评论文章：

回顾学习函数和语义信息测度的进化

--进而理解深度学习

英文原文发表于 Entropy：https://www.mdpi.com/1099-4300/25/5/802

鲁晨光 2023.5

摘要：最近深度学习出现了一股以互信息神经估计 (MINE)和信息噪声对比估计(InforNCE)为代表的潮流。在这个潮流中，相似函数和估计互信息分别用作学习函数和目标函数。巧合的是，估计互信息和笔者 30 年前提出的语义互信息在本质上完全相同。本文首先回顾了语义信息测度和学习函数的进化史，然后简单地介绍了笔者的语义信息 G 理论——包括信息率-保真度函数 $R(G)$(G 表示给定的语义互信息，$R(G)$是信息率-失真函数 $R(D)$的推广)——及其应用于多标签学习，最大互信息分类，和混合模型。然后，文章讨论了：从 $R(G)$函数和 G 理论的角度，我们应当怎样理解语义互信息和香农互信息之间的关系、构成语义互信息的两个广义熵 (模糊熵和覆盖熵)、自动编码器(AutoEncoder)、吉布斯分布(即玻尔茨曼-吉布斯分布)、划分函数…。一个重要结论是：混合模型和有限玻尔茨曼机(RBM)收敛是因为语义互信息最大化的同时香农互信息最小化，使得信息效率 G/R 接近 1。一个潜在机会是：使用高斯信道混合模型预训练深度神经网络(DNN)的隐含层，不需要考虑梯度，从而简化学习。文章也讨论了语义信息测度作为奖励函数(反映控制结果的合目的性)用于强化学习。语义信息 G 理论有助于解释深度学习，但是远远不够，因为深度学习还涉及特征抽取和模糊推理。语义信息论和深度学习相结合应能加速双方发展。

关键词：深度学习，学习函数，语义信息测度，估计互信息， 最大互信息，广义熵，相似函数, SoftMax 函数，有限玻尔茨曼机。

目录

1. 引言 2
2. 语义信息测度的进化 5
2.1. 使用真值函数或相似函数逼近香农互信息 5
2.2. 涉及语义信息测度的历史事件 7
3. 学习函数的进化 10
3.1 从似然函数到相似函数或真值函数 11
3.2. 语义相似度定义 13
3.3. 相似函数和语义信息测度用于深度学习 14
4. 语义信息 G 理论及机器学习应用 15
4.1. P-T 概率框架和语义信息 G 测度 15
4.2. 真值函数的优化和概率预测 17
4.3. 信息率保真函数 $R(G)$ 17
4.4. 信道匹配算法用于机器学习 18
4.4.1. 用于多标签学习 18
4.4.2. 不可见实例的最大互信息分类 19
4.4.3. 解释和改进用于混合模型的 EM 算法 20

5. 讨论 1：澄清一些问题 22
5.1. 互信息最大化是好目标？ 22
5.2. 解释深度神经网络：用 $R(G)$函数还是信息瓶颈？ 22
5.3. 理解吉布斯分布、划分函数、最小互信息匹配和 RBM 23
5.4. 理解模糊熵，覆盖熵和失真 24
5.5. 评价学习方法：信息准则还是正确率准则？ 25
6. 讨论 2：探索新的方法 26
6.1. 用香农信道训练高斯语义信道 26
6.2. 高斯信道混合模型和信道混合模型机 26
6.3. 用统计方法求不同单词之间的相似度 27
6.3. 合目的信息和信息价值作为奖惩函数用于强化学习 28
6.4. 语义信息 G 理论的局限性 30
7. 结论 31
英文缩写和原文 31
附录 A. 关于形式语义 32
附录 B. 增值熵和信息价值定义 32
参考文献： 33

1. 引言

信息理论的学习方法已经出现很久了，主要方法是用似然函数作为学习函数，把它带进交叉熵，然后使用最小交叉熵或最小 Kulback-Leibler (KL)距离准则优化模型参数。但是，深度学习领域最近几年出现了一个新的潮流——以互信息神经估计(MINE)【1】和信息噪声对比估计(InforNCE)【2】为代表。其中，研究者用相似函数构造出逼近香农互信息(ShMI)【3】的信息——我们且称之为估计互信息(EMI)，然后通过最大化估计互信息优化深度神经网络(DNN)。文末缩写表列出了所有缩写和英文原文。

2018 年 1 月，Belghazi 等人(包括 Bengio)【1】发表文章：*MINE: Mutual Information Neural Estimation*(互信息神经估计)，这是使用相似函数构造估计互信息的第一篇文章。2018 年 7 月， Oord 等人发表文章【2】：*Representation Learning with Contrastive Predictive Coding*，其中提出 InforNCE。该文明确提出：使用一个学习函数——希望它正比于 $m(x, y)=P(x|y)/P(x)$——并用该学习函数构造 EMI 或损失函数。MINE 和 InfoNCE 都取得了明显的进步并且鼓励了其他研究者。2019 年，Hjelm 等人(包括 Bengio)【4】，在 MINE 的基础上，提出 Deep InfoMax (DIM)学习方法。其中结合 InfoNCE 方法(见【4】中等式 (5)) 时可以取得更好结果(该文作者认为 InfoNCE 是独立于 MINE 提出来的)。2020 年，Chen 等人(包括 Hinton) 提出的 SimCLR [6]，何凯明等人提出的 MoCo【7】，Grill 等人提出的 BOYL【8】，都显示了强大的学习能力。其中都使用了和 InfoNCE 相似的估计信息或损失函数。

另一方面，相应 Weaver 于 1949 在【9】中提出的号召，有不少人一直研究语义信息论。研究语义通信和语义信息论的潮流也在逐

渐壮大。一方面是因为新一代互联网对语义通信的需求【10,11】——我们需要利用相似度和语义 (而不是香农信道) 做概率预测和压缩数据；另一方面，语义信息论用于机器学习也取得重要进展，特别是：本文作者于 2017 年【12】提出一个简单方法，用它可以从样本分布得到真值函数或相似函数；随后他又提供了一组信道匹配算法，用于多标签学习、不可见实例最大互信息分类和混合模型【12,13】。

巧合的是，上述估计信息测度和作者 30 年前提出语义信息测度在本质上是相同的。1990 年，作者提出用真值函数和相似函数构造语义信息和估计信息测度【14】。1993 年，他出版了专著《广义信息论》。1999 年，他改进了这个理论并在一个英文期刊发表了一个长篇介绍文章(37 页)【16】。自从 2017 年，他采用 P-T 概率框架和模型参数，使得这一理论可以用于机器学习【12,13,17】。

这一语义信息理论【14,15】现在被称之为语义信息 G 理论或简言之，G 理论(G 意思是推广)。相应的语义信息测度被称之为 G 测度。在 G 理论中信息率失真函数被推广为信息率保真度函数 $R(G)$，其中 R 是给定语义互信息 G 时的香农互信息的下限；而 G 是给定香农互信息 R 时的语义互信息的上限。$R(G)$函数可以揭示语义互信息和香农互信息之间的关系，有助于我们理解深度学习。其应用到机器学习从【12,13,18】可见。

语义信息和香农信息的主要区别是前者涉及真假(Floridi 有详细讨论【19】)。根据 Tarski 的关于真的语义概念【20】和 Davidson 的真值条件语义学【21】，一个假设或标签的真值函数确定了它的语义——指形式语义 (参看附录 A) 。

度量给定标签 y_j(常量)提供的关于实例 x (变量) 的语义信息时，我们要用到 x 使 y_j 为真的 (模糊) 真值函数。一般情况下，x 和 y_j 属于不同的论域。如果它们属于同一个论域，y_j 就变成估计 $\hat{x}_j$="x 大约是 x_j"。比如 GPS 指针指示的位置就是估计 $y_j = \hat{x}_j$(3.1 节详述)，它提供估计信息。因为它可能是错的，该信息也是语义信息。因此，可以说，估计信息是语义信息的特例；本文也称估计信息为语义信息，称使用估计信息测度的学习方法为语义信息论的学习方法。

虽然深度学习中估计信息方法非常成功，但是也还存在许多问题。首先，其中用到的一些函数和熵 (它们非常重要) ，但是却没有统一的名称，比如函数 $m(x, y)=P(x, y) / [P(x)P(y)]$和涉及估计互信息的广义熵。特别是作为主角的估计互信息过去一直没有自己的名字，它经常被称之为香农互信息的下限。另一方面，同一名称可能代表非常不同的函数， 比如相似度可能在 0 和 1 之间，比如 $\exp(-d(x, y))$；也可能在 0 和∞之间，比如 $m(x, y)$ 和 $-\log P(c)$ 【22】。为方便交流，我们有必要统一各种函数以及熵和信息测度的名称。

第二，深度学习研究者对许多问题并不能清楚回答。这些问题包含：

- 在有监督、半监督，和无监督学习中，香农互信息和估计互信息之间的关系是怎样的？
- 相似度是概率吗？如果它是，为什么它们不是归一化的？如果不是，为什么它能带进贝叶斯公式，用作概率预测(比如用于像是贝叶斯公式的公式(9))？
- 我们能直接从样本或样本分布得到相似函数，失真函数，真值函数或隶属函数？

而 G 理论可以清楚地回答这些问题。

第三，由于对估计互信息和香农互信息之间的关系认识不够清楚，有些深度学习作者认为【4, 24】: 最大化香农互信息是很好目标；因为估计互信息是香农互信息的下界，我们可以通过最大化估计互信息来最大化香农互信息。然而, Tschannen 等人【17】提出不同看法，认为用互信息作为目标函数是不够的，在有些情况下甚至是不好的；最近的深度学习成功另有原因。笔者认为，他们的质疑值得重视。从语义信息 G 理论的角度看，估计互信息类似于效用，香农互信息类似于成本；最大估计互信息才是目的。有些情况下两者需要同时最大化。而有些情况下，为了提高通信效率，我们需要减少香农互信息。

Tishby 等人【26,27】用信息瓶颈解释深度学习。根据他们的理论，对于结构为 X->T->Y 的神经网络，我们需要使 $I(Y; T)-\beta I(X; T)$ 最大化【26】或使 $I(X; T)-\beta I(Y; T)$最小化【27】。其中 β 是拉格朗日乘子，控制 $I(Y; T)$最大化和 $I(X; T)$最小化之间的权衡。这种解释涉及率失真函数 $R(D)$，很有启发。 然而，在这一理论中，EMI(或参数化互信息)和 ShMI 并没有很好地区分，忽略了使它们彼此靠近是一项重要任务。因此，本文(第 5.2 节)将提供一种有竞争力的解释——使用 $R(G)$函数。

尽管深度学习中存在各种问题，深度神经网络(DNN)有强有力的特征抽取能力，结合估计信息测度从而提供了有说服力的应用。对比之下，语义信息论的机器学习应用比较简单，没有结合 DNN。从实用的角度看, G 理论的应用缺少说服力。各种信息理论(包括 G 理论)也不足以解释深度学习中许多巧妙方法。

现在是到了合并两股潮流，让它们相互学习相互取长补短的时候了！

推广的熵有很多种，然而本文涉及的广义熵，不改变香农熵的结构，仅仅改变 log 右边的函数。除了经典信息理论的学习【28,29】中的交叉熵方法，研究者也提出其他熵，比如 Reny 熵【30,31】, Tsallis 熵【32,33】, correntropy【34】，基于矩阵的熵【35】，以及相关的方法用于广义信息理论的学习。 图 1 显示了经典信息理论的学习、广义的信息理论的学习和语义信息理论的学习之间的区别和联系。

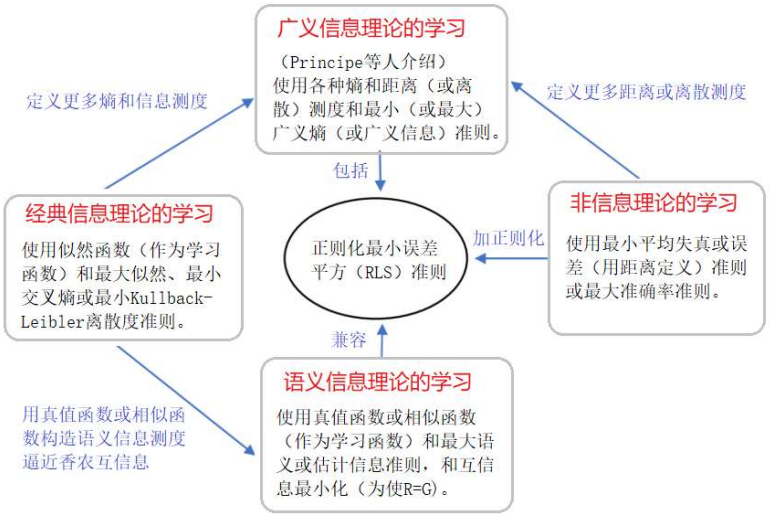


图 1.四种学习之间的区别和联系

语义信息理论的学习和非信息理论的学习之间的主要区别是，后者使用真或正确率作为评价准则，而前者使用逼真度(Truthlikeness【36】),即向真的成功逼近或其对数作为评价准则。真独立于事实的先验概率分布 $P(x)$，而逼真度和 $P(x)$相关。两个准则应有它们各自的用途。语义信息理论的学习和广义信息理论的学习的主要差别是：前者仅仅用真值函数或相似函数推广香农熵和信息。

虽然很多方法使用参数或 SoftMax 函数构造互信息，但是如果没有使用相似函数或真值函数，还不能说是语义信息理论的学习方法。这些熵和信息测度也有其应用价值。但不是本文要讨论的。

除了上面提到的语义信息测度，还有其他人（比如 Floridi【31】和钟义信【32】）提出的语义信息测度。但是作者没有看到这些测度的机器学习应用。

本文主要目的：

- 回顾语义信息测度发展史和学习函数进化史；
- 澄清语义互信息和香农互信息之间的关系；
- 促进语义信息论和深度学习的融合和发展。

本文后面章节内容是：

- 回顾语义信息测度的进化史；
- 回顾学习函数的进化史；
- 介绍笔者建立的语义信息 G 理论及机器学习应用；
- 讨论深度学习中语义信息和香农信息最大化问题，以及和深度学习相关的一些问题；
- 讨论几个值得探索的新方法以及 G 理论的局限性；
- 结论——包括机遇和挑战。

2. 语义信息测度的进化

2.1. 使用真值函数或相似函数逼近香农互信息

香农互信息公式是：

$$\begin{aligned} I(X;Y) &= \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \\ &= H(Y) - H(Y \mid X) = H(X) - H(X \mid Y). \end{aligned} \tag{1}$$

其中 x_i 是一个实例，y_j 是一个标签，X 和 Y 是两个随机变量。

为了说明估计互信息，我们先看香农互信息的核心部分：

$$I(x; y_j) = \log \frac{P(x, y_j)}{P(x)P(y_j)} \quad , \tag{2}$$

其中 x_i 是变量，y_j 是常量。深度学习研究者发现，我们能用参数构造相似函数，然后用相似函数构造逼近香农互信息的估计互信息。相

似函数的主要特点是：

$$S(x;y_j) \propto m(x,y_j) = \frac{P(x,y_j)}{P(x)P(y_j)} \propto P(y_j \mid x), \tag{3}$$

上式也定义了 $m(x, y_j)$。它是一个重要函数，但是作者没见过它的名字。我们称之为关联函数(relatedness function)。我们也可以用参数构造失真函数 $d(x, y)$，用 $d(x, y)$表示相似函数，使得

$$S(x;y_j) = \exp[-d(x,y_j)] \propto m(x,y_j). \tag{4}$$

然后我们可以用 $S(x, y_j)$或 $\exp(-(d(x, y_j))$表示估计信息:

$$I_\theta(xk;yk) = \log\frac{S(xk,yk)}{\sum_{l=1}^{N} S(xl,yl)} = \log\frac{\exp(-d(xk,yk))}{\sum_{l=1}^{N}\exp(-d(xl,yl))}, \tag{5}$$

其中 k 表示一个样本$\{(x_k, y_k) \mid k=1,2,\ldots,N\}$中第 k 个样例的序号。估计互信息被表示为：

$$I_\theta(X;Y) = \frac{1}{N}\sum_k I_\theta(xk;yk). \tag{6}$$

而信息论研究者通常用样本分布 $P(x, y)$求平均。估计信息和估计互信息被表达为

$$\begin{aligned} I_\theta(x_i;y_j) &= \log\frac{S(x_i,y_j)}{\sum_l P(x_l)S(x_l,y_j)} \\ &= \log\frac{\exp(-d(x_i,y_j))}{\sum_l P(x_l)\exp(-d(x_l,y_j))}, \end{aligned} \tag{7}$$

$$I_\theta(X;Y) = \sum_j\sum_j P(x_i,y_j) I_\theta(x_i;y_j). \tag{8}$$

注意划分函数中多了 $P(x_i)$。但是(7)和(5)是等价的。和似然函数 $P(x \mid \theta_j)$比（θ_j表示 y_j和有关的参数），相似函数独立于信源 $P(x)$。当信源和信道改变之后，相似函数作为预测模型仍然可用。它和新的信源 $P'(x)$一起形成新的模型预测或似然函数:

$$P'(x \mid \theta_j) = \frac{P'(x)S(x,y_j)}{\sum_l P'(x_l)S(x_l,y_j)}. \tag{9}$$

公式(5)和(6)适合样本尺寸不太大的情况；而公式(7)和(8)适合样于样本尺寸远大于实例种类的情况。但是，后一种表达可以反映信源的变化，更加便于澄清语义互信息和香农互信息之间关系。

2.2. 涉及语义信息测度的历史事件

作者尽自己所知，列出和语义信息测度有关的重大历史事件。

1931 年, Popper 在《科学发现的逻辑》【39】 (P.96)一书中提出：科学假设的逻辑概率越小，如果经得起检验, (语义)信息量就越大。可以说, Popper 是语义信息最早研究者(如 Floridi 所言【19】)。随后他还提出逻辑概率公理系统。他强调有两种概率 (统计概率和逻辑) 同时存在([39] (pp. 252-258)。但是他没有建立同时包含两者的概率系统。

1948 年香农发表了著名的文章：《通信的数学理论》【3】。

香农信息论为优化通信编码提供了强大的指导思想和工具。然而，香农只用了统计概率——基于频率解释。文中提出信息率保真函数，后来被改称为信息率失真函数 $R(D)$ (R 是给定平均失真上限 D 时的最小互信息)。

1949, Weaver 和香农合作出版了《The Mathematical Theory of Communication》【9】其中包含香农已发表的著名文章和 Weaver 的文章，后者标题是 "Recent Contributions to The Mathematical Theory of Communication"。Weaver 在这篇文章里提出通信的三个水平：

水平 A. 通信符号怎样被准确传递？

水平 B. 传递的符号怎样精确地传送想要表达的含义？

水平 C. 接收的含义怎样有效地以希望的方式影响实施？

水平 A 只涉及 Shannon 信息，水平 B 涉及香农信息和语义信息，水平 C 还涉及信息效用或价值。

上个世纪，很多人拒绝研究语义信息，主要理由是香农说过这样的话【9】(P. 3)：

"通常，消息含有意义，也就是说，它们指代概念实体或(根据某些系统)与某些物理或概念实体相关联。而通信的这些语义方面与工程问题无关。"

实际上，香农只是限制其理论应用范围，并不是反对研究语义信息。如果他反对，怎么会同意把自己的文章和 Weaver 的那篇文章放一起出版呢？最近十年来，随着人工智能和互联网的发展，社会对语义通信的需求越来越迫切，越来越多学者开始研究语义通信【10,11】，包括语义信息论。

1951 年, Kullback 和 Leibler【40】提出 Kullback-Leibler (KL) 离散度(divergence)，又叫 KL 信息或相对熵：

$$D_{\mathrm{KL}} = H(P \| Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}. \tag{10}$$

KL 信息可看做香农互信息在给定 $Y = y_j$ 时的特例，但是香农互信息也可以看成是两个分布 $P(x, y)$和 $P(x)P(y)$之间的 KL 信息。

因为香农信息不能度量语义信息, 1952 年, Carnap (卡尔纳普) 和 Bar-Hillel 【41】 提出一个语义信息公式：

$$I_{\mathrm{CB}} = \log[1/m_p]. \tag{11}$$

其中 m_p 是一个命题 p 的逻辑概率。这个公式反映了 Popper 的部分思想——逻辑概率越小，信息量越大。但是, I_{CB} 并不能反映假设是否经得

起检验，所以这一公式是不实用的。另外，逻辑概率 m_p 是根据假设的外延包含的被划分论域的子论域的相对数目定义的，和实例的先验概率无关，这也是不合理的。

1957 年, Shepard【41】提出用距离表示相似度: $S(x, y)=\exp(-d(x, y))$。

1959 年，香农【43】提供了一个无记忆二元信源的的信息率失真函数 $R(D)$的表达式和一般无记忆信源的 $R(D)$的参数解：

$$\begin{aligned} D(s) &= \sum_i \sum_j d_{ij} P(x_i) P(y_j) \exp(s d_{ij}) / Z_i, \\ R(s) &= sD(s) - \sum_i P(x_i) \ln Z_i, \end{aligned} \tag{12}$$

其中 $d_{ij} = d(x_i, y_j)$, $s \le 0$, Z_i 是划分函数。在笔者看来, $\exp(sd_{ij})]$ 就是真值函数，限失真时的最小互信息可以看做是语义互信息[18]。

1967 年,Theil【44】提出广义 KL 信息公式：

$$I_{Theil} = \sum_i P(x_i) \log \frac{R(x_i)}{Q(x_i)}, \tag{13}$$

其中包含三条函数曲线。$R(x)$ 可看做是对 $P(x)$的预测, $Q(x)$是先验概率分布。这样, I_{Theil} 就表示预测信息。

1965, Zadeh 提出模糊集合论【45】。一个实例 x_i 在模糊集合 A_j 中的隶属度被表示为 $M_{Aj}(x_i)$。本文作者于 1993 年解释【15】：隶属函数 $M_{Aj}(x)$就是一个谓词 y_j="x 在 A_j 中"的真值函数；如果我们假设存在一个典型 x_j(即柏拉图的理念，它可能不在 A_j 中)使 y_j 为真，即 $M_{Aj}(x_j)=1$, 那么，隶属度 $A_j(x_i)$就是 x_i 和 x_j 之间的相似度。

1972 年, De Luca 和 Termini【46】用单个隶属函数定义模糊熵，我们简称它为 DT 模糊熵。

1974 年, Akaike【47】把模型参数带进 KL 公式，证明最大似然估计等价于最小 KL 信息估计。虽然他没有明确用交叉熵，但是所用的对数似然度等于负的交叉熵乘以样本尺寸：

$$\log P(\mathbf{X} \mid \theta) = N \sum_i P(x_i) \log P(x_i \mid \theta) = -NH(X \mid \theta). \tag{14}$$

1981 年, Thomos 举例说明【48】，我们可以把隶属函数带进贝叶斯公式，做概率预测。根据 Dubois 和 Prade 的文章 [49]，也有其他人差不多同时提出这样的公式。

1983 年, Donsker 和 Varadhan【50】提出：KL 信息可表示为：

$$I_{\text{DV}} = \sum_i P(x_i) T(x_i) - \log \sum_k Q(x_k) \exp[T(x_k)] \le H(P \| Q) = D_{KL}. \tag{15}$$

I_{DV} 后来被称之为 Donsker-Varadhan 表示【1】，我们且称之为 DV-KL 信息。他们提出这一公式可能是因为受到香农的信息率失真函数或 Jibbs 分布或 Boltzmann-Jibbs 分布的启发。为了理解这一公式，我们把 $P(x)$换成 $P(x \mid y_j)$，把 $Q(x)$换成 $P(x)$。于是 KL 信息就变成：

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

为表示 KL 信息，我们只需要找到 $T(x_i)$使得 $\exp(T(x_i)]\propto P(yj|x)$或 $\exp[(T(x_i)]\propto P(y_j|x)/P(y_j)$。DV-KL 信息后来发展为互信息神经估计 MINE【1】。但是，指数或负指数函数一般是对称的，而 $P(y_j|x)$未必是对称的，要使两者相等是不容易的。

1983 年，汪培庄【51】提出隶属度的统计解释——即随机集落影理论。用机器学习的语言来说，我们可以用一个样本，其中每个样例包含一个标签和数量不等的相似实例，即 $S=\{(x_{t1}, x_{t2}, \ldots; y_t) \mid t=1,2,\ldots,N)\}$, x_i 在含有标签 y_j 的样例中的概率就是相应的隶属度 $M_{Aj}(x_i)$。作者后来把随机集落影发展为随机点落影，从样本分布得到隶属函数【17】。

1986 年, Aczel 和 Forte【52】提出广义熵，现在叫做交叉熵，其中给出不等式：

$$-\sum_i P_i\log Q_i\geq-\sum_i P_i\log P_i=H(P), \tag{17}$$

并且通过 $\ln x\leq x-1$证明了上式。他们解释，这一公式也反映香农无失真编码定理——香农熵是平均码长下限。

也是 1986 年, Zadeh 定义了模糊事件的概率【53】，它等于隶属函数的均值。本文将说明，这一概率就是我们常用的逻辑概率和划分函数。

1990 年，笔者提出：同时用真值函数和逻辑概率表达语义信息量【14】，这样可以反映假设是否经得起检验，克服 Carnap-Bar-Hillel 语义信息公式的缺点。为了区别于其他语义信息测度，作者后来称之为语义信息 G 测度或 G 测度。公式是：

$$I_{LU}(x;y_j)=\log\frac{Q(A_j\mid x)}{Q(A_j)}, \tag{18}$$

其中 A_j 是模糊集合, y_j= “x 属于 A_j”是一假设, $Q(A_j|x)$是 A_j 的隶属函数，也是 y_j 的（模糊）真值函数。$Q(A_j)$是 y_j 的逻辑概率，也是 Zadeh 提出的模糊事件的概率。这一公式就能反映假设是否经得起检验，因而反映 Popper 思想。如果总有 $Q(A_j|x)=1$，上面语义信息就变成 Carnap 和 Bar-Hillel 的语义信息。

其实作者一开始就是为了度量色觉信息而推广香农信息公式的。笔者曾提出色觉的译码模型【54,55】，其中互补彩色是三对：红绿，蓝黄，绿绛， 而不像流行的阶段模型——其中彩色是两对。为了辩护译码模(证明它分辨率更高，提供信息更多)，他希望用信息公式度量色觉信息。为此他把色觉看做是对颜色的估计，即 $y_j=\hat{x}_j$，并引入分辨率函数，即相似函数，得到度量色觉的语义信息公式。后来他才发现把分辨率函数换成真值函数，我们也能度量自然语言的信息。因为用统计概率求平均，这一信息测度能够反映：错误的概率预测会减少语义信息。

对 $I_{Lu}(x;y_j)$求平均，就得到广义 Kullback-Leibeler(KL)信息(下式中间部分)， 如 Theil 的广义 KL 信息，它等于语义 KL 信息(下式右边部分)， 即：

$$I_{LU}(X;y_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i\mid A_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{Q(A_j\mid x_i)}{Q(A_j)}. \tag{19}$$

上面公式用到带隶属函数的贝叶斯公式（详见 4.1 节）。不过作者当时并不知道 Thomas 等人的研究。如果 $Q(A_j|x)$用负的损失函数表示，语义 KL 信息就变成 DV-KL 信息：

$$I_{LU}(X;y_j)=-\sum_i P(x_i\mid y_j)d(x_i,y_j)+\log[1/Q(A_j)]. \tag{20}$$

不过作者后来才知道 Theil 的公式，最近才知道 DM-KL 信息。但是，并不是任何 KL 信息或语义 KL 信息都能用指数（或负指数)函数表示为 DV-KL 信息。比如 y_j=“老年人”的真值函数是不对称的，需要用 Logistic 函数表示。所以, DV-KL 信息只是语义 KL 信息的特例。

1993 年，笔者推广香农的信息率失真函数 $R(D)$【15】，即用 $I_{LU}(x;y)$代替失真函数 $d(x, y)$,得到信息率保真函数 $R(G)$ (G 表示语义互信息下限)。$R(G)$函数揭示了香农互信息 R 和语义互信息 G 之间的匹配关系。他同时把真值矩阵解释为广义信道，提出两个信道相互匹配的概念。同时他研究了如何根据颜色的视觉分辨率和 $R(G)$函数压缩图像数据【15,16】。

2017 年，笔者提出 P-T 概率框架【12】，用 T 代替 Q，用 θ_j 代替 A_j。2020 年，他讨论了 P-T 概率框架和 Popper 理论的关系【17】。θ 不仅代表模糊集合，也代表模型参数。使用 P-T 概率框架，我们就能方便地把 G 测度用于机器学习。语义互信息和几个广义熵的关系是：

$$I(X;Y_\theta)=\sum_i\sum_j P(x_i)P(y_j\mid x_i)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{21}$$

$$=H(Y_\theta)-H(Y_\theta\mid X)=H(X)=H(X\mid Y_\theta),$$

其中三种广义熵是：

$$H(X\mid Y_\theta)=-\sum_j\sum_i P(x_i,y_j)\log P(x_i\mid\theta_j), \tag{22}$$

$$H(Y_\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i), \tag{23}$$

$$H(Y_\theta)=-\sum_j P(y_j)\log T(\theta_j). \tag{24}$$

$H(X|Y_\theta)$可谓预测熵，也是交叉熵。关于另外两个广义熵，作者建议称 $H(Y_\theta|X)$为模糊熵，称 $H(Y_\theta)$为覆盖熵（详见 5.4 节讨论）。

和 DT 模糊熵比, $H(Y_\theta|X)$作为模糊熵更有普遍性，因为如果标签由 n 个减少为互补的两个 y_1 和 y_0，并且 $P(y_1|x)$和 $P(y_0|x)$等于两个隶属函数 $T(\theta_1|x)$和 $T(\theta_0|x)$，上面模糊熵就退化为 DT 模糊熵。更多模糊熵和模糊信息测度见【56】。作者认为有些用模糊熵构造的语义信息测度是不妥当的，主要原因它们不能反映假设是否经得起检验。

2021 年，作者用-log$T(\theta_j|x)$定义失真函数 $d(x, y_j)$，即令 $\overline{d}$ 等于模糊熵 $H(Y_\theta|X)$(D 等于其上限)，从而推广信息率失真函数和最大熵方法【18】。这样，约束条件就更加灵活。

最近几年，深度学习研究者用特征矢量的点积表示相似函数，把相似函数带进语义或估计信息公式，使语义信息测度有了更丰富更实用的形式。详见下节。

3. 学习函数的进化

3.1 从似然函数到相似函数或真值函数

本节回顾学习函数和优化准则，以便读者理解为什么我们最终选择相似函数和估计信息测度。

机器学习的任务是，先用样本优化学习函数，再用学习函数做概率预测或分类。优化学习函数有多种准则：比如最大正确率，最大似然度，最大互信息，和最大估计互信息，最小失真，最小平方误差，正则化最小平方误差(RLS)，最小 KL 距离，最小交叉熵…准则。 有些准则是等价的，比如最大似然，最小交叉熵，最大估计互信息准则是等价的。有些准则是类似的或兼容的，比如 RLS 准则和最小 EMI 准则。

不同的学习函数逼近下面不同的概率分布:

- $P(x)$: 实例 x 的先验分布，反映信源；逼近它的是 $P_\theta(x)$(或 $P(x|\theta)$);
- $P(y)$: 标签 y 的概率分布，反映信宿；逼近它的是 $P_\theta(y_j)$;
- $P(x|y_j)$: x 的后验分布，逼近它的学习函数是似然函数 $P(x|\theta_j)$或 $P(x|y_j, \theta)$；
- $P(y_j|x)$: 可谓转移概率函数，逼近它的学习函数是 $P(\theta_j|x)$——也是 Fisher 提出的, Fisher 称之为反概率【57,58】。我们后面称之为反概率函数。可用 Logistic 函数表示它。
- $P(y|x_j)$: y 的后验概率分布。因为 $P(y|x)=P(y)P(x|y)/P(x_i)$，贝叶斯推断用 $P(\theta)P(x|y, \theta)/P_\theta(x)$【58】 (贝叶斯后验) 逼近它。
- $P(x, y)$: 联合概率分布，逼近它的是似然函数 $P(x, y|\theta)$.
- $m(x, y_j)=P(x, y_j)/[P(x)P(y_j)]=P(x|y_j)/P(x)=P(yj|x)/P(y_j)$：关联 (度) 函数，逼近它的函数我们记为 $m_\theta(x, y_j)$，并称之为逼真 (度) 函数 (truthlikeness function)。它在 0 和∞之间变化。
- $m(x, y_j)/\max(m(x, y_j))$，我们用真值函数 $T(\theta_j|x)$或相似函数 $S(x, y_j)$逼近它。真值函数是逼真函数的纵向归一化——其最大值是 1。

作者以为，学习函数从似然函数进化到逼真函数和真值函数(或相似函数)，优化准则也从最大似然或交叉熵准则进化到最大语义(或估计) 信息准则。虽然两者等价，但表达形式不同，含义不同。下面回顾学习函数进化的原因。

因为样本分布 $P(x|y_j)$可能是间断的或不平滑的，我们不能用它做概率预测。为此, Fisher【57】提出用似然函数 $P(x|\theta_j)$逼近它——使用最大似然准则。然后我们用似然函数做概率预测或分类。

最大似然方法的缺点是，当 $P(x)$改变后，似然函数就需要重新学习。另外, $P(x|y_j)$常常是不规则的，很难用一个函数逼近。因此，反概率函数 $P(\theta_j|x)$得到采用。使用这一函数，当 $P(x)$变成 $P'(x)$的时候，我们可以通过贝叶斯公式能得到新的似然函数 $P'(x|\theta_j)$。另外，我们也可以利用 $P(\theta_j|x)$分类 (使用最大正确率准则)

Logistic 函数就是典型的反概率函数。当 y 只有两种 y_1 和其否定 y_0 时，我们可以用一对带参数的 Logistic 函数逼近 $P(y_1|x)$和 $P(y_0|x)$。但是它有两个缺点：

1)当 y 的数目 $n>2$ 时，构造反概率函数是困难的，因为概率分布

是归一化的 (对于每个 x)，即：

$$\sum_j P(y_{\theta j} \mid x_i) = 1, \quad \text{for } i = 1, 2, \ldots \tag{25}$$

因为上式限制，把反概率函数用于多标签学习是困难的。多标签学习的一个权宜的方法是二元关联(Binary Relevance) 【59】，即为每个标签搜集正例和反例，把 n 个多标签学习转换为 n 对标签的单独学习。但是为 n 个标签准备 n 个样本，这是不经济的，特别是用于图像和文本学习时。另外，每个标签单独优化和多个标签放在一起优化，结果是不一样的。

2) 虽然对于二元通信，$P(y_j|x_i, i=j)$反映正确率。但是，当 $n>2$ 时，$P(y_j|x_i, i=j)$并不能反映正确率，特别是对于语义通信。比如，x 表示年龄，y 表示三个标签中的一个：y_0 = ”非成年人”，y_1=”成年人”，y_2=”年轻人”。如果 y_2 很少使用，即 $P(y_2)$和 $P(y_2|x)$就很小。但是，x=20 岁时，用标签 y_2 的正确率应该是 1。这个正确率应该和 $P(x)$及 $P(y)$无关，是语义决定的。

$P(x, y|\theta)$用作学习函数也很常见，比如在 RBM 中【60】和许多信息理论学习中【29】。但是它也有 $P(x|\theta_j)$的缺点：不适合 $P(x)$不规则并且多变的场合。贝叶斯后验也不合适。

所以，我们需要一个学习函数，它正比于 $P(y_j|x)$并且独立于 $P(x)$和 $P(y)$。真值函数和相似函数就是这样的函数。

我们用 GPS 定位作为例子 (参看图 2) 说明为什么我们要用真值 (或相似) 函数 $T(\theta_j|x)$作为学习函数，而不是用 $P(x|\theta_j)$或 $P(x, y|\theta)$。一个 GPS 指针表示一个估计 $y_j = \hat{x}_j$，真实的位置可能有些不同。这时，x 和$\hat{x}_j$ 之间的相似度就是 x 发生时$\hat{x}_j$ 的真值。这也解释了为什么相似度常常被称之为语义相似度【22】。

图 2 显示了一个真实截图，其中虚线、红星、字母是添加的。棕色线表示铁路，GPS 设备在列车上。y_j 并不指在铁路上是因为它不是很准。

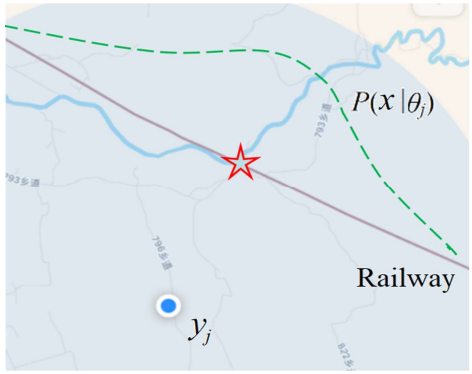


图 1. 图解 GPS 定位带有偏差。我们结合 y_j 和先验知识 $P(x)$做概率预测；红星表示的是最可能位置

GPS 指针的语义可以表达为

$$T(\theta_j|x) = S(x, x_j) = \exp[-(x-x_j)^2/(2\sigma^2)], \tag{26}$$

这里 x_j 是 y_j 指的位置，σ 是平均误差平方的开方 (即 RMS 或标准偏差) 。为了方便，公式中假设 x 是一维的。根据公式 (9)，我们可

以预测：红星代表最为可能的位置。

但是，如果我们使用 $P(x|\theta_j)$ 或 $P(x, y|\theta_j)$作为学习函数和预测模型，$P(x)$变化时，模型就不能利用变化的 $P(x)$，因而缺少可迁移性。另外，使用真值函数，我们也可以从样本学到 GPS 指针的系统偏差和精度【13】。而用 $P(x|\theta_j)$或 $P(x, y|\theta_j)$不行，因为系统偏差和精度独立于 $P(x)$和 $P(y)$。考虑带有 GPS 地图的小车在公路上，$P(x)$更加复杂多变，$P(x|\theta_j)$也是。但是 $T(\theta_j|x)$ 可以是不变的高斯函数。

上面例子解释了为什么深度学习研究者提出用相似函数作为学习函数的原因。4.2 节将证明：相似函数就是逼近 $m(x, y_j)/\max[m(x, y_j)]$的函数。

3.2. 语义相似度定义

本文相似度指语义相似度【22】，比如 19 岁和 20 岁的人年龄相似，雪和水相似，message 和 information 相似。但是不包括仅仅由于偶然原因的外表相似。比如雪和盐之间和 “message”和“massage”之间的相似性。

语义相似度可以在两个感觉的，几何的，生物学的，语言的，…对象之间。最常见的相似度是使用距离定义的。距离也反映失真或误差。继 Shepard【42】, Ashby 和 Perrin 进一步讨论了相似度、距离和感觉及相关的概率预测【61】。图像和生物特征之间的相似性通常用两个特征矢量的点积或余弦表示【2,4,6】。

单词、语句和文本之间的语义相似度研究较多【22】。有人利用语义结构定义相似度，比如：1995 年，Reslik【22】提出使用 WordNet 的结构定义其中任意两个单词之间的相似度，后来其他人提出了改进形式【62】。也有人用统计数据定义相似度，比如，在隐含语义分析中【64】，研究者使用来自统计的矢量表示相似度；也有人使用点互信息【65】(Pointwise Munital Information，即 $I(x_i; y_j)$，见等式(2))作为相似度。因为这一相似度在-∞和∞之间变化，有人提出改进的 PMI 相似度——在 0 和 1 之间变化【65】。还有人通过定义语义距离定义语义相似度。更多语义相似度和语义距离测度见【66】。

有人认为语义相似性包含相关性(relatedness)，有人认为两者不同。比如汉堡和热狗相似，而汉堡和炸薯条相关。关于语义相似性是否包含相关性，存在争议【67】。作者认为，因为相关性反映时空位置相似性，认为相关的两者是相似的，是实用的，也是合理的。

作者提出的真值函数也可以被理解为相似函数（x_i 和典型 x_j 相似）。我们可以用距离或失真定义它，然后用样本分布优化它。作者用 $T(\theta_j|x)$逼近 $m(x, y_j)/\max[m(x, y_j)]$，当样本很大时，两者会相等。

相似函数 $S(x, y)$可以直接作为学习函数，产生 x 的后验概率分布。我们也可以使用这样的变换 $\exp[kS(x, y)]$作为新的相似函数或真值函数【2,6】（其中 k 是参数，可以放大或缩小相似范围）。对于那些适合放在 exp( )函数中的相似函数，比如矢量的点积或余弦，我们最好把它们理解为保真度，记为 $f_d(x, y_j)$。假设最大可能的保真度是 $f_{d\max}$，我们用它定义失真函数：$d(x, y_j) = k[f_{d\max} - f_d(x, y_j)]$，然后用

exp[-$d(x, y_j)$]作为相似函数。不过，用不用这一转换，SoftMax 函数或逼真函数的值是不变的。

语义相似度研究也提示我们，度量语义信息（包括感觉信息）要用相似函数。机器学习中，重要的是怎样用参数构造相似函数，然后用样本或样本分布优化模型参数。

3.3. 相似函数和语义信息测度用于深度学习

1985， Achley, Hinton,和 Sejnowski 提出 Boltzmann 机【68】，后来他们又提出受限 Boltzmann 机(RBM)【60】。其中用到 SoftMax 函数，用以产生 Gibbs 分布，即预测模型。2006 年，Hinton 等人用 RBM 和反向传播方法构成自动编码器(Autoencoder)【69】和深度信念网(DBN)【70】，取得了极大成功。然而，这是信息理论的学习方法，还不算是语义信息论的学习方法，因为其中 SoftMax 函数逼近的是 $P(x, y)$，而不是用相似函数构成的逼真度。不过，SoftMax 函数的使用给使用相似函数的应用提供了示范。

2010 年，Gutmann 和 Hyvärinen【71】提出噪声对比学习，取得了很好效果。其中使用并不归一化的函数和划分函数。在笔者看来，就是已知 $P(x|y_j)$，而想得到 $P(\theta_j|x)$。因为 $P(y_j|x)/P(y_j) = P(x|y_j) / P(x)$，没有 $P(x)$就通过噪声制造反例，然后合并出 $P(x)$。这样就能优化 $P(\theta_j|x)$。然后，如同用 Binary Relevance 方法，把 N 个标签分类转化成 N 对标签的二分类。

2016，Sohn【72】创造性地提出距离测度学习(Distance Metric Learning)，也使用了 SoftMax 函数，虽然没有明说学习函数正比于 $m(x, y)$，但是距离显然独立于 $P(x)$和 $P(y)$.

2017 年，本文作者得到结论【12】：当 $T(\theta_j|x) \propto P(x|y_j)/P(x)=m(x, y)$或 $T(\theta_j|x) \propto P(y_j|x)$时，语义互信息最大，等于香农互信息。$T(\theta_j|x)$是参数化的 $m(x, y)$的纵向归一化。当样本很大时，我们可以从样本分布得到优化的真值函数：

$$
\begin{aligned}
T(\theta_j \mid x) &= \frac{m_\theta(x, y_j)}{\max(m_\theta(x, y_j))} = \frac{m(x, y_j)}{\max(m(x, y_j))} \\
&= \frac{P(x, y_j)}{P(x)P(y_j)} \Big/ \max\left(\frac{P(x, y_j)}{P(x)P(y_j)}\right) = \frac{P(y_j \mid x)}{\max(P(y_j \mid x))}.
\end{aligned}
\tag{27}
$$

作者同时提出一组真值函数构成一个语义信道$\{T(\theta_j|x), j=1,2,\ldots\}$(以前称之为广义信道【15】）。在【12,13】中，他发展出一组信道匹配算法(包括迭代算法)，用以求解多标签分类，不可见实例最大互信息分类和混合模型。他也证明了混合模型的收敛【73】，推导出新的贝叶斯确证测度【74】和因果确证测度【75】。但是作者没有把语义信息方法应用到神经网络。

2018 年 1 月，Belghazi 等人(包括 Bengio)【1】发表文章：MINE: Mutual Information Neural Estimation(互信息神经估计)，这一方法取得很好结果。MINE 使用推广的 DV-KL 信息测度，其中实际上使用了

学习函数:

$$\exp(T_w(x, y_j)) \propto P(y_j \mid x). \tag{28}$$

虽然 $T_w(x, y_j)$不是负的，但是可理解为保真度函数。

2018 年 7 月， Oord 等人【2】发表文章："Representation Learning with Contrastive Predictive Coding"，明确提出：使用一个学习函数——希望它正比于 $m(x, y)=P(x|y)/P(x)$。文中表示是这样的：

$$f_k(x_{t+k}, c_t) \propto P(x_{t+k} \mid c_t) / P(x_{t+k}), \tag{29}$$

其中 c_t 是从前面的数据得到的特征矢量，x_{t+k} 是要预测的特征矢量。文中 $f_k(x_{t+k}, c_t)$) 就是根据 c_t 预测的 x_{t+k} 和实际的 x_{t+k} 之间的相似函数，用它构造估计信息——这种方法被称之为信息噪声对比估计 InfoNCE。对比学习中的 N 对 Logistic 函数变成 N 个 SoftMax 函数，因而可以直接用于多标签学习。其中 $f_k(x_{t+k}, c_t)$不限于指数或负指数函数——和 MINE 中学习函数不同。笔者认为，用一个像隶属函数那样的函数作为相似函数会更加灵活。

MINE、InfoNCE 和作者的语义信息方法在本质上是一样的，这类语义互信息公式的主要特点是：

- 用正比于 $m(x, y_j)$的函数做学习函数，记为 $S(x, y_j)$，它的最大值一般是 1，它的平均就是划分函数 Z_j。
- x 和 y_j 之间的语义信息就是 $\log[S(x, y_j)/Z_j]$。
- 求平均时仍然用统计概率分布 $P(x|y_j)$或 $P(x, y)$。
- 互信息可用覆盖熵减去模糊熵表示，而模糊熵等于平均失真。

4. 语义信息 G 理论及机器学习应用

4.1. P-T 概率框架和语义信息 G 测度

在 P-T 概率框架中，P 表示统计概率(包括主观概率，比如似然度)，T 表示逻辑概率。我们定义该框架如下。

令 X 是一随机变量，表示一个实例，取值 $x \in U=\{x_1,x_2,\ldots\}$；$Y$ 是一随机变量，表示一个标签或假设，取值 $y \in V=\{y_1,y_2,\ldots\}$。香农信道可用一组转移概率函数 $P(y_j|x)$ (j=1,2,…)表示；而语义信道可以用一组真值函数 $T(y_j|x)$ (j=1,2,…)表示。

设 U 中使 y_j 为真的元素构成一个模糊子集 θ_j(即 y_j="x 在 θ_j 中")，则 θ_j 中的隶属度(记为 $T(\theta_j|x)$)就是命题函数 $y_j(x)$的真值函数 $T(y_j|x)$，即 $T(\theta_j|x)=T(y_j|x)$。y_j 的逻辑概率就是 Zadeh 定义的模糊事件的概率【45】:

$$T(y_j)=T(\theta_j)=\sum_i P(x_i)T(\theta_j\mid x_i). \tag{30}$$

y_j 为真时，x 发生的概率是：

$$P(x\mid\theta_j)=P(x)T(\theta_j\mid x)/T(\theta_j), \tag{31}$$

θ_j 也可以理解为模型参数，这样 $P(x|\theta_j)$就是似然函数。

对于估计，$V=U$, $y_j=\hat{x}_j$="x 大约是 x_j"。如果我们假设相应每个模糊集合 θ_j 存在一个典型或柏拉图的理念 x_j (它可能不在 θ_j 中)，那么，隶属度和真值 $T(y_j|x_i)=T(\theta_j|x_i)$也可以理解为 x_i 和 x_j 之间的相似度。

给 $P(x)$和 $P(x|\theta_j)$，令 $T(\theta_j|x)$的最大值是 1，我们能推导出【13】：

$$T(\theta_j\mid x)=\frac{P(x\mid\theta_j)T(\theta_j)}{P(x)},\ T(\theta_j)=1/\max[P(x\mid\theta_j)/P(x)]. \tag{32}$$

逻辑概率和统计概率的主要区别是：统计概率是横向归一化的(总和是 1)，而逻辑概率是纵向归一化的 (最大值是 1) 。一般情况下，每个 $T(\theta_j|x)$的最大值是 1(对于不同的 x)，并且 $T(\theta_0)+T(\theta_1)+\ldots>1$；而 $P(y_0|x)+P(y_1|x)+\ldots=1$ (对于给定的 x)。

y_j 为真提供关于 x_i 的语义信息是：

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{33}$$

语义信息量能反映假设是否经得起检验，其图解见图 3。

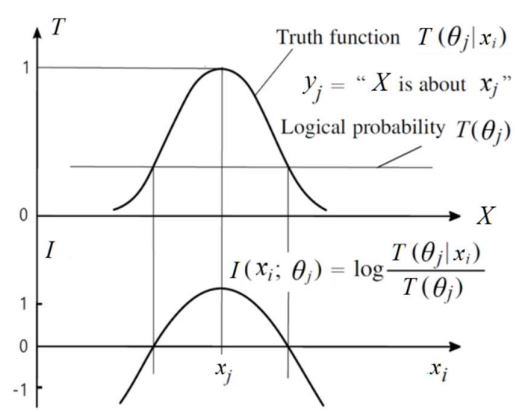


图 3. 语义信息量图解。y_j 传递关于 x_i 的语义信息随偏差增大而减少，偏差越大，消息越少，甚至是负的。

对不同 x 求平均就得到语义 KL 信息 (见公式 (19))，再对不同的 y 求平均就得到语义互信息 $I(X;Y_\theta)$ (见公式 (21))。

4.2. 真值函数的优化和概率预测

就像 $P(y_j|x)(j=1,2,\ldots)$构成一个香农信道，$T(\theta_j|x)(j=1,2,\ldots)$构成一个语义信道。当香农信道匹配语义信道，即 $T(\theta_j|x)\propto P(y_j|x)\propto P(x|\theta_j)/P(x)$时，或 $P(x|\theta_j)=P(x|y_j)$时，语义 KL 信息和语义互信息达到最大。如果样本足够大，就有：

$$T(\theta_j\mid x)=\frac{m(x,y_j)}{mm_j},\qquad m(x,y_j)=\frac{P(x,y_j)}{P(x)P(y_j)},\ mm_j=\max(m(x,y_j)). \tag{34}$$

其中 mm_j 是函数 $m(x,y_j)$的最大值. 作者已经证明【17】，上式和和汪培庄的随机集落影理论【51】兼容，我们可认为上式中 $T(\theta_j|x)$来自随机点落影。

我们称 $m(x, y_j)$是关联函数，它在 0 和∞之间变化。注意：关联度是对称的，而真值函数或隶属函数一般是不对称的，即 $T(\theta_{xi}|y_j)\neq T(\theta_j|x_i)$，这是因为 $mm_i=\max[m(x_i, y)]$不一定等于 mm_j。如果我们用 $m(x, y)$矩阵中最大者代替 mm_j 和 mm_i，真值函数就和失真函数一样是对称的。如果那样，一个真值函数最大值可能小于 1，因而不便于用负指数函数表示。但是，不同实例之间的相似函数 $S(x, \hat{x}_j)$应该是对称的，即 $S(x_j, \hat{x}_i)=S(x_i, \hat{x}_j)$。用对称的失真函数表示的真值函数也是对称的。

笔者认为 $m(x,y_j)$是非常重要的函数。从计算的角度看，先有 $P(x, y_j)$后有 $m(x,y_j)$，但是从哲学的角度看，先有 $m(x,y_j)$后有 $P(x,y_j)$。我们用 $m_\theta(x, y_j)$逼近 $m(x, y_j)$，称 $m_\theta(x, y_j)$是逼真函数。

如果样本不够大，我们要用语义 KL 信息公式优化真值函数：

$$T^*(\theta_j\mid x)=\arg\max_{T(\theta_j|x)}\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{35}$$

用 $m(x, y)$做概率预测比用贝叶斯公式更加简单，因为：

$$P(x\mid y_j)=P(x)m(x,y_j),\quad P(y\mid x_i)=m(x_i,y)P(y). \tag{36}$$

用 $m_\theta(x, y)$类似。可惜人脑记住逼真度不容易。不过人脑记住真值函数要容易些。所以我们需要 $T(y_j|x)$或 $S(x, y_j)$。使用真值函数或相似函数，在 $P(x)$改变时，我们也能做概率预测（见等式 (31)）。

4.3. 信息率保真函数 $R(G)$

如果我们用 $I_{ij} = I(x_i; \theta_j)$代替 $d_{ij} = d(x_i, y_j)$，用语义互信息下限 G 代替平均失真 $\bar{d}$ 的上限 D，那么信息率失真函数 $R(D)$就变成信息率保真函数 $R(G)$【13, 18】(参看图 4)。仿照信息率失真函数的推导，我们得到带参数 s 的 $R(G)$函数:

$$G(s)=\sum_i\sum_j P(x_i)P(y_j\mid x_i)I_{ij}=\sum_i\sum_j I_{ij}P(x_i)P(y_j)\exp(sI_{ij})/Z_i,$$

$$R(s)=sG(s)-\sum_i P(x_i)\log Z_i,\quad Z_i=\sum_k P(y_k)\left[m_\theta(x_i,y_j)\right]^s. \tag{37}$$

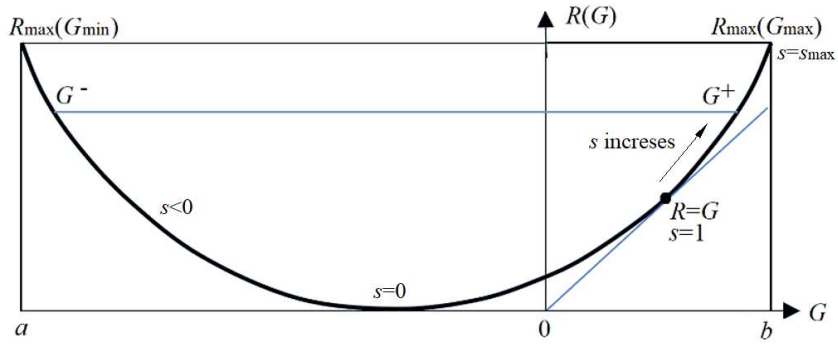


图 4. 二元通信的信息率保真度函数 $R(G)$。任何 $R(G)$函数都是碗状的. 其中有一个 $R(G)=G$ 的点 (s=1)。给定 R 有两个反函数 $G^+(R)$和 $G^-(R)$。

要得到和约束条件相匹配的香农信道，需要通过最小互信息迭代，即重复下面两个公式：

$$\begin{aligned}&P(y_j\mid x_i)=P(y_j)\left[m_\theta(x_i,y_j)\right]^s\Big/Z_i,\ i=1,2,...;j=1,2,...\\&P(y_j)=\sum_i P(x_i)P(y_j\mid x_i),j=1,2,...\end{aligned} \tag{38}$$

$R(G)$函数中的参数 $s=\mathrm{d}R/\mathrm{d}G$ 在右边部分是正的，而 $R(D)$ 函数中 s 总是负的。当 s=1 时, $G=R$, 意味着语义信道匹配香农信道。G/R 意味着信息效率。$R(G)$函数可以应于根据视觉分辨率的图像压缩【16】, 语义压缩【18】, 不可见实例最大互信息分类的收敛证明【13】, 和混合模型收敛证明【73】。关于 $R(G)$函数的更多讨论见【13,18】。

4.4. 信道匹配算法用于机器学习

4.4.1. 用于多标签学习

考虑多标签学习——一种有监督学习。从样本$\{(x_n, y_n), n=1,2,\ldots,N\}$, 我们能得到样本分布 $P(x, y)$。如果样本足够大, 我们可以简单地令 $T(y_j|x)=P(y_j|x)/\max(P(y_j|x))$, $j=1,2,\ldots$; 否则, 我们用语义 KL 信息公式(等式(35))优化 $T(y_j|x)$.

对于多标签分类, 我们可以用分类器：

$$y_j{}^*=h(x)=\arg\max_{y_j} I(x;\theta_j)=\arg\max_{y_j}\log\frac{T(\theta_j\mid x)}{T(\theta_j)}. \tag{39}$$

二元关联方法(Binary Relevance【59】)是不必要的。更多讨论见【13】。

4.4.2. 不可见实例的最大互信息分类

这种分类属于半监督学习。图 5 显示的是一个简单例子——用以说明医学检验和信号检测。Z 是取值 $z \in C$ 的随机变量。概率分布 $P(x)$和 $P(z|x)$ 是给定的。任务是求导致最大香农互信息 $I(X;Y)$的分类器 $y = h(z)$.

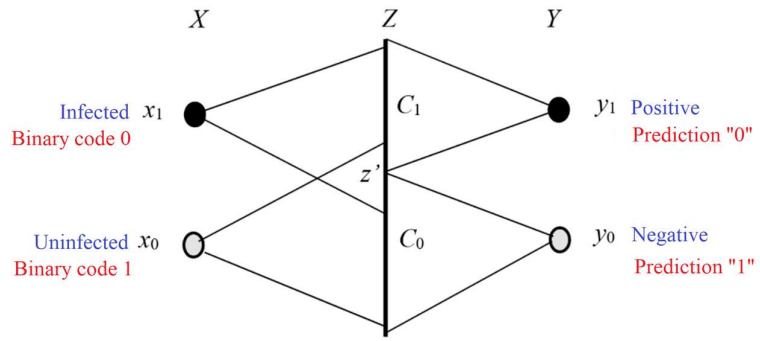


图 5. 图解医学检验和信号检测。我们根据 $z \in C_j$ 选择 y_j。任务是发现导致最大互信息的划分点 z'。

下面方法不限于二分类。设 C_j 是 C 的子集，并且 $y_j = f(z|z \in C_j)$。$S=\{C_1, C_2, \ldots\}$ 是 C 的一个划分。我们的目的是找到：

$$S^* = \arg\max_S I(X;Y_\theta|S) = \arg\max_S \sum_j \sum_i P(C_j)P(x_i|C_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}. \tag{40}$$

先假设一个划分，然后做下面迭代。

匹配 1：让语义信道匹配香农信道，并设置奖励函数。

首先我们根据给定的 S 获得香农信道：

$$P(y_j|x) = \sum_{z_k \in C_j} P(z_k|x),\ j = 1,2,\ldots,n. \tag{41}$$

从 $P(y|x)$我们能得到 $T(\theta_j|x)$。$P(x)$和 $T(\theta_j|x)$，我们能得到 $T(\theta_j)$(或得到 $m_\theta(x,y) = m(x,y)$)和 $I(x;\theta_j)$。然后，对于每个 z，我们得到条件信息或奖励函数：

$$I(X;\theta_j|z) = \sum_i P(x_i|z)I(x_i;\theta_j)\ ,\quad j=0,1,\ldots,n. \tag{42}$$

匹配 2：令香农信道匹配语义信道——通过分类器：

$$y_j^* = f(z) = \arg\max_{y_j} I(X;\theta_j|z),\ j=0,1,\ldots,n. \tag{43}$$

重复匹配 1 和匹配 2 直至 S 不再变化。收敛的 S 就是我们要找的 S^*。笔者已经在【13】(3.3 节)通过 $R(G)$函数解释了上述迭代收敛。

图 6 显示了一个例子。详细数据见【13】(4.2 节)。图 6a 中两条水平线表示初始划分。图 6d 显示互信息随迭代次数变化。

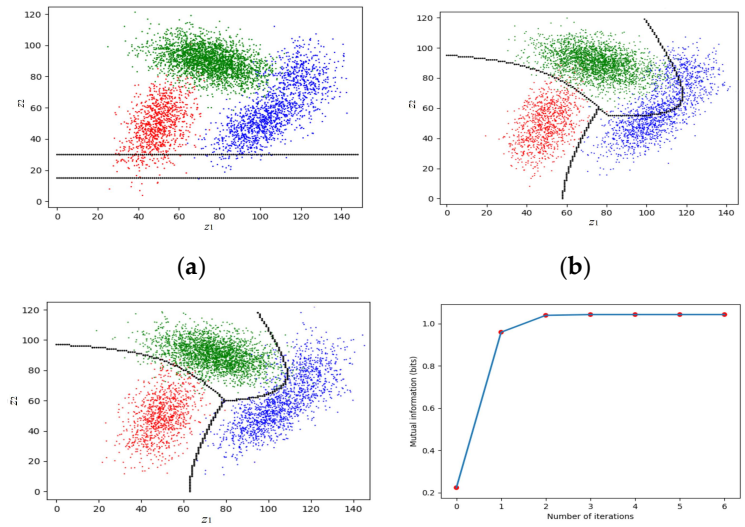


图 6. 最大互信息分类——用一个很坏的初始划分。收敛快速稳定(两次就接近最大互信息)，无需考虑梯度。 (a) 很坏的初始划分。(**b**)首次迭代之后的划分。(**c**) 二次迭代后的划分。(**d**) 互信息随迭代次数变化.

4.4.3. 解释和改进用于混合模型的 EM 算法

我们知道 $P(x)=\sum_j P(y_j)P(x|y_j)$。给定样本分布 $P(x)$，我们使用 $P_\theta(x)=\sum_j P(y_j)P(x|\theta_j)$逼近 $P(x)$，使得相对熵 $H(P\|P_\theta)$ 接近 0。迭代之前我们设置初始的 $P(x|\theta_j)$和 $P(y_j)$, j=1,2,… 每个迭代包含下面两个匹配步骤(详见【13,73】):

匹配 1：令香农信道 $P(y|x)$匹配语义信道 $T(y_j|x)$，通过重复下面两个公式 n 或更少次：

$$P(y_j|x)=P(y_j)P(x|\theta_j)/P_\theta(x),\ P_\theta(x)=\sum_j P(y_j)P(x|\theta_j),$$
$$P^{+1}(y_j)=\sum_x P(x_i)P(y_j|x_i). \tag{44}$$

上面所有步骤中，改变 θ 的第一步, R 和 G 可能增大也可能减小，其他步骤都不改变 θ，都会减小香农互信息。

匹配 2：令语义信道匹配香农信道，通过

$$P(x|\theta_j^{+1})=P(x)P(x|\theta_j)/P_\theta(x),\ P_\theta(x)=\sum_j P(y_j)P(x_i|\theta_j). \tag{45}$$

直至相对熵或 θ 不能改进为止。

为了证明迭代收敛，我们能推导出【13】:

$$H(P\|P_\theta)=R-G+H(P^{+1}(y)\|P(y)). \tag{46}$$

这一公式也能用来解释深度学习预训练。

因为匹配 2 最大化 G，且匹配 1 最小化 R 和 $H(P^{+1}(y)\|P(y))$, $H(P\|P_\theta)$

能接近 0。

这一算法可谓 EnM 算法，和 EM 算法比，每个迭代中，我们重复等式 (44) n 或更少次 (比如 $n\leq 3$) 使得 $P^{+1}(y)\approx P(y)$。EnM 算法在大多数情况下比 EM 算法表现更好。收敛证明可以帮助我们避免盲目改进。

图 7 显示一个高斯混合模型的例子，用以比较 EM 和 E3M 算法。真实模型参数是：$(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1)) = (100, 125, 10, 10, 0.7)$.

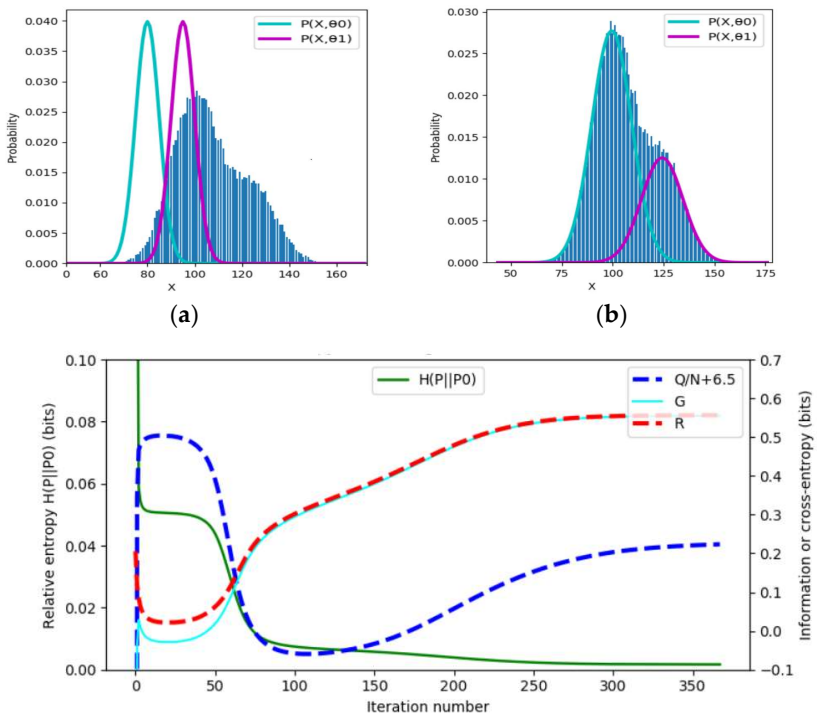


图 7. 用一个不易收敛的例子比较 EM 和 E3M 算法。EM 算法需要大约 340 次迭代，而 E3M 算法需要大约 240 次迭代。收敛过程中，完整数据对数似然度 Q 并不是单调增加的。$H(P||P_\theta)$随 $R-G$ 减小。(**a**)两个初始部件(compponents)的均值 $(\mu_1, \mu_2) = (80, 95)$。(**b**)两个部件全局收敛。(**c**) Q, R, G, 和 $H(P||P_\theta)$随迭代次数变化。初始参数是: $(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1)) = (80, 95, 5, 5, 0.5)$).

这个例子揭示：EM 算法收敛是因为语义(或预测)互信息 G 和香农互信息 R 相互靠近，并不是因为完整数据对数似然度 Q 不断增大(流行观点)。详细讨论见【73】中的 3.3 节。

注意：需要多少次迭代主要取决于初始参数。如果我们根据公平竞争原则【73】使用初始化$(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1)) = (105, 120, 5, 5, 0.5)$，EM 算法只需要 4 次迭代，E3M 算法只需要 3 次迭代。文献【73】提供了一个初始化地图，它能告诉一对初始均值(μ_1, μ_2)好不好。

文献【13,73】提供了更多例子用以检验 EnM 算法。

5. 讨论 1：澄清一些问题

5.1. 互信息最大化是好目标？

在有些研究者的文章中，比如在【3,24】中，他们实际做的是最大化估计互信息，但是解释的时候却说是最大化互信息（香农互信息）。他们认为估计互信息是香农互信息的下界，我们可以通过增大下界来增大香农互信息。对此，Tschannent 等人质疑"最大互信息是好的目标？"【25】笔者认为，他们的质疑值得重视。

一般情况下，最大估计或语义互信息才是目的，有时候我们可以通过增大其上界（香农互信息）来增大估计互信息。下面我们分几个场景来谈香农互信息和语义或估计互信息的最大化。

对于有监督学习，比如多标签学习和分类（简记为：$X \to Y$）。在学习阶段，样本分布 $P(x, y)$约束了香农互信息 $I(X; Y)$，它是语义互信息的上限——这个上限是不能变大的。不过在分类阶段，我们根据逼真(Truthlikeness)函数 $m_\theta(x, y_j)$, j=1,2,…分类，使香农互信息和语义互信息同时最大化。从信息率保真度函数的角度看，就是令 $s \to \infty$，使得 $P(y_j|x)$=1 或 0（见等式（38）），R 和 G 同时到达 $R(G)$曲线的右边最高点。当 s 从 1 增大时，语义通信效率 G/R 是减小的（参看图 4）。有时候我们要在最大 G 和最大 G/R 之间权衡。

对于半监督学习，比如不可见实例的最大互信息分类（简记为 $X \to Z \to Y$，见图 5），语义互信息是目标函数，香农互信息是其上界，我们通过最大化香农互信息来提高语义互信息上界，最后使两者同时最大化。

为什么语义通信中香农互信息不适单独作为目标函数？因为它和对错无关。比如，我们根据年龄 x 把所有人分为 4 类，即四个标签：y_1="小孩"，y_1="年轻人"，y_3="成年人"，y_4="老年人"。分类使得 $P(y_1)=P(y_2)=P(y_3)=P(y_4)=1/4$ 时，香农互信息最大。但是这并不是我们的目的。另外，即使我们对调两个年龄集合(比如{小孩}和{老年人})的标签，也不改变香农互信息。但是标签的不正确使用，语义互信息会大大减少，甚至是负的。因此，只是最大化香农互信息是有问题的。然而，语义互信息最大化能保证失真小，预测信息多；最大语义信息准则等价于最大似然准则，并且兼容 RLS 准则。另外，香农互信息最大化不是一个好的目标还因为使用香农后验熵作为损失函数是不适当的(见 5.4 节)。

5.2. 解释深度神经网络：用 $R(G)$函数还是信息瓶颈？

我们以 AutoEncoder 为例讨论香农互信息和语义互信息在自监督学习中的变化。AutoEncoder【69】的结构是 $X \to Y \to \hat{X}$。X 是一个输入图像，Y 是 X 的低维表示，$\hat{X}$是 X 的估计。成功的学习方法是：首先预训练（比如用 RBM）从 X 到 Y 的网络参数，如混合模型，最小化相对熵 $H(P\|P_\theta)$；然后微调(Fine-tune) $\hat{X}$和 X 之间的网络参数——最小化用$\hat{X}$代表 X 的损失(用 RLS 准则)，或最大化估计互信息 $I(X; \hat{X}_\theta)$。

从语义信息 G 理论角度看，预训练就是求解混合模型，使得

R=G 且相对熵接近 0。微调就是最大化 R 和 G(不要求 R=G)。从 $R(G)$ 函数看 (见图 4)，预训练就是找到 s=1 的点，微调在减小 X 和$\hat{X}$之间的失真的同时增大 R 和 G，使得 s=s_{max}。

是否可以直接通过增大 s 来增大 $I(X;\ \hat{X}_\theta)$? 理论上是可行的。设 $d(x,\ \hat{x}_j)=\exp[-|x-x_j|^2/(2\sigma^2)]$且 $T(\theta_j|x)=\exp[-sd(x,\ y_j)]$ (s>0)。增大 s 就是缩小了真值函数 (或模糊集合 θ_j 的隶属函数) 的覆盖范围。如果 x 和$\hat{x}$都是模糊集合 θ_j 中的元素，两者之间的平均失真或距离也会减小。实际上增大 s 是否足够，需要检验。

根据信息瓶颈理论【27】，要优化一个深度神经网，比如深度信念网，其结构是 $X \to T_1 \to \ldots \to T_m \to Y$，我们需要最小化 $I(T_{i-1};\ T_i) - \beta I(Y;\ T_{i-1}|T_i)$ (i = 1, 2, …, m)，和求解 $R(D)$函数类似。这个想法很有启发。然而，本文作者相信每个隐含层 (T_i)都需要语义互信息最大化和香农互信息最小化. 理由是：

- 深度神经网通常需要预训练和微调。在预训练段，RBM 用于每个隐含层；
- RBM 类似于 EM 算法用于混合模型【78,79】。作者已经证明(参看等式 46)，我们需要最大化语义互信息和最小化香农互信息，从而使混合模型收敛。

另外，信息瓶颈理论认为我们需要最大化 X 和$\hat{X}$或 Y 和$\hat{Y}$之间的香农互信息，而 G 理论建议最大化它们之间的语义互信息。

5.3. 理解吉布斯分布、划分函数、最小互信息匹配和 RBM

吉布斯分布可以表示为相格或状态上的分布

$$P(e_k|T)=\exp(-\frac{e_k}{kT})/Z,\ \ Z=\sum_l \exp(-\frac{e_l}{kT})。\quad (47)$$

也可以表示为能量上的分布【43】:

$$P(e_i|T)=P(e_i)\exp(-\frac{e_i}{kT})/Z,\ \ Z=\sum_l P(e_l)\exp(-\frac{e_l}{kT}).\quad (48)$$

对于机器学习，要预测 x_i 发生的后验概率，我们要用和后一个公式类似的公式:

$$P(x_i|\theta_j)=P(x_i)T(\theta_j|x_i)/T(\theta_j),\ T(\theta_j)=\sum_k P(x_k)T(\theta_j|x_k).\quad (49)$$

其中逻辑概率 $T(\theta_j)$就是划分函数 Z。

如果我们将 $T(\theta_j|x) = m(x, y_j)/mm_j$ 代入 $T(\theta_j)$就得到：

$$T(\theta_j)=\sum_i P(x_i)m(x_i,y_j)/mm_j=\sum_i P(x_i|y_j)/mm_j=1/mm_j.\quad (50)$$

从上面两个等式，我们得到 $P(x|\theta_j)$=$P(x|y_j)$。可见 $T(\theta_j)$中求和只是为了求出 $1/mm_j$，以便和分子中 $1/mm_j$ 同时消除，使得 $P(x|\theta_j)$逼近 $P(x)m(x, y_j)$=$P(x|y_j)$。至于 mm_j 究竟多大，并不重要。

有人用 SoftMax 函数表示学习函数 $P(x, y|\theta_j)$和概率预测，比如

$P(x, y|\theta_j)$=exp(-$d(x, y_j)$]和

$$P(x|\theta_j)=\frac{P(x,y_j|\theta)}{\sum_k P(x_k,y_j|\theta)}=\frac{\exp(-d(x,y_j))}{\sum_k \exp(-d(x_k,y_j))}. \quad (51)$$

这很像是语义信息方法中的 $m_\theta(x, y_j)$, 但是本质不同。语义信息方法中,

$$P(x|\theta_j)=P(x)m_\theta(x,y_j)=\frac{P(x)\exp(-d(x,y_j))}{\sum_k P(x_k)\exp(-d(x_k,y_j))}. \quad (52)$$

其中有 $P(x)$ and $P(x_k)$.

给定 $P(x)$和语义信道 $T(y|x)$, 我们可以令 $P(y_j|x)=kT(y_j|x)$ 做最小互信息匹配, 然后得到 $P(y_j)=kT(y_j)$, $k=1/\sum_j T(y_j)$。于是：

$$\begin{aligned} P(y_j)&=T(y_j)/\sum_j T(y_j),\\ P(y_j|x)&=T(y_j|x)/\sum_j T(y_j). \end{aligned} \quad (53)$$

注意：如果 $T(y|x)$也需要优化，我们只能用最小互信息迭代(见 6.2 节)。使用上面公式，$T(y|x)$不能被进一步优化。

另外, 给定 $m(x, y)$, $P(x)$和 $P(y)$是相关的, 我们不能简单地令: $P(x, y)$=p$(x)m(x, y)P(y)$, 因为这样得到的新的 $P(y)$可能不是归一化的, 即

$$\begin{aligned} P^{+1}(y_j)&=\sum_i P(x_i)m(x_i,y_j)P(y_j),\\ \sum_j P^{+1}(y_j)&\neq 1. \end{aligned} \quad (54)$$

不过, 我们可以固定 $P(x)$和 $P(y)$中的一个, 用上述方法得到另一个。比如，给定 $P(x)$和 $m(x, y)$, 我们首先得到 $T(y_j|x)=m(x, y_j)/mm_j$, 然后使用等式 (53)得到 $P(y)$.

等式(54)能帮助我们理解 RBM。一个结构为 V —> H 的 RBM [60] 包含一组参数 $\theta=\{a_i, w_{ij}, b_j|i=1, 2, \ldots, n; j=1, 2, \ldots, m\}$。参数 a_i, w_{ij}, 和 b_j 分别关联到 $P(v_i)$, $m_\theta(v_i, h_j)$和 $P(h_j)$。优化权重$\{w_{ij}\}$改进 $m_\theta(v, h)$ ，即最大化语义互信息；优化$\{b_j\}$改进 $P(h)$，即最小化香农互信息。交替优化使得语义互信息和香农互信息相互接近。

5.4. 理解模糊熵, 覆盖熵和失真

我们先假设 θ 是清晰集合, 看覆盖熵 $H(Y_\theta)$的性质。

我们以年龄和相应的分类标签为例说明覆盖熵。假设年龄 x 以 15, 30, 55 为界, 划分为四个清晰子集, 它们的标签是 y_1=”小孩”, y_2=”年轻人”, y_3=”中年人”, y_4=”老年人”。这时四个子集构成 U 的一个划分。如果以 18 岁为界把 U 划分成两个子集, 添加标签 y_5=”成年人” 和 y_6=”非成年人”, 六个子集就构成 U 的一个覆盖。

笔者曾证明[16], 覆盖熵 $H(Y_\theta)$代表给定 $P(y)$, 解码 y 产生 x 时的最小香农互信息。约束条件是：如果 $x\notin\theta_j$, $P(x|y_j)$=0。有人称这

一最小香农互信息为复杂性失真【77】。一个简单的证明方法是利用信息率失真函数。我们定义失真函数：

$$d(x, y_j) = \begin{cases} 0, \ x \in \theta_j, \\ \infty, \ x \notin \theta_j. \end{cases} \tag{55}$$

那么就以得到

$$R(D=0) = H(Y_\theta) = -\sum_j P(y_j) \log T(\theta_j). \tag{56}$$

如果这些子集构成 U 的一个划分，覆盖熵就变成香农熵。

如果这些子集是模糊的，约束条件是 $T(\theta_j|x) < 1$ 时，$P(x|y_j) \le P(x|\theta_j)$。那么最小 ShMI 就等于覆盖熵减去模糊熵，即 $R = I(X;Y) = H(Y_\theta) - H(Y_\theta|X)$ [18].

香农失真函数也确定了一个语义信息量。给定信源 $P(x)$和失真函数 $d(x, y_j)$，我们可以得到真值函数 $T(\theta_j|x)=\exp(-d(x, y_j))$和逻辑概率 $T(\theta_j)=Z_j$，再令 $P(y_j|x)=kT(\theta_j|x)$，就得到 $P(y_j|x)=\exp(-d(x, y_j))/\sum_j T(\theta_j)$。然后就有 $I(X;Y)=H(Y_\theta)-\bar{d}=H(X;Y_\theta)$。可见：限失真时的最小香农互信息就等于语义互信息；语义互信息随平均失真增加而减少。

我们也看到很多人用香农后验熵 $H(Y|X)$表示平均失真（或损失），或用$-\log P(y|x_i)$表示失真。按这种用法，$P(y|x_i)$小于 1 就存在失真。但是，当标签增加，使得它们外延构成 U 的覆盖或模糊覆盖时，$P(y|x_i)$可能很小，但是并不意味着失真。比如对于年龄的例子，当上面 6 个标签同时使用时，y_6="x 是非成年人"很少被使用，因而 $P(y_6|x<18)$很小。$-\log P(y_6|x<18)$很大。但是从语义信息角度看，$T(\theta_6|x<18)=1$，不存在失真。所以，用 $H(Y|X)$表示失真是不合理的。另外，我们可以这样理解：标签是否失真和它被使用的频率没有必然关系，而 $P(y|x_i)$和 $P(y)$有关。

用 KL 距离表示失真或损失【27】也有同样问题，而用语义 KL 信息表示负的损失合适。

因为语义互信息可以写成：覆盖熵（最大可能语义信息）减去平均失真，它可以作为负的损失函数。

5.5. 评价学习方法：信息准则还是正确率准则？

在笔者看来，机器学习领域存在一个奇怪的现象：大家用最大似然，RLS，最大互信息，最大估计信息…准则优化模型参数，但是最后却用开放数据集上的分类正确率证明自己的方法好。既然如此，为什么不用正确率准则优化模型？有人解释说，正确率准则不能保证分类器有鲁棒性，一种分类器可能适合一种信源 $P(x)$而不合适另一种信源。但是，用开放数据集就能检验分类器的鲁棒性吗？

笔者认为，使用信息准则（包括最大似然，最小交叉熵，最大估计互信息准则），不仅因为要提高鲁棒性，还因为要减少小概率事件

漏报，即增大小概率事件的检出率，因为小概率事件漏报带来的信息损失更大，经济损失往往也更大。比如对于医学检验、Xray 图像识别，天气预报、地震预报等等。

如何解决学习准则和评价准则的不一致性？也用信息准则评价各种方法当然最好，但是实际上不可行。为此笔者建议：在分类准确率不是很高的情况下，除了整体正确率，我们还应该检查小概率事件正确率，然后用两个正确率构成的矢量评价学习方法。

6. 讨论 2：探索新的方法

6.1. 用香农信道训练高斯语义信道

用样本或样本分布训练用 SoftMax 函数表示的真值函数是复杂的，通常用梯度下降方法。如果我们能把 SoftMax 函数学习转换为高斯函数学习，就可以像 EM 算法用于高斯混合模型那样，无需考虑梯度，学习就会非常简单。

假设 $P(y_j|x)$正比于高斯函数。那么 $P(y_j|x)/\sum_k P(y_j|x_k)$就是高斯分布。我们可以假设 $P(x)$ 是等概率的，即 $P(x)=1/|U|$ ($|U|$是论域 U 中元素数量)，然后用 $P(y_j|x)$优化 $T(\theta_j|x)$ —— 通过最大化语义 KL 信息:

$$
\begin{aligned}
I(X;\theta_j) &= \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \\
&= \sum_i \frac{P(y_j|x_i)}{\sum_k P(y_j|x_k)} \log \frac{T(\theta_j \mid x_i)}{\sum_k T(\theta_j|x_k)} + \log|U|.
\end{aligned}
\tag{57}
$$

显然，当 $T(\theta_j|x) \propto P(y_j|x)$时，$I(X;\theta_j)$达到最大，这意味着我们能用 $P(y_j|x)$的期望 μ_j 和标准差 σ_j 作为 $T^*(\theta_j|x)$的期望和标准差。

另外，如果我们只知道 $P(x|y_j)$和 $P(x)$，我们可以用 $m(x, y_j) = P(x|y_j)/P(x)$代替 $P(y_j|x)$，然后优化高斯真值函数。这需要不存在使 $P(x)=0$ 的 x。为此，我们可以用接近 $P(x)$的不间断分布取代 $P(x)$.

上述优化高斯真值函数的方法不需要梯度下降.

6.2. 高斯信道混合模型和信道混合模型机

到目前为止，大家只用过用似然函数(比如高斯似然函数)构成的混合模型。然而，我们也能用高斯真值函数构造**高斯信道混合模型**(GCMM)。

我们也能用 EnM 算法求解 GCMM。匹配 1 和匹配 2 变为：

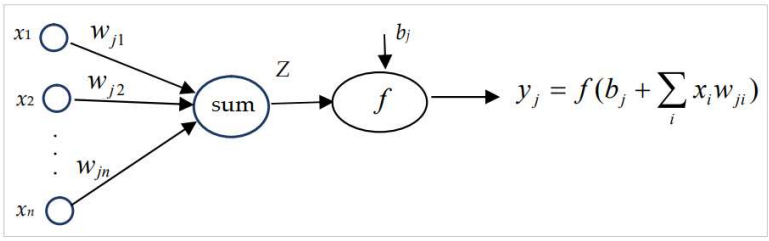
匹配 1：让香农信道匹配语义信道——通过用 $P(x|\theta_j) = P(x)T(\theta_j|x)/T(\theta_j)$并且重复等式(44) n 次。

匹配 2：让语义信道匹配香农信道——通过赋值：

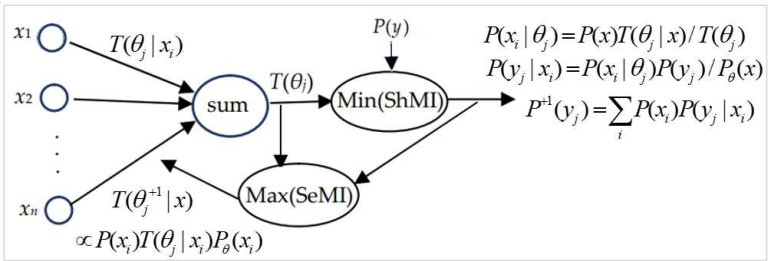
$$
\begin{aligned}
T(\theta_j^{+1} \mid x) &= \exp\left[\frac{-(x-\mu_j)^2}{2\sigma_j^{\,2}}\right] \propto \frac{P(y_j \mid x)}{P(y_j)} \\
&= \frac{P(x)T(\theta_j \mid x)}{T(\theta_j)P_\theta(x)} \propto \frac{P(x)T(\theta_j \mid x)}{P_\theta(x)},
\end{aligned} \tag{58}
$$

这意味着我们能用 $P(y_j|x)$或 $P(x)T(\theta_j|x)/P_\theta(x)$的期望和标准偏差作为 $T(\theta_j^{+1}|x)$的期望和标准偏差。

我们也能建立像信道混合模型一样工作的**信道混合模型机(CMMM)**。图 8 包括一个典型的神经元和 **CMMM** 的一个神经元。**图 8b 中，真值** $T(\theta_j|x_i)$用作权重 w_{ji}。输入可以是 x_i, 一个矢量 **x**, 或一个分布 $P(\mathbf{x})$, 为此，我们用不同的方法求出相同的 $T(\theta_j)$.



(a)



(b)

图 8. 比较典型的神经元和信道混合模型机(**CMMM**)的神经元。(**a**)一个典型的神经元。(**b**) CMMM 的一个神经元及其优化.

图 8a 中，典型神经元中 f 是激活函数。如果 $b_j = P(y_j) - T(\theta_j) \le 0$，且 f 是 Relu 函数, 则两个神经元等价.

对于 CMMM，我们只需要通过重复赋值优化网络参数。初始真值或权重可以在 0 和正数 c 之间——因为用 $cT(\theta_j|x)$ 代替 $T(\theta_j|x)$ 并不改变 $P(x|\theta_j)$.

作者初步试验表明：CMMM 工作如预期那样，收敛快速。优化权重只需要 3-4 次迭代。我们需要更多试验检验是否能用 CMMM 或高斯 CMMM(权重分布限于高斯函数)代替 RBM。

6.3. 用统计方法求不同单词之间的相似度

考虑求两个单词 x 和 y_j 之间的相似度。我们可以简单地从样本分布得到：

$$S(x, y_j) = \begin{cases} 0, & \text{if } P(x, y_j) = 0, \\ \dfrac{m(x, y_j)}{\max(m(x, y))}, & \text{otherwise.} \end{cases} \tag{59}$$

其中 max[$m(x, y)$]是所有可能的 $m(x, y)$中最大者。这样得到的相似度是对称的。这一相似度可谓相对逼真度 (Relative Trughlikeness) 。这种方法的优点是：

- 不需要借助于 WordNet 那样的语义结构；
- 和改进的 PMI 相似度【65】类似，在 0 和 1 之间变化；
- 适合做概率预测。

其缺点是，语义可能不明确。我们可以把 x 和 y 分别限制在不同的较小论域内，得到不对称的真值函数(见等式 34)，从而更好反映语义。

基于 WordNet 结构建立的相似度 (比如 Resnik 相似度【23】) 考虑了一个具有较小逻辑概率的标签含有更多信息内容，类似于逼真度 $m_\theta(x, y)$。 相对逼真度 $S(x, y_j)$要和 $P(x)$一起转换成逼真度，然后才好和 Resnik 之类相似度比较。逼真度可以直接用于分类，而本文定义的相似度特别适合概率预测。它用于分类时，需要结合 $P(x)$得到逼真度 $m_\theta(x, y)$。

隐含语义分析 (LSA) 已经成功地用于自然语言处理【64】，语义信息方法如何与 LSA 结合是否能得到更好结果， 值得探索。

6.3. 合目的信息和信息价值作为奖惩函数用于强化学习

我们可以用语义信息测度度量合目的信息——关于结果怎样符合目的【80】。比如，一个小车上的乘客和驾驶员从 GPS 地图得到的关于离目的地还有多远的信息。对于乘客，是合目的信息，而对于驾驶员是用于控制的反馈信息。如果我们训练机器驾驶，合目的信息就可以作为强化学习的奖励函数。

控制系统一般用误差作为反馈信号，但是维纳提出要用熵反映结果的不确定性。有时候我们需要兼顾两者，就是既要准确(均值误差小)，又要精确(标准偏差小)。G 测度就符合这一目的。

研究者已经把香农信息论用于约束控制和强化学习【81,82】，香农互信息(或 KL 离散度 D_{KL})就代表控制复杂性，给定失真限制 D，互信息 R 越小则效率越高。把语义信息 G 理论用于约束控制和强化学习，语义互信息就是奖励函数，G/R 就反映控制效率。

我们以驾车到达目的地为例说明强化学习。为此需要考虑下面问题:

- 选择动作 a 达到目的;
- 系统状态从 $P(x)$变为 $P(x|a)$；
- 设置奖励函数， 它是目标，$P(x)$和 $P(x|a)$的函数。

强化学习需要解决的问题：

1. 怎样通过学习得到 $P(x|a)$或 $P(x|a, h)$ (h 表示历史)；

2. 如何根据当前状态和奖励函数选择适当的 a；
3. 怎样经济地实现目的，即：怎样权衡奖励最大化和控制效率最大化。

下面我们假设 $P(x|a)$已经学习好了，要解决的是问题 2 和 3.

要控制不确定事件，有时候，目标不是一个点，而是一个模糊范围——可以用真值函数表示，比如，某国平均死亡年龄是 50 岁，它设立目标：y_j=”人口死亡年龄最好不低于 60 岁”。假定控制方法是改进医疗条件。这时香农信息反映控制成本，语义信息反映控制效用。虽然我们可以花较大的成本把平均死亡年龄提高到 80 岁或更高，但是从控制效率的角度看，是不经济的。因为死亡年龄到达 62 岁左右就有较大的合目的信息。作者在【80】中讨论了这个例子。

假设目的用 Logistic 函数表示：$T(\theta_j|x)$=1//[1+exp(-0.8(x-60))](见图 9)，先验分布 $P(x)$是正态的 (μ=570，σ=10)，一种医疗条件 a_j 的控制结果是 $P(x|a_j)$ (也是正态的)，则控制 a_j 的合目的信息或奖励函数是：

$$I(X;\theta_j\,|\,a_j)=\sum_i P(x_i\,|\,a_j)\log\frac{T(\theta_j\,|\,x_i)}{T(\theta_j)}=\sum_i P(x_i\,|\,a_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}. \tag{60}$$

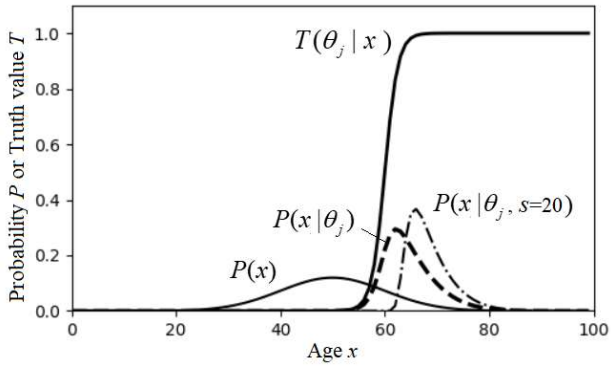


图 9. 图解人口死亡年龄控制，用以说明度量合目的信息。$P(x|a_j)$ 逼近 $P(x|\theta_j) = P(x|\theta_j, s = 1)$时信息效率 $G/R = 1$。当 $P(x|a_j)$逼近 $P(x|\theta_j, s = 20)$时，G 和 R 接近最大可能值。

我们能选择 a_j 或其 μ_j 和 σ_j 使得 $P(x|a_j)$逼近 $P(x|\theta_j)$(即令两者 μ 和 σ 尽可能相同)，于是得到最小 KL 信息 $I(X; a_j)$，使得信息效率 $I(X; \theta_j|a_j)/I(X; a_j)$接近 1。另外，我们可以用求 $R(G)$函数使用的方法增大合目的信息，即使用

$$P(x_i\,|\,a_j)\approx P(x_i)T(\theta_j\,|\,x_i)^s\Big/\sum_k P(x_k)T(\theta_j\,|\,x_i)^s. \tag{61}$$

通过增大 s 增大 $I(X; \theta_j|a_j)$和 $I(X; a_j)$。这时候信息效率会有所降低。我们可以在最大信息效率和最大奖励之间平衡。

作者在【80】中提供了这个例子的计算结果。假定 $Y=y_j$，则语义互信息 G 就等于 $I(X; \theta_j|a_j)$，香农互信息 R 就等于 $I(X; a_j)$。结果表明：最

大信息效率是 0.95(在 $P(x|a_j)$接近 $P(x|\theta_j$时)。当 s 从 1 增加的 20 时，合目的信息从 2.08 比特增加的 3.13 比特，但是信息效率 G/R 从 0.95 下降 0.8。为了权衡 G 和 G/R，我们最好选择 s 介于 5 和 15 之间.

如果目标有多种，我们还需要根据等式(38)优化 $P(a_j)$ (j=1, 2, …)。

强化学习也可以用信息价值作为奖励函数。

Cover 和 Thomas 曾讨论过用信息理论的方法优化投资组合【83】。笔者曾讨论过利用用语义或预测信息优化投资组合【84】，并得到下面增值熵公式和信息价值公式(详见附录 B)：

$$H_v(\mathbf{R},\mathbf{q}\mid\theta_j)=\sum_i P(\mathbf{x}_i\mid\theta_j)\log R_i(\mathbf{q}), \tag{62}$$

$$V(\mathbf{X};\theta_j)=\sum_i P(\mathbf{x}_i\mid\theta_j)\log(R_i(\mathbf{q}^{**})/R_i(\mathbf{q}^{*})), \tag{63}$$

其中 $\mathbf{q}$ 是投资比例矢量，$R_i(\mathbf{q})$ 是第 i 种价格矢量出现时投资组合的回报，$\mathbf{q}^*$是按价格矢量的先验概率分布 $P(\mathbf{x})$优化的投资比例，$\mathbf{q}^{**}$是按概率预测 $P(\mathbf{x}_i|\theta_j)$优化的投资比例。预测的信息价值可以作为奖励函数，用以优化概率预测和决策。不过要计算实际的信息价值 (在学习阶段)，我们需要用 $P(\mathbf{x}|y_j)$代替等式(63)中的 $P(\mathbf{x}|\theta_j)$.

从增值熵公式可以推导出几个实用公式。比如，设有一种投资会导致两个可能结果：收益率 $r_1<0$ 的概率是 P_1；收益率 $r_2>0$ 的概率是 P_2，则最优投资比例是

$$q^*=E/|r_1r_2|, \tag{64}$$

其中 E 是期望收益。如果 $r_1=-1$，则 $q^*=P_2-P_1/r_2$，这就是著名的 Kelly 公式。上面假设无风险收益率 $r_0=0$，否则有

$$q^*=E_0R_0/|r_{10}r_{20}|, \tag{65}$$

其中 $E_0=E-r_0$, $r_{10}=r_1-r_0$, $r_{20}=r_2-r_0$。

然而，$V(\mathbf{X};\theta_j)$只是特殊情况下的信息价值。其他情况下的信息价值还有待进一步探讨。

6.4. 语义信息 G 理论的局限性

基于语义信息 G 理论，我们似乎可以把信道匹配算法(用于多标签学习，最大互信息分类和混合模型)用于深度神经网络——为各种学习任务。然而，深度神经网络不仅包含统计推理，还隐含了逻辑推理——通过偏置和激活函数(比如 Rule 函数就是笔者建立色觉机制模型用到的模糊逻辑运算【54,55】)。而语义信息理论方法只涉及统计部分。虽然它有助于解释深度学习，但是远远不足以指导我们建立高效的深度神经网络。

目前，深度神经网络在特征抽取，自监督学习，强化学习等方面，显示出强大的生命力。 G 理论作为完整的语义信息理论也是不够的。我们需要的是结合特征抽取和模糊推理发展它,使它跟上深度学习的步伐。

7. 结论

通过回顾语义信息论和学习函数的进化史，我们能看出：平行发展的两个潮流异途同归，最后都使用真值（或相似）函数和最大语义（或估计）信息准则优化语义通信或机器学习。而估计信息测度就是笔者发展的 G 理论中的语义信息测度的特例；相似函数就是 G 理论所用的相似函数或混淆概率函数，是真值函数的一种。相似函数和真值函数正比于 $m(x, y)=P(x, y)/[P(x)P(y)]$，其最大值是 1。和似然函数及反概率函数（比如 Logistic 函数）比，相似函数独立于先验概率分布 $P(x)$和 $P(y)$，具有很好的可迁徙性，更适合多标签学习。

用 G 理论分析机器学习，我们得到结论：香农互信息 R 是语义互信息 G 的上限；G/R 反映信息效率；最大化 G 而不是 R 才是重要的；R 有时候需要最大化（作为 G 的上限），有时候是需要最小化(为了提高语义通信效率）。

文中定义了一些重要函数(比如 $m(x, y)$)和两种广义熵（模糊熵 $H(Y_\theta|X)$和覆盖熵 $H(Y_\theta)$）；还把吉布斯分布解释为语义贝叶斯预测，把划分函数解释为逻辑概率；把能量函数解释为失真函数。

本文澄清了：在大多数情况下，语义互信息最大化(而不是香农互信息最大化)才是重要的。它用 $R(G)$函数解释了自动编码器(AutoEncoder)和类似学习任务：预训练使 X 和 Y 之间的香农互信息和语义互信息相等，即 $G=R\ (s=1)$，微调(Fine-Turn)同时增大 R 和 $G(s\text{—}>\infty)$。和信息瓶颈理论的解释不同，本文解释：RBM 做深度神经网的预训练时，它分别优化每个隐含层——最小化香农互信息并最大化语义互信息。

G 理论和学习函数和语义信息测度的进化提醒我们目前存在两个潜在机会：一是我们能用高斯信道混合模型或信道混合模型机预训练 DNN 的隐含层，而不需要梯度下降和反向传播。在微调阶段，我们能通过增大 $R(G)$函数中的 s 增大语义互信息和香农互信息。这种方法应能显著地简化和加速深度学习。另一个潜在机会是：我们能用 G 测度作为奖励函数优化强化学习。G 测度可以帮助我们权衡合目的信息和信息效率。未来一段时间，涉及这两个机会的课题特别值得研究。

然而，深度学习还涉及特征抽取和模糊推理，语义信息 G 理论还远远跟不上深度学习的要求。因此，结合 G 理论和深度学习得到实用的 DNN，这是一个巨大的挑战。不过 G 理论有希望成为未来统一的语义信息理论的基础部分。语义信息理论和深度学习应能相互促进，使得 DNN 更好解释.

英文缩写和原文

BOYL	Bootstrap Your Own Latent
DIM	Deep InfoMax (Information Maximization)
DNN	Deep Neural Network
DT	DeLuca-Termini
DV	Donsker-Varadhan
EM	Expectation-Maximization
EMI	Estimated Mutual Information
EnM	Expectation-n-Maximization
GCMM	Gaussian Channel Mixture Model
CMMM	Channel Mixture Model Machine
GPS	Global Positioning System
G theory	Semantic information G theory (G means generalization)
InfoNCE	Information Noise Contrast Estimation
ITL	Information-Theoretic Learning
KL	Kullback–Leibler
LSA	Latent Semantic Analysis
MaxMI	Maximum Mutual Information
MinMI	Minimum Mutual Information
MINE	Mutual Information Neural Estimation
MoCo	Momentum Contrast
PMI	Pointwise Mutual Information
SeMI	Semantic Mutual Information
ShMI	Shannon's Mutual Information
SimCLR	A simple framework for contrastive learning of visual representations

附录 A. 关于形式语义

一个标签或外概念的语义包含其外延和内涵。内涵包括各种属性。但是形式语义只包括外延和确定外延的属性。比如"老年人"内涵包括：年龄大，衰老，退休，接近死亡等。但是形式语义只包含"老年人"的外延（即所指）和外延覆盖的年龄(年龄属性)。

附录 B. 增值熵和信息价值定义

设变量 **x** 是一组证券的价格矢量，常量是 $\mathbf{x}_i=(x_{i1}, x_{i2},\ldots)$, $i=1,2,\ldots$。当前价格矢量是 $\mathbf{x}_0=(x_{01}, x_{02}, \ldots)$，回报矢量是 $\mathbf{R}_i=(R_0, x_{i1}/x_{01}, x_{i2}/x_{02},\ldots)$，其中 $R_0=1+r_0$, r_0 是无风险收益率。投资比例矢量是 $\mathbf{q}=(q_0, q_1, q_2,\ldots)$ 。则投资回报是 **q** 的函数，即：

$$R_i(\mathbf{q}) = \sum_k q_k r_{ik} \cdot \tag{a1}$$

设概率预测是 $P(\mathbf{x}|\theta_j)$，增值熵(即期望的翻倍率，如果 log 是 $\log_2$)是：

$$H_v(\mathbf{R},\mathbf{q}\,|\,\theta_j) = \sum_i P(\mathbf{x}_i\,|\,\theta_j)\log R_i(\mathbf{q}), \tag{a2}$$

其中 **q** 表示决策，它也是一个学习函数。我们可以通过优化 **q** 增大增值熵。预测的信息价值是

$$V(\mathbf{X};\theta_j) = \sum_i P(x_i\,|\,\theta_j)\log(R_i(\mathbf{q}^{**})/R_i(\mathbf{q}^{*})), \tag{a3}$$

其中 $\mathbf{q}^*$是根据先验分布 $P(x)$优化的投资比例， $\mathbf{q}^{**}$是根据预测 $P(\mathbf{x}|\theta_j)$优化的投资比例。预测的信息价值可以作为奖励函数，用以优化概率预测和决策。不过要计算实际的信息价值（在学习阶段），我们需要用 $P(\mathbf{x}|y_j)$代替 $P(\mathbf{x}|\theta_j)$.

参考文献：

1. Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. MINE: Mutual information neural estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1–44. https://doi.org/10.48550/arXiv.1801.04062.
2. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. Available online: https://arxiv.org/abs/1807.03748 (accessed on 10 January 2023).
3. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–429 and 623–656.
4. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Trischler, A.; Bengio, Y. Learning Deep Representations by Mutual Information Estimation and Maximization. Available online: https://arxiv.org/abs/1808.06670 (accessed on 22 December 2022).
5. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views. Available online: https://arxiv.org/abs/1906.00910 (accessed on 1 February 2023).
6. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, ICML, PMLR 119, Virtual Event, 13–18 July 2020; pp. 1575–1585.
7. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975.
8. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; Piot, B. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
9. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1963.
10. Bao, J.; Basu, P.; Dean, M.; Partridge, C.; Swami, A.; Leland, W.; Hendler, J.A. Towards a theory of semantic communication. In Proceedings of the 2011 IEEE 1st International Network Science Workshop, West Point, NSW, USA, 22–24 June 2011; pp. 110–117. https://doi.org/10.1109/NSW.2011.6004632.
11. Strinati, E.C.; Barbarossa, S. 6G networks: Beyond Shannon towards semantic and goal-oriented communications. *Comput. Netw.* **2021**, *190*, 107930. https://doi.org/10.1016/J.COMNET.2021.107930.
12. Lu, C. Channels' matching algorithm for mixture models. In *Intelligence Science I, Proceedings of the ICIS 2017, Beijing, China, 27 September 2017*; Shi, Z.Z., Goertel, B., Feng, J.L., Eds.; Springer: Cham, Switzerland, 2017; pp. 321–332. https://doi.org/10.1007/978-3-319-68121-4_35.
13. Lu, C. Semantic information G theory and logical Bayesian inference for machine learning. *Information* ***2019****, 10, 261*. https://doi.org/10.3390/info10080261.
14. 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/ (accessed on 5 March 2019).

15. 鲁晨光，《广义信息论》，中国科技大学出版社，合肥，中国，1993; ISBN 7-312-00501-2.

16. Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* **1999**, *28*, 453–490.

17. Lu, C. The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning. *Philosophies* **2020**, *5*, 25.

18. Lu, C. Using the Semantic Information G Measure to Explain and Extend Rate-Distortion Functions and Maximum Entropy Distributions. *Entropy* **2021**, *23*, 1050. https://doi.org/10.3390/e23081050.

19. Floridi, L. Semantic conceptions of information. In *Stanford Encyclopedia of Philosophy*; Stanford University: Stanford, CA, USA, 2005. Available online: http://seop.illc.uva.nl/entries/information-semantic/ (accessed on 1 March 2023).

20. *Tarski, A.* The semantic conception of truth: And the foundations of semantics. *Philos. Phenomenol. Res.* **1994**, *4*, 341–376.

21. Davidson, D. Truth and meaning. *Synthese* **1967**, *17*, 304–323.

22. Semantic Similarity. In Wikipedia: The Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Semantic_similarity (accessed on 10 February 2023).

23. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence. *arXiv* **1995**, *1*, 448–453.

24. Poole, B.; Ozair, S.; Oord, A.V.D.; Alemi, A.; Tucker, G. On Variational Bounds of Mutual Information. In International Conference on Machine Learning. Available online: https://arxiv.org/abs/1905.06922 (accessed on 10 February 2023).

25. Tschannen, M.; Djolonga, J.; Rubenstein, P.K.; Gelly, S.; Luci, M. On Mutual Information Maximization for Representation Learning. Available online: https://arxiv.org/pdf/1907.13625.pdf (accessed on 23 February 2023).

26. Tishby, N.; Pereira, F.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.

27. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5.

28. Hu, B.G. Information theory and its relation to machine learning. In *Proceedings of the 2015 Chinese Intelligent Automation Conference*; Lecture Notes in Electrical Engineering; Deng, Z., Li, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 336. https://doi.org/10.1007/978-3-662-46469-4_1.

29. Xu, X.; Huang, S.-L.; Zheng, L.; Wornell, G.W. An information-theoretic interpretation to deep neural networks. *Entropy* **2022**, *24*, 135. https://doi.org/10.3390/e24010135.

30. Rényi, A. On measures of information and entropy. *Proc. Fourth Berkeley Symp. Math. Stat. Probab.* **1960**, *4*, 547–561.

31. Principe, J.C. *Information-theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer Publishing Company: New York, NY, USA, 2010.

32. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. https://doi.org/10.1007/BF01016429.

33. Irshad, M.R.; Maya, R.; Buono, F.; Longobardi, M. Kernel estimation of cumulative residual Tsallis entropy and its dynamic version under ρ-mixing dependent data. *Entropy* **2022**, *24*, 9. https://doi.org/10.3390/e24010009.

34. Liu, W.; Pokharel, P.P.; Principe, J.C. Correntropy: A localized similarity measure. In Proceedings of the 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006; IEEE: Piscataway, NY, USA, 2006.

35. Yu, S.; Giraldo, L.S.; Principe, J. Information-Theoretic Methods in Deep Neural Networks: Recent Advances and Emerging Opportunities. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Survey Track, Montreal, QC, Canada, 19–27 August 2021; pp. 4669–4678. https://doi.org/10.24963/ijcai.2021/633.

36. Oddie, G. Truthlikeness, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Zalta, Edward N., Ed. Available online: https://plato.stanford.edu/archives/win2016/entries/truthlikeness/ (assessed on 18 May 2020).

37. Floridi, L. Outline of a theory of strongly semantic information. *Minds Mach*. **2004**, *14*, 197–221.

38. Zhong, Y. A theory of semantic information. *Proceedings* **2017**, *1*, 129. https://doi.org/10.3390/IS4SI-2017-04000.

39. Popper, K. *Logik Der Forschung: Zur Erkenntnistheorie Der Modernen Naturwissenschaft*; Springer: Vienna, Austria, 1935; English translation: *The Logic of Scientific Discovery*, 1st ed.; Hutchinson: London, UK, 1959.

40. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

41. Carnap, R.; Bar-Hillel, Y. *An Outline of a Theory of Semantic Information*; Technical Report No. 247; Research Laboratory of Electronics, MIT: Cambridge, MA, USA, 1952.

42. Shepard, R.N. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika* **1957**, *22*, 325–345.

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

44. Theil, H. *Economics and Information Theory*; North-Holland Pub. Co.: Amsterdam, The Netherlands; Rand McNally: Chicago, IL, USA, 1967.

45. Zadeh, L.A. Fuzzy Sets. *Inf. Control.* **1965**, *8*, 338–353.

46. De Luca, A.; Termini, S. A definition of a non-probabilistic entropy in setting of fuzzy sets. *Inf. Control.* **1972**, *20*, 301–312.

47. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control*. **1974**, *19*, 716–723.

48. Thomas, S.F. Possibilistic uncertainty and statistical inference. In Proceedings of the ORSA/TIMS Meeting, Houston, TX, USA, 12–14 October 1981.

49. Dubois, D.; Prade, H. Fuzzy sets and probability: Misunderstandings, bridges and gaps. In Proceedings of the 1993 Second IEEE International Conference on Fuzzy Systems, San Francisco, CA, USA, 28 March 1993.

50. Donsker, M.; Varadhan, S. Asymptotic evaluation of certain Markov process expectations for large time IV. *Commun. Pure Appl. Math.* **1983**, *36*, 183–212.

51. Wang, P.Z. From the fuzzy statistics to the falling fandom subsets. In *Advances in Fuzzy Sets, Possibility Theory and Applications*; Wang, P.P., Ed.; Plenum Press: New York, NY, USA, 1983; pp. 81–96.

52. Aczel, J.; Forte, B. Generalized entropies and the maximum entropy principle, In *Bayesian Entropy and Bayesian Methods in Applied Statistics*; Justice, J.H., Ed.; Cambridge University Press: Cambridge, UK, 1986; pp. 95–100.

53. Zadeh, L.A. Probability measures of fuzzy events. *J. Math. Anal. Appl.* **1986**, *23*, 421–427.

54. Lu, C. Decoding model of color vision and verifications. *Acta Opt. Sin.* **1989**, *9*, 158–163. (in Chinese).

55. Lu, C. Explaining color evolution, color blindness, and color recognition by the decoding model of color vision, In *11th IFIP TC 12 International Conference, IIP 2020, Hangzhou, China;* Shi, Z., Vadera, S., Chang, E., Eds.; Springer Nature: Cham, Switzerland, 2020; pp. 287–298. Available online: https://www.springer.com/gp/book/9783030469306 (accessed on).

56. Ohlan, A.; Ohlan, R. Fundamentals of fuzzy information measures. In *Generalizations of Fuzzy Information Measures*; Springer: Cham, Switzerland, 2016. https://doi.org/10.1007/978-3-319-45928-8_1.

57. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc.* **1922**, *222*, 309–368.

58. Fienberg, S.E. When did Bayesian Inference become "Bayesian"? *Bayesian Anal.* **2006**, *1*, 1–40.

59. Zhang, M.L.; Li, Y.K.; Liu, X.Y.; Geng, X. Binary Relevance for multi-label learning: An overview. *Front. Comput. Sci.* **2018**, *12*, 191–202.

60. Hinton, G.E. A practical guide to training Restricted Boltzmann Machines. In *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science*; Montavon, G., Orr, G.B., Müller, K.R., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7700; pp. 599–619. https://doi.org/10.1007/978-3-642-35289-8_32.

61. Ashby, F.G.; Perrin, N.A. Toward a unified theory of similarity and recognition. *Psychol. Rev. 1988, 95, 124–150*. https://doi.org/10.1037/0033-295X.95.1.124.

62. Banu, A.; Fatima, S.S.; Khan, K.U.R. Information content based semantic similarity measure for concepts subsumed by multiple concepts. *Int. J. Web Appl.* **2015**, *7*, 85–94.

63. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2005**, *38*, 188–230. https://doi.org/10.1002/aris.1440380105.

64. Church, K.W.; Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.

65. Islam, A.; Inkpen, D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data* **2008**, *2*, 1–25.

66. Chandrasekaran, D.; Mago, V. Evolution of Semantic Similarity—A Survey. Available online: https://arxiv.org/abs/2004.13820 (accessed on 3 March 2023).

67. Costa, T.; Leal, J.P. Semantic measures: How similar? How related? In *Web Engineering, Proceedings of the ICWE 2016, Lugano, Switzerland, 6–9 June 2016*; Bozzon, A., Cudre-Maroux, P., Pautasso, C., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9671. https://doi.org/10.1007/978-3-319-38791-8_29.

68. Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169.

69. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504-507.

70. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554.

71. Gutmann, M.U.; Hyv ä rinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.* **2012**, *13*, 307–361.

72. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS2016), Barcelona, Spain, 5–10 December 2016; pp. 1857–1865.

73. Lu, C. Understanding and Accelerating EM Algorithm's Convergence by Fair Competition Principle and Rate-Verisimilitude Function. Available online: https://arxiv.org/abs/2104.12592 (accessed on 27 April 2022).

74. Lu, C. Channels' Confirmation and Predictions' Confirmation: From the Medical Test to the Raven Paradox. *Entropy* **2020**, *22*, 384. https://doi.org/10.3390/e22040384.

75. Lu, C. Causal Confirmation Measures: From Simpson's Paradox to COVID-19. *Entropy* **2023**, *25*, 143. https://doi.org/10.3390/e25010143.

76. 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 the 2018 IEEE International Conference on Big Data and Smart Computing, Shanghai, China, 15 January 2018; IEEE Computer Society Press Room: Piscataway, NY, USA, 2018; pp. 15–18.

77. Sow, D.M.; Alexandros Eleftheriadis, A. Complexity distortion theory, *IEEE Trans. Inf. Theory* **2003**, *49*, 604–608. https://doi.org/10.1109/TIT.2002.808135.

78. Nair, V.; Hinton, G. Implicit mixtures of Restricted Boltzmann Machines. In Proceedings of the NIPS'08: Proceedings of the 21st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 8–10 December 2008; pp. 1145–1152.

79. Song, J.; Yuan, C. Learning Boltzmann Machine with EM-like Method. Available online: https://arxiv.org/pdf/1609.01840.pdf (accessed on 10 March 2023).

80. Lu, C. How Semantic Information G Measure Relates to Distortion, Freshness, Purposiveness, and Efficiency. Available online: https://arxiv.org/abs/2304.13502 (accessed on 1 May 2023).

81. Still, S. Information-theoretic approach to interactive learning. *Europhys. Lett.* **2009**, *85*, 28005.

82. Eysenbach, B.; Salakhutdinov, R.; Levine, S. The Information Geometry of Unsupervised Reinforcement Learning. Available online: https://arxiv.org/abs/2110.02719 (accessed on 10 March 2023).

83. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2006.

84. 鲁晨光.《投资组合的熵理论和信息价值——简析股票和期货的风险控制》中国科技大学出版社，合肥，中国, 1997; ISBN7-312-00952-2/F.36.

鲁晨光文章中文翻译见：鲁晨光的广义信息论文章和专著 (survivor99.com),

个人网站见：鲁晨光文集首页 Chenguang Lu first page (survivor99.com)