

其中 $s=\mathrm{d}R/\mathrm{d}D\le 0$ 反映 $R(D)$的斜率。 $P(y|x_i)$的后延分布是

$$P(y\mid x_i)=P(y)\exp(sd_{ij})/\lambda_i, \tag{27}$$

它使 $I(X;\ Y)$ 达到最小值.

因为 $s \le 0$, $\exp(sd_{ij})$的最大值是 1（当 s=0 时）。 一个常用的失真函数是 $d(x_i, y_j)=(y_j-x_i)^2$. 对于这一失真函数， $\exp(sd_{ij})$是一个没有系数的高斯函数。因而，$\exp(sd_{ij})$能被当做真值函数或分布约束函数 $T(\theta_{xi}|y)$——其中 θ_{xi} 是 $\boldsymbol{y}$ （而不是 $\boldsymbol{x}$）上的模糊子集; λ_i 能被当做 x_i 的逻辑概率 $T(\theta_{xi})$ 。 现在我们能看出等式 (27)实际上是一个语义贝叶斯公式（在贝叶斯定理 3 中)；一个 $R(D)$函数可以被表达为带有真值函数的语义互信息公式，它等价于一个 $R(\Theta)$函数 (详见附录 II).

在统计力学中，类似于上述分布 $P(y|x_i)$的是玻尔茨曼分布 [44]

$$P(x_i \mid T) = \exp(-\frac{e_i}{kT}) / Z, \;\; Z = \sum_i \exp(-\frac{e_i}{kT}), \tag{28}$$

其中 $P(x_i|T)$ 一个粒子在第 i 种状态的概率或密度，该粒子的能量是 e_i; T 绝对温度，k 是玻尔茨曼常数, Z 是划分函数。

如果用 e_i 表示第 i 种能量, G_i 是带有 e_i 的状态个数（简并度），G 是所有状态的个数，那么 $P(x_i)=G_i/G$ 就是先验分布。于是, 上面公式变成

$$P(x_i \mid T) = P(x_i)\exp(-\frac{e_i}{kT}) / Z', \;\; Z' = \sum_i P(x_i)\exp(-\frac{e_i}{kT}). \tag{29}$$

现在我们能看出 $\exp[-e_i/(kT)]$ 可以看作真值函数或分布约束函数，Z' 可以看作逻辑概率，等式（29）可以看作是贝叶斯定理 3 中的语义贝叶斯公式.

对于局域平衡系统， 系统的不同区域有不同温度。借助于真值函数或分布约束函数，我[14](pp. 102-103)推导出最小香农互信息 $R(\Theta)$热力学熵之 S 间的关系 （详见附录 III)：

$$R(\Theta) = \ln G - S/(kT). \tag{30}$$

该公式表明，最大熵原理等价于最小互信息原理。

4 P-T 概率框架和 G 理论如何支持波普尔思想

4.1 语义信息测度如何支持波普尔的科学进步思想

早在 1935，波普尔在《科学发现的逻辑》 [7]中就提出：一个假设逻辑概率越小，就越容易被证伪，因而含有更多的经验信息。他说：：

“一个理论的经验信息量或其经验内容随着其证伪度增加而增加。” (p. 96)
“一个叙述的逻辑概率和其证伪度是互补的，它增加的时候证伪度减小。 “(p. 102)

但是，波普尔没有提出语义信息公式。1948 年，香农提出著名的信息理论——使用统计概率[3]. 1950 年, 卡尔纳普 和 Bar-Hillel 提出语义信息测度——使用逻辑概率：

$$I_p=\log(1/m_p), \tag{31}$$

其中 p 是一个命题， m_p 是逻辑概率。 然而，这个公式和实例（它使命题为真或为假）不相关，因此，它能反映检验的严厉性，不能反映 p 是否能很好地经得起检验。

1963 年，波普尔出版了《猜想和反驳》[45]。在这本书中，他更明确肯定科学理论的意义是传递信息。他说：

“凡是包含更大量的经验信息或内容的理论，也即在逻辑上更有力的理论，凡是具有更大的解释力和预测力的理论，从而可以通过把所预测的事实同观察加以比较而经得起更严格检验的理论，就更为可取。总之，我们宁取一种有趣、大胆、信息丰富的理论，而不取一种平庸的理论。” [45] (p.294).

在这本书中，波普尔 还提出用 [45] (p.526)

$$P(e, hb)/P(e, b) \text{ or } \log[P(e, hb)/P(e, b)]$$

表示假设如何经得起严厉检验，其中 e 是支持理论 h 的证据，b 是背景知识，$P(e, hb)$ 和 $P(e, b)$ 在 [45] 中是条件概率——其现代形式是 $P(e|h,b)$和 $P(e|b)$。和卡尔纳普和 Bar-Hillel 的语义信息公式不同，上面公式能反映 e 支持 h 有多好。但是, $P(e, hb)$ 是 e 的条件概率，不易被解释为逻辑概率；因而，我不能说 $\log[P(e, hb)/P(e, b)]$表示语义信息量.

后来其他人又提出不少语义信息或广义信息测度[38,46,47]. 相比之下，我的语义信息测度更加支持波普尔的科学评价理论。我们用一个例子说明其性质。

设 y_j 是 GPS 指针或假设 “x 大概是 x_j”。其真值函数是 $\exp[-(x-x_j)^2/(2\sigma^2)]$ (见图 5)。那么，我们有语义信息(量):

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}=\log[1/T(\theta_j)]-(x_i-x_j)^2/(2\sigma^2). \quad (32)$$

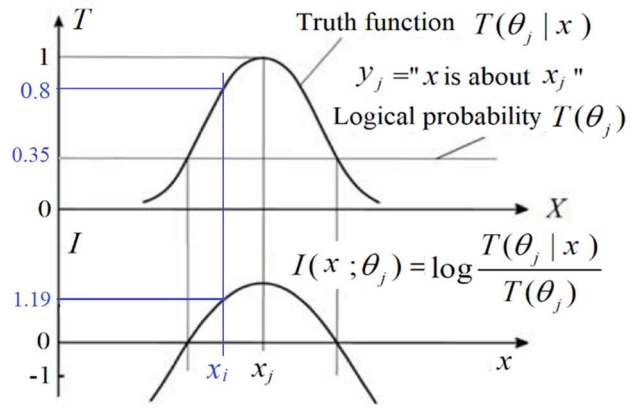


Figure 5. y_j 关于 x_i 的语义信息量可以代表 x_j 反映 x_i 的逼真度。 当实际的 x 是 x_i, 真值是 $T(\theta_j|x_i)$=0.8; 信息是 $I(x_i; \theta_j)$= log(0.8/0.35)=1.19 比特. 如果 x 超出一定范围, 信息是负的.

图 5 显示真值函数和语义信息随 x 变化。$P(x_i)$小意味着 x_i 是意外的; $P(x_i|\theta_j)$大意味着预测是正确的; $\log[P(x_i|\theta_j)/P(x_i)]$表明 y_j 经受 x_i 的检验有多严厉有多好。$T(\theta_j|x_i)$大意味着 y_j 是真的或接近真; $T(\theta_j)$小意味着 y_j 是精确的。 因此, $\log[T(\theta_j|x_i)/T(\theta_j)]$表明 x_j 反映 x_i 的逼真度。意外性、正确性、可检验性、真、精确、逼真、偏差都调和在一个公式中了。

图 5 表明：逻辑概率越小，信息量越大；偏差越大，信息量越小；一个错的假设提供的信息是负的。根据这一信息测度，永真句和矛盾句的信息量为 0。这些结论符合波普尔思想。

4.2 语义信息测度如何支持波普尔的证伪思想

波普尔肯定：逻辑概率较小的假设容易被证伪，如果它经得起检验，它就提供更多信息。计算 $I(x;\theta_j)$的语义信息公式反映了这一思想。波普尔肯定：一个反例可以证伪一个全称假设。广义 KL 信息公式 (19)支持这一观点。一个全称假设的真值函数取值只有 0 或 1。如果存在一个实例 x_i 使得 $T(\theta_j|x_i)=0$, 那么 $I(x_i;\theta_j)$ 就是-∞. 平均信息 $I(X;\theta_j)$也是-∞, 它证伪全称假设.

拉卡托斯（Lakatos）部分接受了库恩（Kuhn）关于证伪的思想。他指出 [48]：在科学实践中，当观察事实不符合一个理论（或假设）时，科学家并不轻易放弃该理论。他们经常增加一些辅助假设到该理论或其内核，使得该理论和辅助假设一起符合更多观察事实。拉卡托斯因此称他改进的证伪是精致证伪。

拉卡托斯是正确的，比如，根据科学的内核，一个 GPS 设备到三个卫星的距离确定了它的位置。然而，要提供准确的定位，厂家还要考虑各种条件因素，包括卫星几何、信号障碍、大气条件等等。但是波普尔也未必就是错的，因为我们能用下面理由解释：

- 波普尔主张科学知识通过重复猜想和反驳得到增长。重复猜想应该包括辅助假设。
- 证伪并不是目的，它只是科学和非科学理论的划界标准。科学的目的是预测经验事实从而提供更多信息。科学家保留一个理论取决于它是否比其他理论提供更多信息，因此，证伪假设不等于放弃假设。

但是，这里仍然存在问题，那就是：一个被证伪的理论怎样能提供比别的理论更多信息？根据平均语义信息公式或广义 KL 信息公式，我们可以通过下面方法提供平均语义信息：

- 增加预测的模糊性或降低预测精度到一个适当的水平；
- 减少我们对一个预测规则或大前提的可信度。

我们可以把一个全称假设变为一个不太确定的全称假设，比如，我们可以把“所有天鹅是白的” 改成 “差不多所有天鹅是白的,” “可能所有天鹅是白的,” 或 “天鹅是白的，确证度是 0.9” 。

GPS 设备是一个定量的例子。GPS 设备用 RMS 或类似指标表示准确性(正确性和精确性)。GPS 地图上围绕 GPS 指针的较大圆圈(见图 4)意味着较大的 RMS 即较低的准确性。如果一个 GPS 设备显示有较多的实际位置超出了圆圈，但是不远，我们就可能并不放弃该 GPS 设备。我们可以假设它有较大的 RMS 并且继续使用它。我们选择最好的假设类似于我们选择更高准确性的 GPS 设备。 如果一个 GPS 设备可能指向错误方向，我们可以通过降低可信度来减少平均信息损失[12]。

4.3 逼真度（verisimilitude）：调和内容方法和相似性方法

在波普尔早期著作中，他强调假设具有小逻辑概率的重要性，没有强调假设的真。但是在波普尔后来的著作《实在论和科学的目的》[49]中，他解释科学的目的在于更好的理论解释——符合事实因而是真的, 即使我们并不知道或不能证明理论是真的. 拉卡托斯 [50] 因此指出：波普尔的科学游戏和科学目的是矛盾的。科学游戏包含大胆猜想和严厉反驳，而猜想不必是真的；然而，科学的目的包含发展真的或似真的理论——关于独立于心的世界的理论。拉卡托斯批评和波普尔的逼真度概念相关。

波普尔 [42] (p.535) 提供了一个逼真度公式:

$$Vs(a)=\begin{cases}0, & \text{如果 } a \text{ 是永真句},\\ -1, & \text{如果 } a \text{ 是矛盾句},\\ [1-P(a)]/[1+P(a)], & \text{否则。}\end{cases} \tag{33}$$

其中 a 是一个命题, 1-$P(a)$意味着信息内容。 这个公式意味着逻辑概率越小，信息内容越多，因此逼真度越高。波普尔希望 $Vs(a)$在-1 和 1 之间变化。但是根据这一公式, $Vs(a)$不会出现在-1 和 0 之间。更严重的问题是，他只利用了逻辑概率 $P(a)$, 没有使用命题 a 的真值, 或者说没有使用结果——它可能比别的结果更接近真[51], 以致于上面公式没有表达预测和结果之间的相似性。波普尔后来承认这一错误因而强调：我们需要符合真实世界的真的解释理论[49]。

现在研究者使用三种方法解释逼真度[52]: 内容方法、结果方法和相似性方法。因为后两个是相关的并且明显不同于内容方法，于是逼真度理论按惯例被分为两个对立阵营: 内容方法和相似性方法 [52]. 内容方法强调检验的严厉性，和相似性方法不同， 后者强调假设的真或接近真。有些研究者认为内容方法和相似性方法是不可调和的[53], 就像拉卡托斯认为波普尔的科学游戏和科学目的是矛盾的。但是也有研究者，比如 Oddie [52], 试图试图结合两种方法。不过他们也认为结合是不容易的。

本文继续 Oddie 等人的努力。使用语义信息测度作为逼真度或似真度（truthlikeness）, 我们能容易地结合内容方法和相似性方法。

在等式 (32)和图 5 中, 真值函数 $T(\theta_j|x)$也就是混淆概率函数，它反映 x 和 x_j 的相似性。 x_i (或 $X=x_i$)是结果, x_i 和 x_j 在特征空间中的距离反映了相似性; $\log[1/T(\theta_j)]$ 表示检验严厉性和潜在的信息内容。使用等式 (32), 我们能容易解释经常谈到的例子[52]：为什么“太阳有 9 个卫星”(实际有 8 个) 比 “太阳有 100 个卫星”有更高的逼真度.

另一个经常谈到的例子是怎样度量天气预报的逼真度[52]。使用语义信息方法，我们能评价更实用的天气预报。我们用矢量 $\mathbf{x} = (h, r, w)$ 表示一种天气，其中 h 是温度，r 是降雨量，w 是风速。 设 $\mathbf{x}_j = (h_j, r_j, w_j)$ 是预测的天气，$\mathbf{x}_i = (h_i, r_i, w_i)$是实际天气 (即结果)。预测是 y_j="$\mathbf{x}$ 大概是 $\mathbf{x}_j$." 为简单起见，我们假设 h, r, w 是相互独立的。我们可用下面高斯真值函数：

$$T(\theta_j|\mathbf{x}) = \exp[-(h-h_j)^2/(2\sigma_h^2)-(r-r_j)^2/(2\sigma_r^2)-(w-w_j)^2/(2\sigma_w^2)]. \quad (34)$$

这一真值函数符合相似性方法核心——相似性取决于两个实例($\mathbf{x}$ 和 $\mathbf{x}_j$)之间的距离[52]。如果结果是 $\mathbf{x}_i$, 那么命题 $y_j(\mathbf{x}_i)$ 的真值 $T(\theta_j|\mathbf{x}_i)$ 就是相似性，信息 $I(\mathbf{x}_i; \theta_j) = \log[T(\theta_j|\mathbf{x}_i)/T(\theta_j)]$就是逼真度——它差不多具有使用三种方法所希望的所有性质。

逼真度 $I(\mathbf{x}_i; \theta_j)$ 也和先验分布 $P(\mathbf{x})$相关. 不常见天气的正确预测会比普通预测有更高的逼真度——在两种预测都正确的情况下.

现在我们可以解释：正则化误差平方准则越来越流行是因为它类似于最大逼真度准则(参看等式 (21))。

5 P-T 概率框架和语义信息 G 理论用于确证

5.1 确证的目的：为不确定三段论优化大前提的可信度

确证经常被解释为评估证据对假设的冲击（增量确证）[11, 54]或证据对假设的支持(绝对确证) [55, 56]；确证度就是证据或数据支持的可信度 [57]。确证研究者只研究用确证测度评估假设。已经存在各种各样确证测度 [11]。

下面我们用 e 代替 y 表示证据；用 h 代替 x，表示要支持的假设；用 h_0 表示 h_1 的否定；用 e_0 表示 e_1 的否定；用 $c(e_1, h_1)$表示一个确证测度。在流行的方法中，逻辑概率和统计概率不加区分，都用 P 表示。流行的关于“如果 e_1 那么 h_1”确证测度包括 [11, 58]：

- $D(e_1, h_1)=P(h_1|e_1)-P(h_1)$ (卡尔纳普, 1962),
- $M(e_1, h_1) = P(e_1|h_1)-P(e_1)$ (Mortimer, 1988),
- $R(e_1, h_1) = \log[P(h_1|e_1)/P(h_1)]$ (Horwich, 1982),
- $C(e_1, h_1) = P(h_1, e_1)-P(e_1)P(h_1)$ (卡尔纳普,1962),
- $$Z(h_1,e_1) = \begin{cases} [P(h_1 \mid e_1) - P(h_1)] / P(h_0), \text{ as } P(h_1 \mid e_1) \geq P(h_1), \\ [P(h_1 \mid e_1) - P(h_1)] / P(h_1), \text{ otherwise}, \end{cases}$$

  (Shortliffe and Buchanan, 1975；Crupi et al., 2007),
- $S(e_1, h_1) = P(h_1|e_1)-P(h_1|e_0)$ (Christensen, 1999),
- $N(e_1, h_1) = P(e_1|h_1)-P(e_1|h_0)$ (Nozik, 1981),
- $L(e_1, h_1) = \log[P(e_1|h_1)/P(e_1|h_0)]$ (Good, 1984), and
- $F(e_1, h_1) = [P(e_1|h_1)-P(e_1|h_0)]/[P(e_1|h_1)+P(e_1|h_0)]$

(Kemeny and Oppenheim, 1952).

我也提出两种不同确证测度[58]。下面我联系科学推理简要介绍我的确证研究。

在我的确证研究中，任务、目的和方法和流行的理论中的任务、目的和方法有所不同。

- **确证的任务**:

只有大前提——比如 “如果医学检验是阳性，则被样品被感染” 和 “如果 x 是乌鸦，则 x 是黑的”——需要确证；确证度在-1 和 1 之间变化。一个命题，比如 “汤姆是年老的”, 或一个谓词比如 “x 是年老的” (x 是给定人群中的一个)不需要确证。谓词 “x 是年老的”的真值函数反映“年老的”语义，是由定义或习惯用法确定的。我们对一个命题的可信度是命题的真值，对一个谓词的可信度是谓词的逻辑概率。真值和逻辑概率在 0 和 1 之间而不是-1 和 1 之间。

- **确证的目的:**

确证的目的不只为评估假设（大前提），也为概率预测或不确定三段论。一个三段论需要一个大前提，但是如休谟和波普尔所说，要通过归纳获得绝对正确的全称大前提——其论域是无限的——是不可能的。但是，通过正反例的比例优化我们对这样的大前提的可信度是可能的。使用确证度，我们能做不确定模糊三段论。因此，确证是根据经验做科学推理的重要环节。

- **确证方法:**

我并不直接定义确证测度——像大多数研究者所做的那样。我推导出确证测度——通过优化对大前提的可信度，使用最大语义信息准则或最大似然准则。这种方法也就是统计学习方法，其中证据是样本。

从统计学习的视角看，一个大前提来自一个分类器或一个预测规则。下面我们使用医学检验作为例子解释分类和确证之间的关系（参看图 6）。二元信号检测和使用“老年人”和“非老年人”标签的人群分类是类似的。西瓜和垃圾邮件分类也是类似的。

在图 6 中，h_1 表示感染的样品（或个人）, h_0 表示未感染的样品, e_1 表示阳性，e_0 是阴性。我们可以把 e_1 当作预测 “h 被感染”，把 e_0 当作预测 “h 未被感染”。x 是 h 的观察特征；E_1 和 E_2 是 x 的论域中的两个子集。如果 x 在 E_1 中，我们选择 e_1; 如果 x 在 E_0 中，我们选择 e_0。对于二元信号检测，我们使用信宿中的“0” 或 “1” 预测信源中的 0 或 1 ——根据接收到的模拟信号 x。

需要被确证的两个大前提是 “如果 e= e_1 那么 h=h_1”（记为 $e_1 \rightarrow h_1$）和“如果 e=e_0 那么 h=h_0”(记为 $e_0 \rightarrow h_0$)。一个确证测度记为 $c(e \rightarrow h)$.

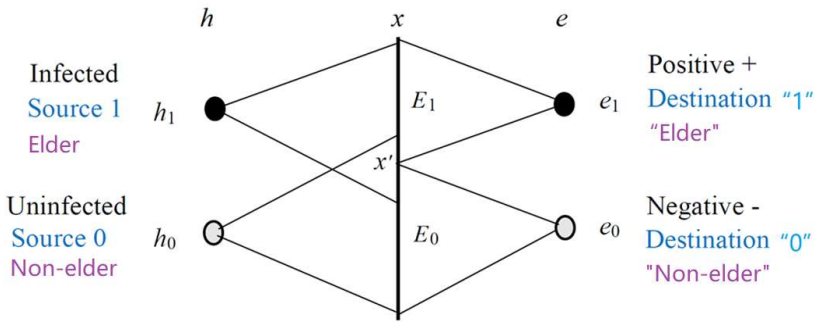


图 6. 图解医学检验和二元分类用以解释确证。如果 x 在 E_1 中，我们用 e_1 预测 "h=h_1"；如果 x 在 E_0 中，我们用 e_0 预测 "h=h_0".

x'确定了一个分类。对于给定的分类或预测规则，我们能获得一个样本，它包括四种样例：(e_1, h_1), (e_0, h_1), (e_1, h_0) 和 (e_0, h_0)。然后我们可以使用四种样例的数目 a, b, c, d (见表 1)构造确证测度。

表 1. 确证用的四种样例个数

	e_0	e_1
h_1	b	a
h_0	d	c

表中 a 是 (e_1, h_1)的个数；b, c, d 同理。绝对确证测度可以表示为函数 $f(a, b, c, d)$，其增量为

$$\Delta f = f(a+\Delta a, b+\Delta b, c+\Delta c, d+\Delta d) - f(a, b, c, d).$$

信息测度和确证测度用于不同任务。要比较两个假设并选择较好的一个，我们使用信息测度。使用最大语义信息准则，我们能找到最好划分点 x' ，它能产生最大互信息分类[12]。要优化对大前提的可信度，我们需要确证测度。使用信息测度是为了分类，而确证测度用在分类之后。

5.2 信道确证测度——评价分类作为信道

一个二元分类确定一个香农信道，它包含四个条件概率，如表 2 所示。

表 2. 二元分类产生的四个条件概率构成一个香农信道.

	e_0 (阴性)	e_1 (阳性)
h_1 (被感染)	$P(e_0\|h_1) = b/(a+b)$	$P(e_1\|h_1) = a/(a+b)$
h_0 (未感染)	$P(e_0\|h_0) = d/(c+d)$	$P(e_1\|h_0) = c/(c+d)$

我们把谓词 $e_1(h)$当做可信部分和不可信部分的组合 (见图 7)。e_1 的可信部分的真值函数是 $T(E_1|h) \in \{0,1\}$。$T(E_1|h_1) = T(E_0|h_0) = 1$；$T(E_1|h_0) = T(E_1|h_0) = 0$。不可信部分是永真句，其真值函数总是 1。然后我们有谓词 $e_1(h)$和 $e_0(h)$的真值函数:

$$T(\theta_{e1}|h) = b_1' + b_1 T(E_1|h); \tag{35}$$

$$T(\theta_{e0}|h) = b_0' + b_0 T(E_0|h). \tag{36}$$

模型参数 b_1 是可信部分的比例；b_1'=1-$|b_1|$ 是不可信部分的比例，也是 $y_1(h_0)$的真值 $T(E_1|h_0)$ (h_0是相反实例)。b_1' 可以当作大前提 $e_1 \rightarrow h_1$ 的不信度。

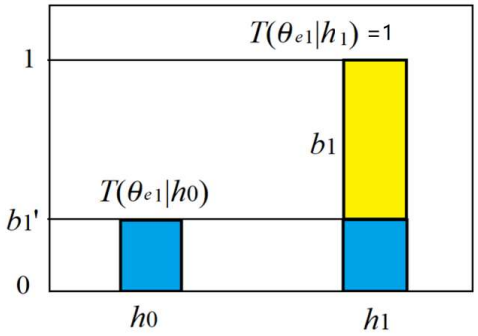


图 7. 真值函数 $T(\theta_{e1}|h)$ 包括可信部分(b_1)和不可信部分($b_1' = 1 - |b_1|$)。

四个真值形成一个语义信道, 如表 3 所示.

表 3. 两个不信度 b_1'和 b_0'确定一个语义信道.

	e_0 **(阴性)**	e_1 **(阳性)**
h_1 (被感染)	$T(\theta_{e0}\|h_1) = b_0'$	$T(\theta_{e1}\|h_1) = 1$
h_0 (未感染)	$T(\theta_{e0}\|h_0) = 1$	$T(\theta_{e1}\|h_0) = b_1'$

我们可以用 $T(\theta_{e1}|h)$做概率预测 $P(h|e_{\theta 1})=P(h)T(\theta_{e1}|h)/T(\theta_{e1})$。根据广义 KL 公式（19），当 $P(h|\theta_{e1})=P(h|e_1)$ 或 $T^*(\theta_{e1}|h) \propto P(e_1|h)$时，平均语义信息量 $I(H; \theta_{e1})$ 达到其最大值。令 $P(h_1|\theta_{e1})=P(h_1|e_1)$, 我们推导出（详见附录 IV）:

$$b_1^*=1-b_1'^*=[P(e_1|h_1)-P(e_1|h_0)]/P(e_1|h_1). \tag{37}$$

考虑 $P(h_1|e_1) < P(h_0|e_1)$, 我们有

$$b_1^*=b_1'^*-1=[P(e_1|h_0)-P(e_1|h_1)]/P(e_1|h_0). \tag{38}$$

结合上面两个等式，我们有

$$\begin{aligned} b_1^* = b^*(e_1 \rightarrow h_1) &= \frac{P(e_1|h_1)-P(e_1|h_0)}{\max(P(e_1|h_1), P(e_1|h_0))} \\ &= \frac{ad-bc}{\max(a(c+d), c(a+b))} = \frac{LR^+ - 1}{\max(LR^+, 1)}, \end{aligned} \tag{39}$$

其中 LR 是似然比, LR^+ 是阳性似然比。Elles 和 Fitelson [54]提出假设对称性: $c(e_1 \rightarrow h_1) = -c(e_1 \rightarrow h_0)$。 因为也有 $c(h_1 \rightarrow e_1) = -c(h_1 \rightarrow e_0)$ （其中两个结果即证据是相反的）, 我称这种对称性为结果对称性[58]。因为

$$b_1^* = b^*(e_1 \to h_0) = \frac{P(e_1 \mid h_0) - P(e_1 \mid h_1)}{\max(P(e_1 \mid h_0), P(e_1 \mid h_1))} = -b^*(e_1 \to h_1), \tag{40}$$

b_1* 具有这种对称性。

用类似方法，我们获得

$$b_0^* = b^*(e_0 \to h_0) = \frac{P(e_0 \mid h_0) - P(e_0 \mid h_1)}{\max(P(e_0 \mid h_0), P(e_0 \mid h_1))} = \frac{LR^- - 1}{\max(LR^-, 1)} \quad . \tag{41}$$

使用上面对称性，我们有 $b^*(e_1 \to h_0) = -b^*(e_1 \to h_1)$ 和 $b^*(e_0 \to h_1) = -b^*(e_0 \to h_0)$.

在医学检验中，似然比用来评估一种检验结果（阳性或阴性）有多好。 Kemeny 和 Oppenheim [55]提出的确证测度 F 是

$$F(e_1 \to h_1) = \frac{P(e_1 \mid h_1) - P(e_1 \mid h_0)}{P(e_1 \mid h_1) + P(e_1 \mid h_0))} = \frac{LR^+ - 1}{LR^+ + 1} \quad . \tag{42}$$

测度 b^*和测度 F 类似，也是似然比的函数, 因此 F 也能用来评估医学检验。和 LR 相比， b^* (或 F) 能更好表示一种检验（具有 b^*）和最好检验 (b^*=1)或最坏检验 (b^*= -1) 的距离。

和 F 相比， b^* 能更好地用于概率预测。比如, 给定 $b_1^* > 0$ 和 $P(h)$, 我们有

$$P(h_1|\theta_{e1}) = P(h_1)/[\, P(h_1) + b_1' * P(h_0)] = P(h_1)/[1 - b_1 * P(h_0)]. \tag{43}$$

这一公式比经典的贝叶斯公式（6）更加简单。如果 $b_1^* = 0$, 则 $P(h_1|\theta_{e1}) = P(h_1)$; 如果 $P(h_1|\theta_{e1}) < 0$, 我们可以利用结果对称性做概率预测[58]。

然而，到目前为止，要评价一个概率预测有多好或处理乌鸦悖论，使用 b^*, F 或其他确证测度仍然有问题。

5.3 预测确证测度 c^*用于澄清乌鸦悖论

统计学不仅使用似然比表示一个检验手段（作为信道）有多好，还使用正确率表示所预测的事件有多大可能发生。测度 F 和 b^* 同似然比 LR 一样，不能反映概率预测的质量。比如 $b_1^* > 0$ 并不意味 $P(h_1|\theta_{e1}) > P(h_0|\theta_{e1})$。其他确证测度都有类似问题[58].

现在我们把概率预测 $P(h|\theta_{e1})$看作可信部分（比例是 c_1）和不可信部分(比例是 c_1')的组合，如图 8 所示。我们称 c_1 是规则 $e_1 \to h_1$ 作为预测的可信度。

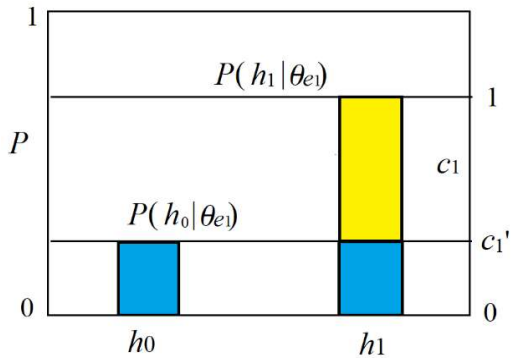


图 8. 似然函数 $P(h|\theta_{e1})$可看作是可信部分加不可信部分。

在预测符合事实时，即 $P(h|\theta_{e1})= P(h|e_1)$时，可信度 c_1 变成确证度 c_1*。然后我们推导出预测确证测度

$$\begin{aligned} c_1^* &= c^*(e_1 \to h_1) = \frac{P(h_1 \mid e_1) - P(h_0 \mid e_1)}{\max(P(h_1 \mid e_1), P(h_0 \mid e_1))} \\ &= \frac{2P(h_1 \mid e_1) - 1}{\max(P(h_1 \mid e_1), 1 - P(h_1 \mid e_1))} = \frac{2CR_1 - 1}{\max(CR_1, 1 - CR_1)}. \end{aligned} \tag{44}$$

其中 $CR_1 = P(h_1|\theta_{e1}) = P(h_1|e_1)$ 是规则 $e_1 \to h_1$ 的正确率。等式（44）的分子和分母同乘以 $P(e_1)$, 我们得到

$$c_1^* = c^*(e_1 \to h_1) = \frac{P(h_1, e_1) - P(h_0, e_1)}{\max(P(h_1, e_1), P(h_0, e_1))} = \frac{a - c}{\max(a, c)}. \tag{45}$$

用类似的方式，我们得到

$$c_0^* = c^*(e_0 \to h_0) = \frac{P(h_0, e_0) - P(h_1, e_0)}{\max(P(h_0, e_0), P(h_1, e_0))} = \frac{d - b}{\max(d, b)}. \tag{46}$$

我们同样能证明：$c^*(e_1 \to h_1)$ 也具有结果对称性。利用这一对称性，我们有 $c^*(e_1 \to h_0) = -c^*(e_1 \to h_1)$ 和 $c^*(e_0 \to h_1) = -c^*(e_0 \to h_0)$.

当 $c_1^* > 0$ 时，根据等式（45），我们有规则 $e_1 \to h_1$ 的正确率：

$$CR_1 = P(h_1 \mid \theta_{e_1}) = 1 / (1 + c_1'^*) = 1 / (2 - c_1^*). \tag{47}$$

在 $c^*(e_1 \to h_1) < 0$ 时，我们可以利用结果对称性做概率预测. 但是，在 $P(h)$改变时，我们仍然需要使用 b^* 和 $P(h)$ 做概率预测。

对于医学检验，我们需要两种确证测度。测度 b^*告诉我们一种检验和其他检验比，作为手段或信道的可靠性；而测度 c^*告诉我们一个人被感染的可能性。 对于科学预测，b^*和 c^*有类似意义。

现在我们能用测度 c^*澄清乌鸦悖论。

亨普尔(Hemple) [59] 提出确证悖论，即乌鸦悖论。根据经典逻辑中的等价条件，“如果 x 是乌鸦，则 x 是黑的” (规则 I)等价于 “如果 x 不是黑的, 则 x 不是乌鸦” (规则 II)。一支白粉笔支持规则 II; 因而也支持规则 I。但是，根据 Nicod 准则 [60], 一只黑乌鸦支持规则 I, 一个非黑的乌鸦反对规则 I；一个不是乌鸦的东西——比如一只黑猫或一只白粉笔——和规则 I 不相关。因此，在等价条件和 Nicod 准则之间存在悖论。

为了澄清乌鸦悖论，有些研究者包括亨普尔 [59], 肯定等价条件而否定 Nicod 准则；有些研究者——比如 Scheffler 和 Goodman [61]——肯定 Nicod 准则而否定等价条件。有些研究者并不全部否定或肯定等价条件或 Nicod 准则.

图 9 显示一个样本——用于大前提 “如果 x 是乌鸦，则 x 是黑的”的确证。

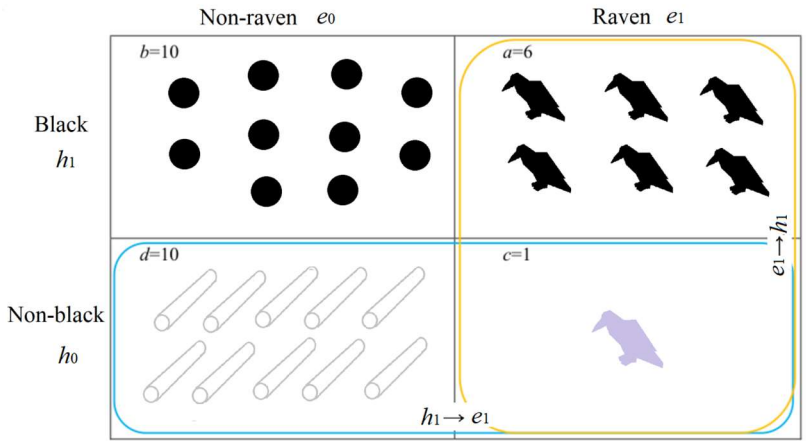


图 9. 用一个样本检查不同的确证测度是否能澄清乌鸦悖论。

首先，我们用测度 F 看我们能否用它消除乌鸦悖论。$F(e_1 \rightarrow h_1)$和 $F(h_0 \rightarrow e_0)$之间的差别是：他们的反例（个数）是相同的(c=1), 然而, 他们的正例个数是不同的。在 d 增大到 $d+\Delta d$ 时，$F(e_1 \rightarrow h_1) = (ad-bc)/(ad+bc+2ac)$ 和 $F(h_0 \rightarrow e_0) = (ad-bc)/(ad+bc+2dc)$ 不等地增大。因此，虽然测度 F 否定等价条件，它仍然肯定 Δd 同时影响 $F(e_1 \rightarrow h_1)$ 和 $F(h_0 \rightarrow e_0)$。所以测度 F 并不符合 Nicod 准则。

在现存的所有确证测度中，只有 c^*确保 $f(a, b, c, d)$只受 a 和 c 影响。我们有 $c^*(e_1 \rightarrow h_1) = (6-1)/6=5/6$.

然而，很多研究者仍然认为 Nicod 准则是不正确的。他们认为这一准则符合我们的直觉只是因为确证测度 $c(e_1 \rightarrow h_1)$ 明显地随 a 增大并且稍许随 d 增大。比如, Fitelson 和 Hawthorne [62] 相信似然比 LR 可以用来解释一个黑乌鸦能比一个非黑非乌鸦物体更有力地支持“乌鸦是黑的”。这是真的吗？

对于上面例子，LR, F 和 b^*确能使 Δa=1 引起的增量 Δf 大于 Δd=1 引起的增量；然而，在 a=d=20 和 b=c=10 时，除了 c^*, 没有一个测度能使 $\Delta f/\Delta a > \Delta f/\Delta d$ (见[58]中表 13), 这意味着，流行的所有确证测度中没有一个能用来解释一只黑乌鸦能比一只白粉笔更有力地支持 “乌鸦是黑的”。

因为 $c^*(e_1 \rightarrow h_1) = (a\text{-}c)/\max(a, c)$ 且 $c^*(h_0 \rightarrow e_0) = (d\text{-}c)/\max(d, c)$, 等价条件并不成立, 测度 c^* 完全符合 Nicod 准则，因此，根据测度 c^*, 乌鸦悖论不再存在。

在 Sally Clark 案[②]中，法官根据两个孩子死亡(e_1)判定母亲有罪(h_1)。如果使用 c^*测度，$c^*(e_1 \rightarrow h_1)$将小于 0，因而不能判定该母亲有罪. 而使用其他确证测度则不容易得到这样的结论。

5.4 确证测度 F, b^*和 c^*如何兼容波普尔的证伪思想

波普尔肯定一个反例能证伪一个全称假设或大前提。然而，对于不确定大前提，反例如何影响确证度？确证测度 F, b^*和 c^*能反映：较少反例的存在比较多正例的存在更重要。

波普尔 肯定一个反例能证伪一个全称假设，据此，我们可以解释：对于严格全称假设，没有反例更重要。现在对于不严格全称假设或大前提确证，我们可以解释较少的反例比较多正例更重要。因此，确证测度 F, b^*和 c^*和波普尔证伪思想兼容。

Scheffler 和 Goodman [61] 曾提出基于波普尔证伪思想的**选择确证**。他们相信，黑乌鸦（数目是 a）支持“乌鸦是黑的”是因为它否定 "乌鸦不是黑的。" 他们的理由——关于为什么非黑的乌鸦（数目是 c）支持相反的假设“乌鸦不是黑的”—— 是因为非黑的乌鸦否定“乌鸦是黑的”。他们的解释是有意义的。但是他们没有提出相应的确证测度。测度 $c^*(e_1 \rightarrow h_1)$ 就是他们需要的。

现在我们能发现，正是因为确证，我们可以并不放弃一个被证伪的大前提，并连同优化的可信度保留它，以便获得更多信息。

更多关于确证的讨论见[58].

6 归纳、推理、模糊三段论和模糊逻辑

6.1 从新的视角看归纳

从统计学习的视角看，归纳包括:

- 为概率预测归纳：用样本分布优化似然函数；
- 为标签外延归纳：用样本分布优化真值函数；
- 为大前提的确证度归纳：在分类后产生大前提时，用正反例比例优化大前提的确证度。

从上述视角看，归纳包含确证，确证是归纳的一部分。 只有推理规则——即包含两个谓词的大前提“如果-那么”——需要确证。

上面归纳方法也包含预测模型（似然函数或真值函数）的选择和猜想。求解混合模型和最大互信息分类的迭代算法就反映重复的猜想和反驳[12]。因此，这些归纳方法和波普尔的证伪思想兼容。

---

② 网上搜索“一个被错误的统计证据毁掉一生的母亲”可见。

6.2 贝叶斯推理作为三段论的不同形式

根据上面分析，贝叶斯推理有多种形式，如表 4 所示。其中加重字体部分是新增加的——基于 P-T 概率框架。

表 4. 贝叶斯推理的不同形式

推理方式	**大前提或模型**	**小前提或证据**	**结论**	**解释**
在两个实例之间	$P(y_j\|x)$	x_i (X=x_i)	$P(y_j\|x_i)$	条件统计概率
		y_j, $P(x)$	$P(x\|y_j)=P(x)P(y_j\|x)/P(y_j)$ $P(y_j)=\sum_i P(y_j\|x_i)\ P(x_i)$	贝叶斯定理 2 (贝叶斯预测)
在两个集合之间	$T(\theta_2\|\theta)$	y_1 (is true)	$T(\theta_2\|\theta_1)$	条件逻辑概率
		y_2, $T(\theta)$	$T(\theta\|\theta_2)= T(\theta_2\|\theta)T(\theta)/T(\theta_2)$, $T(\theta_2)=T(\theta_2\|\theta)T(\theta)+T(\theta_2\|\theta')T(\theta')$	贝叶斯定理 1 (θ' 是 θ 的补集)
在一个实例和一个集合(或模型)之间	$\boldsymbol{T(\theta_j\|x)}$	$\boldsymbol{X=x_i}$ **or** $\boldsymbol{P(x)}$	$\boldsymbol{T(\theta_j\|x_i)}$ **or** $\boldsymbol{T(\theta_j)=\sum_i T(\theta_j\|x_i)P(x_i)}$	**命题真值**
		$\boldsymbol{y_j}$ **is true,** $\boldsymbol{P(x)}$	$\boldsymbol{P(x\|\theta_j)=P(x)T(\theta_j\|x)/T(\theta_j)}$, $\boldsymbol{T(\theta_j)=\sum_i T(\theta_j\|x_i)P(x_i)}$	**贝叶斯定理 3 中的语义贝叶斯预测**
	$P(x\|\theta_j)$	x_i or $\mathbf{D}_j$	$P(x_i\|\theta_j)$ or $P(\mathbf{D}_j\|\theta_j)$	似然度
		$\boldsymbol{P(x)}$	$\boldsymbol{T(\theta_j\|x)=[P(x\|\theta_j)/P(x)]}$ $\boldsymbol{/\max[P(x\|\theta_j)/P(x)]}$	**贝叶斯定理 3 中的新的贝叶斯公式**
归纳：用样本分布训练预测模型	$P(x\|\theta_j)$	$P(x\|y_j)$	$P^*(x\|\theta_j)$ (optimized $P(x\|\theta_j)$ with $P(x\|y_j)$)	似然推断
	$P(x\|\theta)$ 和 $P(\theta)$	$P(x\|y)$ or $\mathbf{D}$	$P(\theta\|\boldsymbol{D})=P(\theta)P(\boldsymbol{D}\|\theta)/$ $\sum_j P(\theta_j)P(\boldsymbol{D}\|\theta_j)$	贝叶斯推断(Bayesian Inference)
	$\boldsymbol{T(\theta_j\|x_j)}$	$\boldsymbol{P(x\|y_j)}$ **and** $\boldsymbol{P(x)}$	$\boldsymbol{T^*(\theta_j\|x)}$ $\boldsymbol{=P(x\|y_j)/P(x)/\max[P(x\|y_j)/P(x)]}$ $\boldsymbol{=P(y_j\|x)/\max[P(y_j\|x)]}$	**逻辑贝叶斯推断** (Logical Bayesian Inference)
用信道确证度	$\boldsymbol{b_1^*=b^*(e_1\to h_1)>0}$	$\boldsymbol{h}$ **or** $\boldsymbol{P(h)}$	$\boldsymbol{T(\theta_{e1}\|h_1)=1}$, $\boldsymbol{T(\theta_{e1}\|h_0)=1-\|b_1^*\|}$; $\boldsymbol{T(\theta_{e1})=P(h_1)+(1-b_1^*)P(h_0)}$	**真值和逻辑概率**
		$\boldsymbol{e_1}$ **(**$\boldsymbol{e_1}$ **is true),** $\boldsymbol{P(h)}$	$\boldsymbol{P(h_1\|\theta_{e1})=P(h_1)/T(\theta_{e1})}$, $\boldsymbol{P(h_0\|\theta_{e1})=(1-b_1^*)P(h_0)/T(\theta_{e1})}$;	**使用 b^*的模糊三段论**
用预测确证度	$\boldsymbol{c_1^*=c^*(e_1\to h_1)>0}$	$\boldsymbol{e_1}$ **(**$\boldsymbol{e_1}$ **is true)**	$\boldsymbol{P(h_1\|\theta_{e1})=1/(2-c_1^*)}$, $\boldsymbol{P(h_0\|\theta_{e1})=(1-c_1^*)/(2-c_1^*)}$	**使用 c^*的模糊三段论**

*加重字体表示贝叶斯推理新形式。

重要的是：使用真值函数或确证度的推理要能和统计推理（使用贝叶斯定理 2）兼容。在 $T^*(\theta_j|x)\propto P(y_j|x)$或 $T(\theta_{ej}|h)\propto P(e_j|h)$时，新的推理结果和来自经典统计的推理结果相同。

另一方面，对于给定 e_1, 模糊三段论和经典三段论兼容，因为在 $b_1^*=1$ 或 $c_1^*=1$ 时，结果变成 $P(h_1) = 1$ 和 $P(h_0) = 0$.

但是上面三段论也有其局限性。使用 $b^*(e_1\to h_1)$的模糊三段论是一个经典三段论的推广。模糊三段论是：

- 大前提是 $b^*(e_1\to h_1) = b_1^*$
- 小前提是 e_1 和 $P(h)$,
- 结论是 $P(h|\theta_{e1}) = P(h_1)/[P(h_1) + (1-b_1^*)P(h_0)] = P(h_1)/[1-b_1^*P(h_0)]$.

当 b_1*=1 时，如果小前提变成 “x 在 E_2 中且 E_2 在 E_1 中”, 那么这一三段论就变成 Barbara (AAA-1) [63], 且结论是 $P(h_1|\theta_{e1}) = 1$。 所以上述模糊三段论可看作模糊 Barbara.

因为对于信道确证和预测确证，等价条件都不成立，如果小前提是 $h = h_0$, 我们只能使用反确证测度 $b^*(h_0 \to e_0)$ 或 $c^*(h_0 \to e_0)$作为大前提，进而获得结论 (详见 [58])。

6.3 模糊逻辑：期望和问题

真值函数也是隶属函数。模糊逻辑能简化隶属函数和真值函数的统计。

假定三个模糊集合 A、B、C 的隶属函数分别是 $a(x)$、$b(x)$、$c(x)$；A、B、C 的逻辑表达式是 $F(A, B, C)$，其中有三个运算符 $\cap, \cup, ^c$；真值函数的逻辑表达式是 $f(a(x), b(x), c(x))$，其中有三个运算符 $\wedge, \vee, \bar{}$. 运算符 $\cap$ 和 $\wedge$ 可以省略。则一共有 2^8 个不同的逻辑表达式。为了简化统计，我们希望

$$T(F(A, B, C)|x) = f(a(x), b(x), c(x)). \tag{48}$$

如果上面等式成立，我们只需要优化三个真值函数 $a(x)$、$b(x)$、$c(x)$，并计算各种逻辑表达式 $F(A, B, C)$的真值函数 $f(a(x), b(x), c(x))$。进一步，我们希望 $f(...)$中的模糊逻辑和布尔代数以及柯氏概率系统兼容。

然而，一般说来，隶属函数的逻辑运算并不和布尔代数兼容。扎德的模糊逻辑被定义为:

$$\begin{aligned} a(x) \wedge b(x) &= \min(a(x), b(x)), \\ a(x) \vee b(x) &= \max(a(x), b(x)), \\ a(x) &= 1\text{-}a(x). \end{aligned} \tag{49}$$

根据这一定义，互补律并不成立，因为

$$a(x) \wedge \bar{a}(x) = \min(a(x), 1\text{-}a(x)) \neq 0, \quad a(x) \vee \bar{a}(x) = \max(a(x), 1\text{-}a(x)) \neq 1.$$

我们也能看到其他定义。考虑到两个谓词“x 在 A 中” 和 “x 在 B 中”之间的相关性，汪培庄等人[64]定义:

$$a(x) \wedge b(x) = \begin{cases} \min(a(x), b(x)), & \text{正相关}, \\ a(x)b(x), & \text{相互独立}, \\ \max(0, a(x)+b(x)-1), & \text{负相关}. \end{cases} \tag{50}$$

$$a(x) \vee b(x) = \begin{cases} \max(a(x), b(x)), & \text{正相关}, \\ a(x)+b(x)-a(x)b(x), & \text{相互独立}, \\ \min(1, a(x)+b(x)), & \text{负相关}. \end{cases} \tag{51}$$

如果两个谓词总是正相关的，上面运算就变成扎德的运算。因为 A 和 A^c 是负相关的, 于是有

$a(x) \wedge \bar{a}(x)=\max(0, a(x)+1-a(x)-1) \equiv 0, \quad a(x) \vee \bar{a}(x) = \min(1, a(x)+1-a(x)) \equiv 1.$

因而互补律是成立的。

为了建立一个对称的色觉机制模型[65, 66]，我于上世纪 80 年代提出——用直集 $[0,1]\times\mathcal{X}$ 上的布尔代数定义 $\mathcal{X}$ 上的模糊准布尔代数。这一色觉机制模型被解释为模糊三八译码器，因此被称为译码模型(见附录 V)．其实用性能被下面事实证明：国际照明协会(CIE)于 2006 年推荐了和我的模型几乎相同的对称颜色模型用于颜色转换[67]。使用译码模型，我们能容易解释色盲和色觉进化——通过视锥细胞敏感曲线的分裂和合并 [68].

我们能用模糊准布尔代数简化自然语言真值函数的统计。这中情况下，我们需要区分原子标签和复合标签，假定表示这些标签外延的随机集合同时变宽或变窄。

设三个原子标签“年轻人”、“成年人”和“老年人”的真值函数分别是 $u(x)$、a(x)和 $e(x)$。三个复合标签的真值函数是：

$$T(\text{“小孩”}|x) = T(\text{“非年轻人” 且“非成年人”}|x) = [\overline{u}\,\overline{a}] = 1\text{-}\max(u(x), a(x)),$$

$$T(\text{“年轻人” 和 “非成年人”}|x) = [u\,\bar{a}] = \max(0, u(x)\text{-}a(x)),$$

$$T(\text{“中年人”}|x) = T(\text{“成年人”且 “非年轻人” 且 “非老年人”}|x)$$

$$= [a\,\overline{u}\,\overline{e}] = \max(0, a(x)\text{-}\max(u(x),e(x)).$$

然而，不同标签或叙述之间的相关性常常是复杂的。当我们选择一组模糊逻辑运算时，我们需要权衡合理性和可行性。

7 讨论

7.1 P-T 概率框架如何经受多种理论应用的检验

P-T 概率框架包含统计概率和逻辑概率；逻辑概率包含真值函数，它也就是隶属函数，能反映假设或标签的外延。使用真值函数 $T(\theta_j|x)$ 和先验分布 $P(x)$, 我们能产生似然函数 $P(x|\theta_j)$，并能用样本分布 $P(x|y_j)$训练 $T(\theta_j|x)$和 $P(x|\theta_j)$ 。然后我们能让机器像人脑一样用概念外延推理。

使用基于这一框架的语义信息方法，我们能用逻辑概率表达信息量: $I(x_i; \theta_j) = \log[T(\theta_j|x)/T(\theta_j)]$。我们也能用似然函数表达预测信息，因为 $I(x_i; \theta_j) = \log[T(\theta_j|x)/T(\theta_j)] = \log[P(x_i|\theta_j)/P(x_i)]$。在我们计算平均语义信息的时候，我们也需要统计概率，比如用 $P(x|y_j)$表示样本分布。

在流行的统计学习方法中，只用到统计概率（包含主观概率）。对于二元分类，我们能用一对 Logistic 函数。但是对于多标签学习和分类，缺少简单方法[40]。现在使用(模糊)真值函数，我们可以使用逻辑贝叶斯推断[12]求解多个标签的外延，从而便于分类。 使用统计概率构成的香农信道和逻辑概率构成的语义信道，我们可以让两个信道相互匹配，实现最大互信息分类，加速混合模型收敛，并提供更好的收敛证明[12]。

3.3 节显示：带有真值函数和逻辑概率的语义贝叶斯公式已经出现在信息率失真理论和统计力学中。因此，P-T 概率框架不仅能用于通信或认识论，也能用于控制或本体论。

关于科学理论或假设的评估，使用逻辑概率，我们能用语义信息量 $I(x; \theta_j)$表示逼真度和经受检验严厉性有多好——两者可以相互转换。使用统计概率，我们能表达样本分布如何检验假设。

关于确证，使用真值函数，我们能方便地表达大前提的可信度。用统计概率，我们能推导出确证度——即用样本分布优化的可信度。

关于贝叶斯推理, 使用 P-T 概率框架，我们能有更多形式的推理（见表 4），比如

- 使用贝叶斯定理 3 的两种推理,
- 逻辑贝叶斯推断——从样本分布得到优化的真值函数，
- 模糊三段论——使用大前提的确证度。

简言之，使用 P-T 概率框架，我们能更加方便地解决许多问题。

7.2 怎样推广逻辑到概率理论？

杰尼斯[16] 总结道：概率论是逻辑的推广。这一结论不同于下面结论：概率是逻辑的推广。 前者包含后者。前者强调概率是科学推理的重要工具，推理包括：贝叶斯推理，最大似然估计，最大熵方法，等等。虽然杰尼斯用平均真值解释概率 [16](p.52), 这一解释并不是推广逻辑。 当一个罐子里包含很多只有两种颜色（白和红）的球时，我们可以用真值 R_i 表示第 i 次取出的红球，用 $\overline{R}_i = W_i$ 表示第 i 次取出的黑球。然而，如果球有更多颜色，这种解释就没那么自然了。这种情况下，频率主义的解释更加简单。

扎德的模糊集合论是经典逻辑的重要推广。使用（模糊）真值函数或隶属函数，我们能让机器学习人脑使用模糊外延推理。这一推理是杰尼斯总结的推理的重要补充。

使用柯氏公理系统，有些研究者希望发展概率逻辑——使用逻辑表达式 $f(T(A), T(B),\ldots)$， 使得

$$T(F(A, B, \ldots))= f(T(A), T(B),\ldots). \tag{52}$$

这一任务是非常困难的，因为要避免概率逻辑和统计推理的矛盾是非常困难的。根据柯氏公理系统，我们有

$$T(A \cup B)=T(A)+T(B)-T(A\cap B). \tag{53}$$

但是我们并不能从 $T(A)$ 和 $T(B)$获得 $T(A\cap B)$ ，因为 $T(A\cap B)$ 和 $P(x)$相关，或者和两个谓词 “x 在 A 中”和“x 在 B 中”的相关性相关。运算符号 “∧”和 “∨" 只能用于真值运算而不能用于逻辑概率运算。比如,A 是集合 {年轻人},B 是集合 {成年人}。假设当人群包含高中生和他们的父母时，先验分布是 $P_1(x)$; 当人群包含高中生和士兵

时，先验分布是 $P_2(x)$。$T_1(A)$是从 $P_1(x)$和 $T(A|x)$得到的, $T_2(A)$是从 $P_2(x)$ 和 $T(A|x)$, 其他同理。$T_1(A\cap B)$ 应当远比 $T_2(A\cap B)$小, 即使 $T_1(A)=T_2(A)$且 $T_1(B)=T_2(B)$.

要得到 $T(A\cap B)$，比等式 (52)更好的方法是，首先得到真值函数，比如 $T(A\cap B|x)=T(A|x)\wedge T(B|x)$， 然后我们有

$$T(A\cap B)=\sum_i P(x_i)T(A\cap B\,|\,x_i). \tag{54}$$

要得到 $T(F(A, B, \ldots))$类似，我们应当首先用模糊逻辑得到 $f(T(A|x), T(B|x), \ldots)$。

表 4 中的各种贝叶斯推理作为经典逻辑的推广和统计推理兼容。比如, 贝叶斯定理 3 和贝叶斯定理 2 兼容。如果我们把一种概率逻辑用于模糊或不确定三段论，我们最好检查推理是否和统计推理兼容。

赖欣巴哈 [9]和 Adams [26]的概率逻辑中有

$$P(p\Rightarrow q) = 1\text{-}P(p)+P(pq), \tag{55}$$

它是数理逻辑中 $p\Rightarrow q$ 的推广，其中 p 和 q 是两个命题序列或两个谓词。我的推广不同。我用确证测度 $b^*(e_1\rightarrow h_1)$ 和 $c^*(e_1\rightarrow h_1)$ （可能是负的）表示推广的大前提。表 4 中大多数推广的三段论和贝叶斯公式相关。测度 $c^*(p\rightarrow q)$ 是 $P(q|p)$的函数(见等式 (44)); 它们是兼容的。用 $P(q|p)$作为评估模糊大前提的测度应该也是合理的， 但是 $P(q|p)$和 $P(p\Rightarrow q)$是不同的，我们能证明

$$P(q|p)\le P(p\Rightarrow q) = 1\text{-}P(p)+P(pq). \tag{56}$$

证明见附录 VI。 $P(p\Rightarrow q)$可能比 $P(q|p)$大很多。比如 $P(p)=0.1$ 且 $P(pq)=0.02$ 时，$P(q|p)=0.2 < P(p\Rightarrow q)=0.92$.

使用等式 (55)是因为在数理逻辑中，$(p\Rightarrow q)=\overline{p}\vee q=\overline{p}\vee pq$ 。然而，不管多少实验支持 $p\Rightarrow q$, “如果 p 则 q” 的可信度都不可能是 1，如休谟(Hume) 和 波普尔指出[7]。 当有反例存在时，$p\Rightarrow q$ 变成 $P(p\Rightarrow q) = 1\text{-}P(p)+P(pq)$——这作为评估模糊大前提也是不合适的。

另外，从 $P(q|p)$和 $P(p)=1$ 获得 $P(q)$或 $P(pq)$要比从 $P(p\Rightarrow q)$ 和 $P(p)=1$ 获得它 [9, 26]要容易得多。如果 $P(p)<1$, 根据统计推理，我们还需要 $P(q\,|\,\overline{p})$ ，从而获得 $P(q) = P(p)P(q|p)+[1\text{-}P(p)]\,P(q\,|\,\overline{p})$ 。使用一种概率逻辑，我们需要小心结果是否和统计推理兼容。

7.3 比较真值函数和费希尔的反概率函数

在 P-T 概率框架中，真值函数起到重要作用。

用似然函数 $P(x|\theta_j)$（θ_j 是常数）作为推断工具叫似然推断（推断参数）， 用贝叶斯后验 $P(\theta|\mathbf{D})$ (给定数据或好样本 **D** 时)作为推断工具叫贝叶斯推断。使用真值函数 $T(\theta_j|x)$作为推断工具叫逻辑贝叶斯推断[12].

Fisher 称参数化的转移概率函数 $P(\theta_j|x)$为反概率 [5]。因为 x 是变量，我们最好称 $P(\theta_j|x)$为反概率函数。根据贝叶斯定理 2，我们有

$$P(\theta_j|x)=P(\theta_j)\,P(x|\theta_j)/P(x), \tag{57}$$

$$P(x|\theta_j)=P(x_i)\,P(\theta_j|x)/P(\theta_j). \tag{58}$$

使用反概率函数 $P(\theta_j|x)$时，我们同样能利用先验知识 $P(x)$ 。 当 $P(x)$改变时，$P(\theta_j|x)$ 仍然能用来做概率预测。但是为什么 Fisher 和其他研究者放弃 $P(\theta_j|x)$ 作为推断工具？

当 n=2, 我们能容易地用参数构造 $P(\theta_j|x), j$=1,2。比如，我们能使用一对 Logistic 函数作为反概率函数. 不幸的是，当 n>2 时, 要构造 $P(\theta_j|x), j$=1,2,…,n 是非常困难的，因为对于每个 x 存在归一化限制$\Sigma_j P(\theta_j|x)$=1。这就是为什么多标签分类常常要转换为多个二标签分类[40].

$P(\theta_j|x)$ 和 $P(y_j|x)$作为预测模型也有严重缺点. 在许多情况下，我们只知道 $P(x)$和 $P(x|y_j)$而不知道 $P(\theta_j)$ 或 $P(y_j)$，以至于我们不能获得 $P(y_j|x)$或 $P(\theta_j|x)$。但是，在这些情况下我们能获得 $T^*(\theta_j|x)$。因为没有归一化限制，要构造一组真值函数并用 $P(x)$ 和 $P(x|y_j), j$=1,2,…,n, 训练它们是容易的，尽管没有 $P(y_j)$或 $P(\theta_j)$.

可以总结，真值函数有下面优点:

- 对于不同的 $P(x)$，我们能用优化的真值函数 $T^*(\theta_j|x)$做概率预测，就像用 $P(y_j|x)$或 $P(\theta_j|x)$一样。
- 我们能用不太大的样本训练真值函数，就像训练似然函数一样。

  真值函数能表明假设的语义或标签外延。它也是隶属函数，适合于分类。
- 要训练一个真值函数 $T(\theta_j|x)$, 我们只需要 $P(x)$和 $P(x|y_j)$, 而不需要 $P(y_j)$或 $P(\theta_j)$.
- 令 $T^*(\theta_j|x)\propto P(y_j|x)$, 我们可以在统计和逻辑之间搭起桥梁。

7.4 回答一些问题

如果概率理论是科学的逻辑，任何科学陈述能表达成概率的或不确定陈述吗？P-T 概率框架支持杰尼斯的观点？回答是“是!”高斯真值函数能用来表达带有很小范围不确定性的物理量，就像它用来表达 GPS 指针的语义。语义信息公式可以用来评价物理公式预测的结果。郭嗣宗[69]研究模糊数运算得到结论：模糊数函数的期望等于期望的函数，比如,(模糊 3)*(模糊 5)=(模糊 15)=(模糊 3*5)。据此，概率的或不确定叙述和物理学规律的确定叙述兼容。

关于现存的概率理论、P-T 概率框架和 G 理论是否是科学理论，以及它们是否能被经验事实检验，我的回答是它们不是科学理论，而是科学理论及其应用的工具，它们能被检验——通过它们解决科学理论和应用中问题的能力。

关于评估科学理论时，是否要求这些理论有逻辑一致性约束，我相信一种科学理论不仅需要自身一致， 还需要和其他已经被接受的理论兼容。类似的是，一个概率和逻辑的统一理论作为科学工具，也需要和已经被接受的数学理论（比如统计学）兼容。为此，在 P-T 概率框架中，优化的真值函数（或语义信道）在用于概率预测时

等价于一个转移概率函数（或一个香农信道）；语义互信息的上限是香农互信息；模糊逻辑最好兼容布尔代数。

7.5 需要进一步研究的一些问题

许多研究者在谈论可解释人工智能（AI），然而，意义问题在今天仍然是亟需解决的[70]。P-T 概率框架对可解释的 AI 应该是有帮助的，理由是:

- 人脑用概念外延而不是相关性思考。真值函数能表示标签的（模糊）外延，反映标签的语义；贝叶斯定理 3 表达了如何用概念外延推理。
- 新的确证方法和模糊三段论能表达归纳和推理——用人脑使用的可信度推理，该推理和统计推理兼容。
- 为了机器学习，玻尔茨曼分布已经用于玻尔茨曼机 [71]。借助于语义贝叶斯公式和语义信息方法，我们能联系信息理论更好理解这一分布和正则化最小平方准则。

然而，将语义信息方法用于机器学习时，要解释神经网络仍然不容易。我们能把一个神经网络解释为一个语义信道，它有一组真值函数构成？我们能把信道匹配算法[12]用于神经网络以便实现最大互信息分类？这些问题需要进一步研究。

本文提供了统计和逻辑之间的桥梁，它对 AI 需要的基于统计的语义字典是有帮助的。然而，仍然有大量工作要做，比如设计和优化自然语言中术语——比如“老年人”、“大雨”、“正常体温”——的真值函数。困难的是，这些术语的外延因地而异。比如，“大雨”的外延在沿海地区和沙漠地区是不同的。这些外延和和先验分布 $P(x)$有关。 要统一人工智能中的统计方法和逻辑方法，还需要更多人的努力。

当样本不够大时，上面方法求得的确证度是不可靠的，在这些情况下，我们需要用确证区间代替确证度，我们需要进一步研究。

本文只推广一个基本的三段论——其大前提是“如果 $e=e_1$ 则 $h=h_1$”——到某些有效的模糊三段论。要推广更多三段论是非常复杂的 [63]。我们需要用到更复杂的模糊三段论，我们需要进一步研究。

我们可以使用真值函数作为分布约束函数，从而推广信息和熵测度，把它们用于随机事件的控制和统计力学 (见 3.3 节)。我们需要进一步研究以便得到实用结果。

8 结论

如波普尔指出, 一个假设同时有统计概率和逻辑概率。本文提出 P-T 概率框架——“P”表示统计概率，“T”表示逻辑概率。在此框架中，一个假设的真值函数等价于一个模糊集合的隶属函数。使用新的贝叶斯定理——贝叶斯定理 3, 我们能把似然函数和真值函数从一个转变为另一个，据此我们可以用样本分布训练真值函数。所用最大语义信息准则等价于最大似然准则。统计和逻辑因此被连接。

本文介绍了怎样把 P-T 概率框架用于语义信息 G 理论（或 G 理论）。G 理论是香农信息论的自然推广，它可以用来改进统计学习，并有助于解释信息和热力学熵之间的关系——最小互信息分布等价于最大熵分布。

本文说明了 P-T 概率框架和 G 理论支持 波普尔思想——关于科学进步、假设评价和证伪；语义信息方法可以调和关于逼真度的内容方法和相似性方法[52]。

本文介绍了如何通过语义信息测度推导出信道确证测度 b^* 和预测确证度 c^*。测度 b^*兼容似然比，在-1 和 1 之间变化。它能用来评价医学检验、信号检测、分类。测度 c^* 可以用来评估概率预测，澄清乌鸦悖论。两个确证测度都兼容波普尔证伪思想。

本文提供了几种不同形式的贝叶斯推理——包括使用确证测度 b^*和 c^*的推理，还介绍了一个新的模糊逻辑——用来建立一个对称的色觉机制模型。该模糊逻辑和布尔代数兼容，因此兼容经典逻辑。

P-T 概率框架的上述理论应用揭示了其合理性和实用性。我们应能发现其更广的应用。然而，要结合 AI 中的逻辑方法和统计方法。仍然有许多工作要做。为了把 P-T 概率框架和 G 理论用到深度学习和随机事件的控制得到实用结果，我们需要进一步研究。

附录 I: 贝叶斯定理 3 证明

因为联合概率 $P(x, \theta_j)=P(X=x, X\epsilon\theta_j)= P(x|\theta_j)T(\theta_j) = T(\theta_j|x)P(x)$, 于是有

$$P(x|\theta_j)=P(x)T(\theta_j|x)/T(\theta_j),\ T(\theta_j|x)=T(\theta_j)P(x|\theta_j)/P(x).$$

因为 $P(x|\theta_j)$是水平归一化系数，等式（8）中的 $T(\theta_j)$是$\sum_i P(x_i)\,T(\theta_j|x_i)$。因为 $T(\theta_j|x)$是纵向归一化的（最大值是 1），我们有

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

于是 $T(\theta_j)=1/\max[P(x|\theta_j)/P(x)]$。

附录 II: 一个 $R(D)$ 函数等价于一个 $R(\Theta)$ 函数——都含有真值函数

设 $T(\theta_{xi}|y)=\exp(sd_{ij})$, 我们有

$$\begin{aligned}R(\Theta)=\min_{P(x),\Theta_x} I(X;Y)=I(Y;\Theta_x)&=\sum_i P(x_i)\sum_j P(y_j|x_i)\log\frac{T(\theta_{xi}|y_j)}{T(\theta_{xi})}\\&=\sum_i P(x_i)\sum_j P(y_j|x_i)\log\frac{\exp(sd_{ij})}{\lambda_i}\\&=sD(s)-\sum_i P(x_i)\log\lambda_i=\min_{P(x),D} I(X;Y)=R(D).\end{aligned}$$

附录 III: 信息和热力学熵之间的联系

Helmholtz 自由能是

$$F=E-TS.$$

对于平衡系统，

$$S = E / T - F = E / T - kN \ln Z, \quad Z=\sum_i G_i \exp[-e_i / (kT)]$$

其中 N 是粒子数。对于局域平衡系统,

$$S = \sum_j \frac{E_j}{T_j} + k\sum_j N_j \ln Z_j$$

其中 T_j 是第 j 个区域 (y_j)的温度, N_j 是第 j 个区域的粒子个数。我们考虑给定分布约束函数 $T(\theta_j|x)=\exp[-e_i/(kT_j)]$ (j=1,2,…)时的最小香农互信息 $R(\Theta)$。y_j 的逻辑概率是 $T(\theta_j)=Z_j/G$, 统计概率是 $P(y_j)=N_j/N$。从附录 II 和上面公式，我们推导出

$$\begin{aligned} R(\Theta) &= -\sum_j P(y_j)\frac{e_j}{kT_j} - \sum_j P(y_j)\ln(Z_j / G) \\ &= -\sum_j \frac{N_j}{N}\frac{(E_j / N_j)}{kT_j} - \sum_j \frac{N_j}{N}\ln(Z_j / G) = \ln G - S/(kN) \end{aligned}$$

其中 $e_j=E_j/N_j$ 是第 j 个区域的一个粒子的平均能量。

附录 IV: b_1*的推导

令 $P(h_1|\theta_{e1})=P(h_1|e_1)$. 从

$$P(h_1 \mid e_{\theta 1}) = \frac{P(h_1)T(\theta_{e1} \mid h_1)}{T(\theta_{e1})} = \frac{P(h_1)}{P(h_1) + b_1' P(h_0)},$$

$$P(h_1 \mid e_1) = \frac{P(h_1)P(e_1 \mid h_1)}{P(h_1)P(e_1 \mid h_1) + P(h_0)P(e_1 \mid h_1)},$$

我们得到 b_1'*= $P(e_1|h_0)/P(e_1|h_1)$ （对于 $P(h_1|e_1) \geq P(h_0|e_1)$）。因此,

$$b_1^*=1-b_1'^*=[P(e_1|h_1)-P(e_1|h_0)]/P(e_1|h_1).$$

附录 V: 图解色觉译码模型中使用的模糊逻辑

从三元色信号 b, g, r, 我们使用模糊译码产生 8 个心理信颜色号输出，如图 A1 所示。

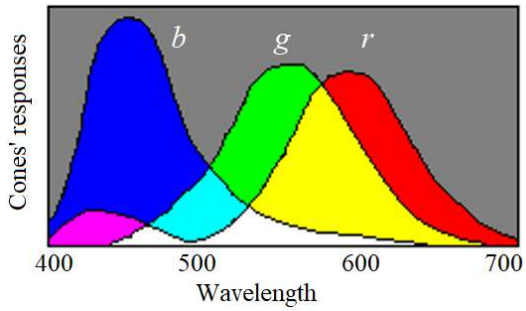


图 A1. 图解译码模型中的模糊逻辑

附录 VI: 证明 $P(q|p) \le P(p \Rightarrow q)$

因为 $P(p)=P(pq)+P(p\bar{q}) \le 1$，我们有 $P(pq) \le 1- P(p\bar{q})$，因此

$$P(q \mid p) = \frac{P(pq)}{P(p)} = \frac{P(pq)}{P(pq)+P(p\bar{q})} \le \frac{1-P(p\bar{q})}{1-P(p\bar{q})+P(p\bar{q})}$$
$$= 1-P(p\bar{q}) = P(\bar{P} \vee Pq) = 1-P(p)+P(pq) = P(p \Rightarrow q).$$

感谢： 我感谢 *Philosophies* 和其特刊 *Science and Logic* 的编辑给我这个机会让我系统地介绍我的这一研究。我也感谢两个审稿人，他们的意见大大改进了本文。我还应该感谢改变我命运的许多人，他们使我有技术的和理论的研究经验，从而写出本文。

参考文献

1. Galavotti M C. The Interpretation of Probability: Still an Open Issue? Philosophies, **2017** 2(3): 20.
2. Hájek A. Interpretations of probability. The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.). Available online: https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/. (accessed on 17 June 2020).
3. Shannon C E. A mathematical theory of communication. Bell Syst. Tech. J. **1948**, 27, 379–429, 623–656.
4. Fisher R A. On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. **1922**, 222, 309–368.
5. Fienberg S E. When did Bayesian Inference become "Bayesian"? Bayesian Anal. **2006** 1(1): 1–40.
6. Popper K R. The propensity interpretation of the calculus of probability and the quantum theory// Körner, S. Ed., The Colston Papers No 9, London: Butterworth Scientific Publications, 1957, 65–70.
7. Popper K. The Logic of Scientific Discovery, London and New York: Routledge Classic, 2002 (德文版出版于 1935).
8. Carnap R. The two concepts of Probability: the problem of probability. Philosophy and Phenomenological Research, **1945**, 5(4): 513–532.
9. Reichenbach H. The Theory of Probability, Berkeley: University of California Press, 1949.

10. Kolmogorov A N. Foundations of Probability, New York: Chelsea Publishing Company, 1950 (德文版 1933 出版).
11. Greco S.; Slowiński R.; Szczech I. Measures of rule interestingness in various perspectives of confirmation. Inf. Sci., **2016**, 346–347, 216–235.
12. Lu C. Semantic information G theory and Logical Bayesian Inference for machine learning. Information, **2019**, 10(8): 261. https://www.mdpi.com/2078-2489/10/8/261
13. Lu C. Shannon equations' reform and applications. BUSEFAL **1990**, 44, 45–52. Available online: https://www.listic.univ-smb.fr/production-scientifique/revue-busefal/version-electronique/ebusefal-44/.
14. 鲁晨光. 广义信息论. 合肥：中国科技大学出版社, 1993.
15. Lu C. A generalization of Shannon's information theory. Int. J. Gen. Syst. **1999**, 28(6): 453–490.
16. Jaynes E T. Probability Theory: The Logic of Science. Cambridge: Cambridge University Press, 2003.
17. Dubois D.; Prade H. Possibility theory, probability theory and multiple-valued logics: A clarification, Annals of Mathematics and Artificial Intelligence, **2001**, 32, 35–66.
18. Carnap R. Logical Foundations of Probability, Chicago: University of Chicago Press, 1962.
19. Zadeh L A. Fuzzy sets. Inf. Control, **1965**, 8, 338–353.
20. Zadeh L A. Probability measures of fuzzy events. J. Math. Anal. Appl. **1986**, 23, 421–427.
21. Zadeh, L. A. Fuzzy set theory and probability theory: What is the relationship? // Lovric, M. Ed., International Encyclopedia of Statistical Science, Heidelberg: Springer, 2011.
22. Wang P Z. From the fuzzy statistics to the falling random subsets. Wang, P. P. Ed. Advances in Fuzzy Sets, Possibility Theory and Applications, New York: Plenum Press,1983, 81–96.
23. 汪培庄. 模糊集合和随机集落影. 北京: 北京师范大学出版社，1985.
24. Keynes J M. A Treatise on Probability, London: Macmillan and Co., 1921.
Demey L.; Kooi B.; Sack J. Logic and probability. Edward N. Z. Ed. The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), https://plato.stanford.edu/archives/sum2019/entries/logic-probability/.
25. Adams E W. A Primer of Probability Logic. Stanford: CSLI Publications, 1998.
26. Dubois D.; Prade H. Fuzzy sets and probability: misunderstandings, bridges and gaps. Proceedings 1993 Second IEEE International Conference on Fuzzy Systems, San Francisco, USA, 1993, 1059-1068.
27. Coletti G.; Scozzafava R. Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Sets and Systems **2004**, 144(1): 227-249.
28. 高庆狮，高小宇，胡月. 概率论基础部分与模糊集合理论的统一定义. 大连理工大学学报，**2006** 46(1): 141-150.
29. von Mises, R, Probability, Statistics and Truth (second revised English edition). London：George Allen and Unwin Ltd, 1957.
30. Tarski A. The semantic conception of truth: and the foundations of semantics. Philos. Phenomenol. Res. **1994**, 4(3): 341–376.
31. Davidson, D. Truth and meaning. Synthese, **1967**, 17(3): 304–323.

32. Carnap R.; Bar-Hillel Y. An Outline of a Theory of Semantic Information. Technical Report No. 247, Research Lab. of Electronics, Cambridge: MIT, 1952.
33. Bayes T.; Price R. An essay towards solving a problem in the doctrine of chance. Philos. Trans. R. Soc. Lond. **1763**, 53, 370–418.
34. Wittgenstein, L. Philosophical Investigations. Oxford: Basil Blackwell Ltd, 1958.
35. Wikipedia contributors, Dempster–Shafer theory. Wikipedia, The Free Encyclopedia. 29 May 2020. Available online: https://en.wikipedia.org/wiki/Dempster%E2%80%93Shafer_theory.
36. Lu C. The third kind of Bayes' Theorem links membership functions to likelihood functions and sampling distributions. Sun F., Liu H., Hu D. Eds. Cognitive Systems and Signal Processing, ICCSIP 2018, Communications in Computer and Information Science, vol 1006. Singapore: Springer, 2019, 268-280.
37. Floridi L. Semantic conceptions of information. Stanford Encyclopedia of Philosophy. Available online: http://seop.illc.uva.nl/entries/information-semantic/.
38. Shannon C E. Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. **1959**(4): 142–163.
39. Zhang M L.; Li Y K.; Liu X Y.; Geng X. Binary relevance for multi-label learning: An overview. Front. Comput. Sci. **2018**, 12(8): 191–202.
40. 鲁晨光. **GPS 信息和限误差信息率与信息率失真及复杂性失真之间的关系.** 成都信息工程学院学报，**2012**, 6, 27–32.
41. Sow D M. Complexity Distortion Theory, IEEE Transactions on Information Theory 2003, 49(3): 604-609.
42. Berger T. Rate Distortion Theory. Enklewood Cliffs: Prentice-Hall, 1971.
43. Wikipedia contributors, Boltzmann distribution. Wikipedia, The Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Boltzmann_distribution.
44. Popper K. Conjectures and Refutations, London and Newyork: Routledge, 2002.
45. Zhong Y X. A theory of semantic information. Proceedings **2017**, 1(3), 129; https://doi.org/10.3390/IS4SI-2017-04000.
46. Klir G. Generalized information theory. Fuzzy Sets Syst. **1991**, 40(1): 127–142.
47. Lakatos I. Falsification and the methodology of scientific research programmes. Harding S.G. Ed. Can Theories be Refuted? Synthese Library, vol 81. Dordrecht: Springer, 1976.
48. Popper K. Realism and the Aim of Science. Bartley III W W Ed., New York: Routledge, 1983.
49. Lakatos I. Popper on demarcation and induction. Schilpp, P. A. Ed., The Philosophy of Karl Popper, La Salle: Open Court:, IL, 1974, 241-273.
50. Tichý P. On Popper's definitions of verisimilitude. British Journal for the Philosophy of Science 1974, 25(2): 155-160.
51. Oddie G. Truthlikeness// Edward N. Z. Ed., The Stanford Encyclopedia of Philosophy, Available online: https://plato.stanford.edu/archives/win2016/entries/truthlikeness/.
52. Zwart S D.; Franssen M. An impossibility theorem for verisimilitude. Synthese, **2007**, 158**(1):** 75–92.
53. Eells E.; Fitelson B. Symmetries and asymmetries in evidential support. Philos. Stud., **2002**, 107(2): 129–142.
54. Kemeny J.; Oppenheim P. Degrees of factual support. Philos. Sci. 1952, 19(4): 307–324.
55. Huber, F. What Is the Point of Confirmation? Philos. Sci., **2005**, 72(5): 1146–1159.

56. Hawthorne J. Inductive Logic// Edward N. Z. Ed. The Stanford Encyclopedia of Philosophy. Available online: https://plato.stanford.edu/archives/spr2018/entries/logic-inductive/.
57. Lu C. Channels' confirmation and predictions' confirmation: from the medical test to the raven paradox. Entropy **2020**, 22(4): 384.
58. Hempel C G. Studies in the logic of confirmation. Mind, **1945**, 54(213): 1–26, 97–121.
59. Nicod J. Le Problème Logique De L'inductio, Paris: Alcan, 1924, 219. (Engl. Transl. The logical problem of induction. In Foundations of Geometry and Induction; Routledge: London, UK, 2000.)
60. Scheffler I.; Goodman N J. Selective confirmation and the ravens: A reply to Foster. J. Philos. **1972**, 69, 78–83.
61. Fitelson, B.; Hawthorne, J. How Bayesian confirmation theory handles the paradox of the ravens. Eells, E., Fetzer, J., Eds., The Place of Probability in Science. Dordrecht: Springer, 2010, 247–276.
62. Wikipedia contributors, Syllogism. Wikipedia, The Free Encyclopedia. Available on line: https://en.wikipedia.org/w/index.php?title=Syllogism&oldid=958696904 .
63. Wang P Z.; Zhang H M.; Ma X. W.; Xu W. Fuzzy set-operations represented by falling shadow theory, In Fuzzy Engineering toward Human Friendly Systems, Proceedings of the International Fuzzy Engineering Symposium '91, Yokohama, Japan, **1991**, 1, 82-90.
64. 鲁晨光. 色觉的译码模型及其验证. 光学学报, 1989, 9(2): 158-163.
65. 鲁晨光. B-模糊集合准布尔代数和广义熵公式. 模糊系统和数学, 1991, 5(1): 76-80.
66. CIELAB, Symmetric colour vision model, CIELAB and colour information technology, 2006. http://130.149.60.45/~farbmetrik/A/FI06E.PDF.
67. Lu C. Explaining color evolution, color blindness, and color recognition by the decoding model of color vision, Shi, Z.; Vadera, **S.;** Chang, E. Eds., 11th IFIP TC 12 International Conference, IIP 2020 (Hangzhou, China). Switzerland: Springer Nature, 2020, 287-298. https://www.springer.com/gp/book/9783030469306
68. 郭嗣宗. 基于结构元理论的模糊数学分析原理. 沈阳: 东北大学出版社, 2004.
69. Froese T.; Taguchi S. The Problem of Meaning in AI and Robotics: Still with Us after All These Years. Philosophies. **2019**, 4(2): 14.
70. Wikipedia contributors, Boltzmann machine. Wikipedia, The Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Boltzmann_machine.

The P–T Probability Framework for Semantic Communication, Falsification, Confirmation, and Bayesian Reasoning

Chenguang Lu

Abstract: Many researchers want to unify probability and logic by defining logical probability or probabilistic logic reasonably. This paper tries to unify statistics and logic so that we can use both statistical probability and logical probability at the same time. For this purpose, this paper proposes the P–T probability framework, which is assembled with Shannon's statistical probability framework

for communication, Kolmogorov's probability axioms for logical probability, and Zadeh's membership functions used as truth functions. Two kinds of probabilities are connected by an extended Bayes' theorem, with which we can convert a likelihood function and a truth function from one to another. Hence, we can train truth functions (in logic) by sampling distributions (in statistics). This probability framework was developed in the author's long-term studies on semantic information, statistical learning, and color vision. This paper first proposes the P–T probability framework and explains different probabilities in it by its applications to semantic information theory. Then, this framework and the semantic information methods are applied to statistical learning, statistical mechanics, hypothesis evaluation (including falsification), confirmation, and Bayesian reasoning. Theoretical applications illustrate the reasonability and practicability of this framework. This framework is helpful for interpretable AI. To interpret neural networks, we need further study.

Keywords: statistical probability; logical probability; semantic information; rate-distortion; Boltzmann distribution; falsification; verisimilitude; confirmation measure; Raven Paradox; Bayesian reasoning