

语义信息论 (Shannon-Lu Theory) 用于有监督、半监督、无监督和强化学习

The Semantic Information Theory (Shannon-Lu Information Theory) for Supervised, Semi-supervised, Unsupervised, and Reinforcement Learning

Chenguang Lu 鲁晨光

Email: Survival99@gmail.com

Homepage: <http://survivor99.com>

自我介绍 Introducing myself

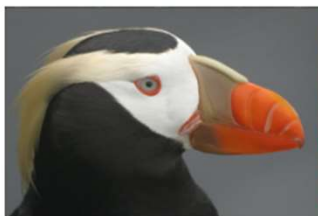
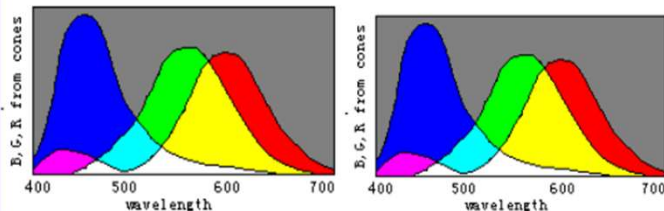
77届南京航空学院学生，毕业设计是逻辑电路，因而想到用三八译码器解释色觉机制；因为使用了模糊逻辑，当过汪培庄教授的访问学者；因为想用分辨率解释色觉信息，进而研究语义信息论。

主页 survivor99.com

鲁晨光探索文集 (中文) 个人主页 Email: [lcguang\(a\)foxmail.com](mailto:lcguang(a)foxmail.com)

Chenguang Lu's Articles and Works (English)

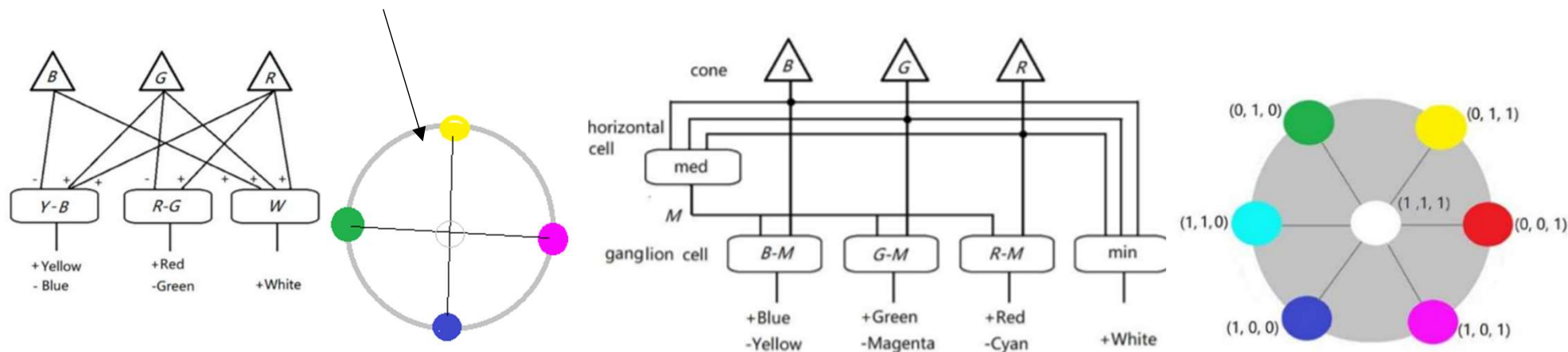
最新研究	研究文章	博客文章	论坛帖子	网站链接	学友网页链接	旧网站(不再更新)
	语义信息论和机器学习 美感和鸟雀华年羽毛 色觉和哲学 财经和经济学	科学网 和讯 ResearchGate ArXiv 博客中国	QQ日志 和讯 天涯 科学网 英文BirdForum	科学网 和讯论坛 天涯 www.csdn.net 知乎	黄有光博客 张学文博客 邓晓明博客 赵南元博客	潜科学 熵和复杂性网站 科学派美学论坛 广义进化论在中国 幸存者
 鲁晨光简历	部分文章	2023英文: 回顾学习函数和语义信息测度的进化——进而理解深度学习 2023英文: 因果确证测度: 从辛普森悖论到COVID-19 2021英文: 需求美学能解释鸟类审美趣味 2021英文: 用语义信息G测度解释和推广信息率-失真函数和最大熵分布(中/英) 2021英文: 用译码模型解释色觉(中/英) 2020英文: P-T概率框架用于语义通信, 证伪, 确证和贝叶斯推理(中/英) 2020发表的英文: 信道确证和预测确证: 从医学检验到乌鸦悖论(中/英) 2019英文: 语义信息G理论和逻辑贝叶斯推理用于机器学习(中/英) 2017 信道匹配算法for最大互信息和最大似然度 2017 兼容Shannon, Popper, Fisher和Zadeh语义信息方法PPT			美感色觉和知识进化——鲁晨光30年科学和哲学研究回顾(2010) PPT文件 用新的财富增量公式统一两种价值理论 作为Shannon和Popper桥梁的广义信息公式 投资滚雪球的数学理论和实践PPT 花脸鸭眼中的环境美和性选择 我终于搞清了鹦鹉华丽外表的秘密 GPS信息和限误差信息率 PPT文件 流行的色觉阶段模型的问题和出路 语义信息公式浅谈和图解 在观鸟论坛的争论	



色觉机制模型：3-8译码模型

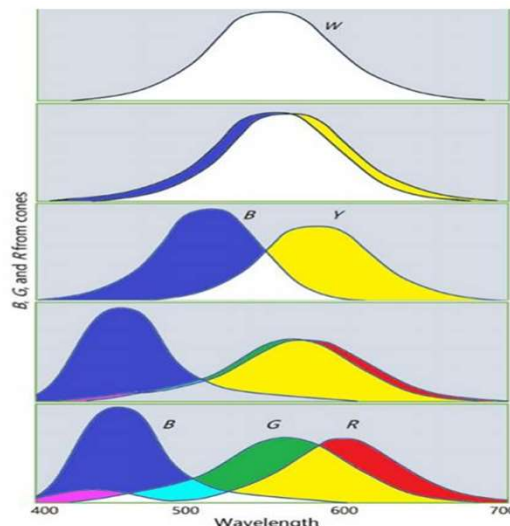
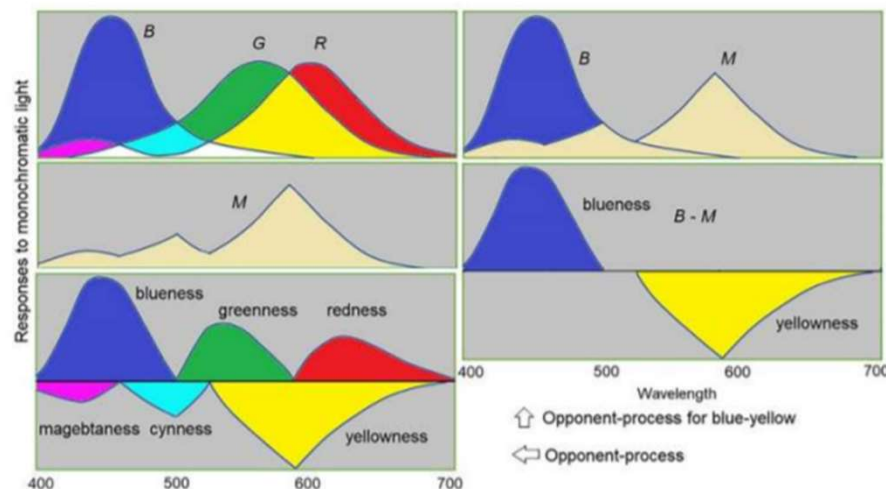
- Young-Helmholtz的三色素理论，解释电视机用的颜色加法很好
- Herring的互补色理论，解释负后象现象很好
- 阶段模型Zone Model

3-8译码模型



- 我提出3-8译码模型，它是对称的，解释互补处理，色盲和色觉进化更好，更适合电脑图像的颜色转换（RGB和HSL系统之间）。文章于1989《光学学报》发表

互补处理



色觉进化

怎样度量色觉信息？

- 为了说明我的模型更好——有三对心理彩色，**分辨率更高，能提供更多信息(非常朴素的观点)**，于是我试图度量色觉信息。而香农公式不行。
- 用负指数函数表示分辨率函数（或相似函数，混淆概率函数）：

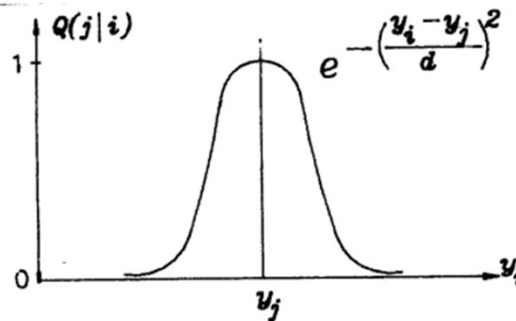


Fig.1 Confusion probability function of y_j with $y_1 = y(x_1)$.

- 用 $\log[Q(j|i)/\text{分子平均}]$ (logSoftmax函数)度量色觉信息.
- 后来发现，把分辨率函数换成隶属函数或真值函数，这个公式也能度量一般的语义信息，**平均信息公式还能反映事实检验**。
- 最早文章于1989年发表在BUSEFUL上。网上有：
- <https://www.univ-smb.fr/listic/production-scientifique/revue-busefal/version-electronique/ebusefal-44/>
- 后来多篇语义信息论文章发表在《通信学报》，Int. J. of General Systems, 《模糊系统和数学》，Entropy, Information. 还发表了专著《广义信息论》，《投资组合的熵理论和信息价值》等。详见我主页。

序言1：香农信息VS语义信息

- 什么是日常语言中的信息或语义信息：

时钟，电子地图上的**GPS**指针，温度表，天气预报，红绿灯，股市指数...，所有疑问句的回答（包括用陈述句和祈使句回答），都提供信息——即语义信息。

- 香农信息和语义，两者相同的是：都是消除不确定性。

- 不同的是：

1) 提供香农信息通过增大概率，而提供语义信息通过缩小或调整概念外延；两者相关，但是一般不等。

2) 香农信息和真假无关；语义信息需要事实检验，和真假有关；

3) 单个事件也提供语义信息（比如**GPS**指针，温度表），但是不提供香农信息。

4) 香农互信息是平均语义信息量的上限，总是正的；而语义信息可能是负的，比如错的预测或谎言提供的信息。天气预报总是报对了或报反了，语义信息量不同，而香农信息量一样多，总是正的。

序言2：信息科学vs信息论 ——区分信息量和信息内容

- 信息论和语义信息论研究的信息都是指信息量， 不是指信息内容。
比如股市涨跌的概率都是 $1/2$ ，预测对了或错了提供的信息量都是 $\log_2=1\text{bit}$ 。
1bite是信息量， 不是信息内容。因为预测涨和预测跌， 信息内容不同。预测气温升降， 信息量可能和预测股市涨跌信息量相同， 但是信息内容更加不同。
很多研究信息科学的人对信息论有抵触情绪，认为怎么你们用**log**就是信息论，我们不用**log**就不是信息论了？

- 我建议：
把（语义）信息论理解为关于（语义）信息量的理论。
用信息科学表示更一般的信息理论（钟义信教授这样做了）。

- 信息科学科学包含：
——关于信息量的理论（包括香农信息论和语义信息论）
——其他方面，特别是关于信息内容的理论（包括信息获取——如遥感，卫星定位，算法等；信息表示——如书籍，展览，标牌；信息存取——如数据库；...）。
信息论和信息科学的其他方面相互支持，但不能相互取代。

序言3：研究语义信息测度的两条路线

- 几乎所有研究语义信息测度的人都认为要把真值或逻辑概率带进语义信息公式（如Carnap和Bar-Hillel, Floridi）。但是路线不同。典型的：
- 1)在模糊熵基础上，如钟义信：

$$I(y_n) = \log 2 + [t_n \log t_n + (1 - t_n) \log (1 - t_n)]$$

- 2) 在Shannon信息公式基础上，如作者，把模型参数，似然函数，真值函数或隶属函数带进香农互信息公式：

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

用语义互信息G代替失真D，把信息率失真函数R(D)推广为信息率逼真度函数R(G)。这样的语义信息理论我称之为G理论或Shannon-Lu理论。因此此理论是香农信息论推广，我不能说是我鲁晨光的理论。

语义信息论数学基础：P-T概率框架

- 同时使用两种概率：统计概率P和逻辑概率T。
- 一个语句的逻辑概率是它被判断为真的概率，而它的统计概率是它被选择的概率。永真句“明天有雨或无雨”逻辑概率是1，而它被选择概率几乎是0。
- 香农信道由转移概率函数构成： $P(y_j|x), j=1,2,\dots$

语义信道由真值函数或隶属函数构成： $T(y_j|x)=T(\theta_j|x), j=1,2,\dots$

谓词 $y_j(x)$ ="x在 θ_j 中"真值函数就是x在模糊集 θ_j 上的隶属度。

- 用真值函数表示形式语义，符合Davidson的真值函数语义学。其实机器学习中负指数函数或高斯函数就是真值函数，最大值是1。
- 逻辑概率即真值的平均，或Zadeh的模糊事件概率：

Zadeh仍然用P，这个机会留给我了

$$T(y_j) = T(\theta_j) = \sum_i P(x_i) T(\theta_j | x_i).$$

- 香农用的贝叶斯公式：

$$P(x | y_j) = P(x) P(y_j | x) / P(y_j),$$

语义贝叶斯公式：

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

上两者可能相等也可能不等。给定 $P(x)$ 和 $P(x|\theta_j)$ ，令 $T(\theta_j|x)$ 的最大值是1，我们能推导出似然函数和真值函数之间的关系：

$$T(\theta_j | x) = \frac{P(x | \theta_j)}{P(x)} \bigg/ \max_x \left[\frac{P(x | \theta_j)}{P(x)} \right]$$

逻辑概率和统计概率的区别

1. 逻辑概率是用“ $\in$ ”定义的，比如 $T(\theta_j)=P(X \in \theta_j)$;
而统计概率是用“ $=$ ”定义的， 比如 $P(y_j)=P(Y=y_j)$.
2. 统计概率是归一化的，而逻辑概率不是归一化的
比如一群不同年龄的人，有标签 y_0 =“未成年人”， y_1 =“成年人”，
 y_2 =“年轻人”， y_3 =“老年人” ...
 $T(y_0) + T(y_1) = 1$ ， 所有标签的逻辑概率之和大于1.
而它们被选择的概率之和等于1: $P(y_0)+P(y_1)+P(y_2)+\dots=1$

给定年龄20岁，选择不同标签的概率（条件概率）相加等于1；而真值相加
 $T(\text{“成年人”} | 20) + T(\text{“年轻人”} | 20) = 2 > 1$
3. 条件逻辑概率就是真值 $T(\theta_j|x)$ ，其最大值是1（纵向归一化）.

三种贝叶斯定理

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

$$T(B|A)=T(A|B)T(B)/T(A), T(A)=T(A|B)T(B)+T(A|B')T(B') \quad (3.7)$$

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

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

Kolmogorov 概率——
逻辑概率, 一个论域 2^U

不同

第二种贝叶斯定理是 Shannon 使用的关于两个统计概率之间关系的定理。

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

$$P(x|y_j)=P(x)P(y_j|x)/P(y_j), P(y_j)=\sum_i P(x_i)P(y_j|x_i) \quad (3.9)$$

频率论 Mises 概率,
——统计概率, 两个论域 U, V

类似地, 也可以对称地求出 $P(y_j|x)$ 。

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

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

$$P(x|A_j)=P(x)T(A_j|x)/T(A_j), T(A_j)=\sum_i P(x_i)T(A_j|x_i) \quad (3.10)$$

两种概率+模糊逻辑
两个论域

$$T(A_j|x)=T(A_j)P(x|A_j)/P(x), T(A_j)=1/\max[P(x|A_j)/P(x)] \quad (3.11)$$

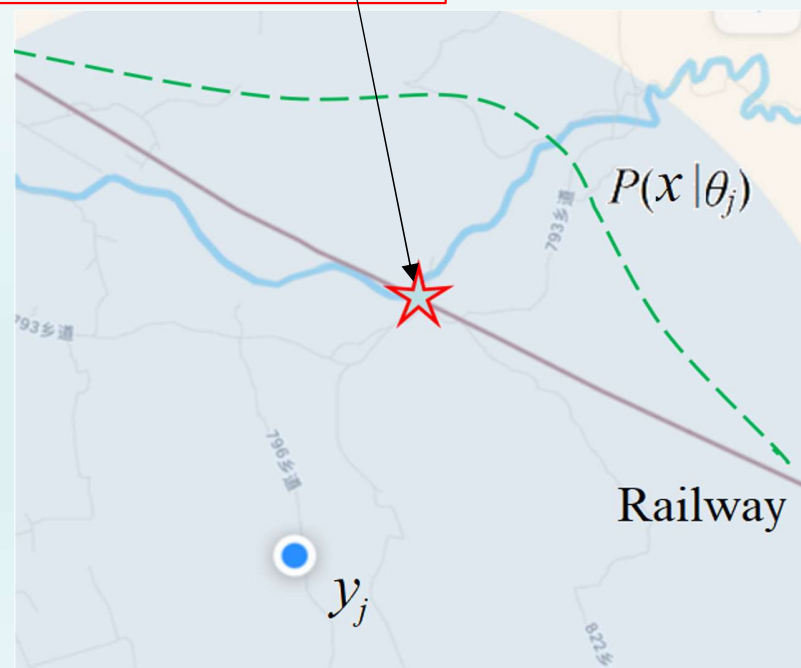
从GPS或北斗定位指针看语义贝叶斯公式

GPS圆圈的数学表示？——是模糊范围，不是概率分布，只能用真值函数或隶属函数表示： $T(\theta_j|x)=\exp[-*]$ 高斯函数

语义贝叶斯公式反映人脑思维： $P(x|\theta_j) = P(x)T(\theta_j|x)/T(\theta_j)$,

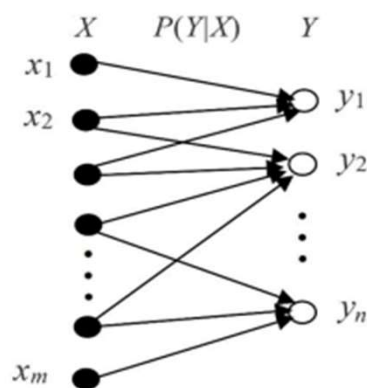


先验：在动车上。
最大可能位置？

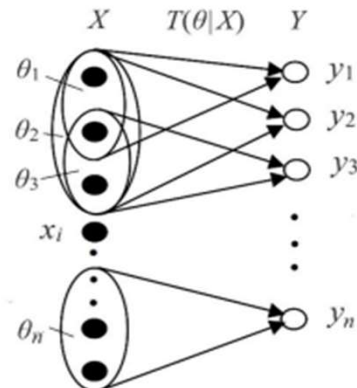


基本思想：香农信道和语义信道相互匹配

点-点



(a) 香农信道。



(b) 语义信道。

点 $\in$ 模糊集合 -- 点

比如标签“老年人”对应年龄上的一模糊集合

香农信道

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

语义信道

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

后面讲到：固定香农信道，令语义信道匹配香农信道，即

$$T(\theta_j|x) \propto P(y_j|x), j=1,2,\dots$$

语义信息量大最大。反过来，则信息效率G/R达最大。

语义信息测度——来自香农信息测度的改进

香农互信息的核心（香农信息量公式）：

$$I(x; y_j) = \log \frac{P(x | y_j)}{P(x)} = \log \frac{P(y_j | x)}{P(y_j)}$$

Carnap 和 Bar-Hillel's 的语义信息测度：

$$I_{CB} = \log[1/m_p] = \log(1/\text{逻辑概率})$$

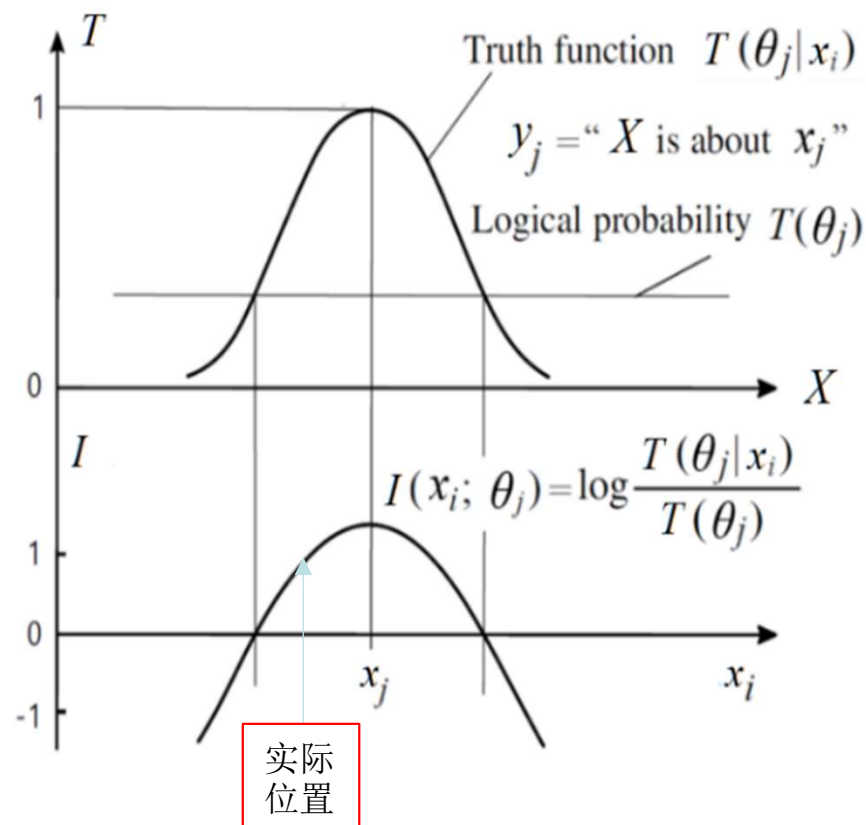
我的语义信息测度：

带进似然函数

$$I(x; \theta_j) = \log \frac{P(x | \theta_j)}{P(x)}$$
$$= \log \frac{T(\theta_j | x)}{T(\theta_j)} = \log \frac{\text{真值}}{\text{逻辑概率}}$$

带进真值函数或误差

偏差越小信息量越大，逻辑概率越小信息量越大。语义信息需要事实检验——用真值函数就行。比如**GPS**指针的语义信息——用实际位置检验。



语义KL (Kullback-Leibler) 信息和最大似然准则

给定 y_j 为真时的平均语义信息或语义KL信息（广义KL公式）：

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

求平均一定要用统计概率！

Log似然度和交叉熵之间的关系：

$$\begin{aligned} \log P(\mathbf{X}_j | \theta_j) &= \log P(x(1), x(2), \dots, x(N) | \theta_j) = \log \prod_i P(x_i | \theta_j)^{N_{ji}} \\ &= N_j \sum_i P(x_i | y_j) \log P(x_i | \theta_j) = -N_j H(X | \theta_j). \end{aligned}$$

语义KL信息和似然度的关系：

$$I(X; \theta_j) = \sum_i P(x_i | y_j) \log \frac{P(x_i | \theta_j)}{P(x_i)} = -H(X | \theta_j) + \sum_i P(x_i | y_j) \log P(x_i)$$

所以最大语义KL信息等价于最小交叉熵，也等价于最大似然度。

语义互信息和信道匹配

再对 y 求平均, 得到语义互信息:

$$I(X; Y_\theta) = \sum_j \sum_i P(y_j) P(x_i | y_j) \log \frac{T(\theta_j | x_i)}{T(\theta_j)} = H(Y_\theta) - H(Y_\theta | X),$$

覆盖熵:
$$H(Y_\theta) = - \sum_j P(y_j) \log T(\theta_j)$$

模糊熵 (等于平均失真):
$$H(Y_\theta | X) = - \sum_j \sum_i P(x_i, y_j) \log T(\theta_j | x_i)$$

语义信道匹配香农信道, 即

优化

$$T(\theta_j | x) = \frac{P(y_j | x)}{\max_x (P(y_j | x))} = \frac{P(x | y_j)}{P(x)} \bigg/ \max_x \left[\frac{P(x | y_j)}{P(x)} \right] \propto P(y_j | x),$$

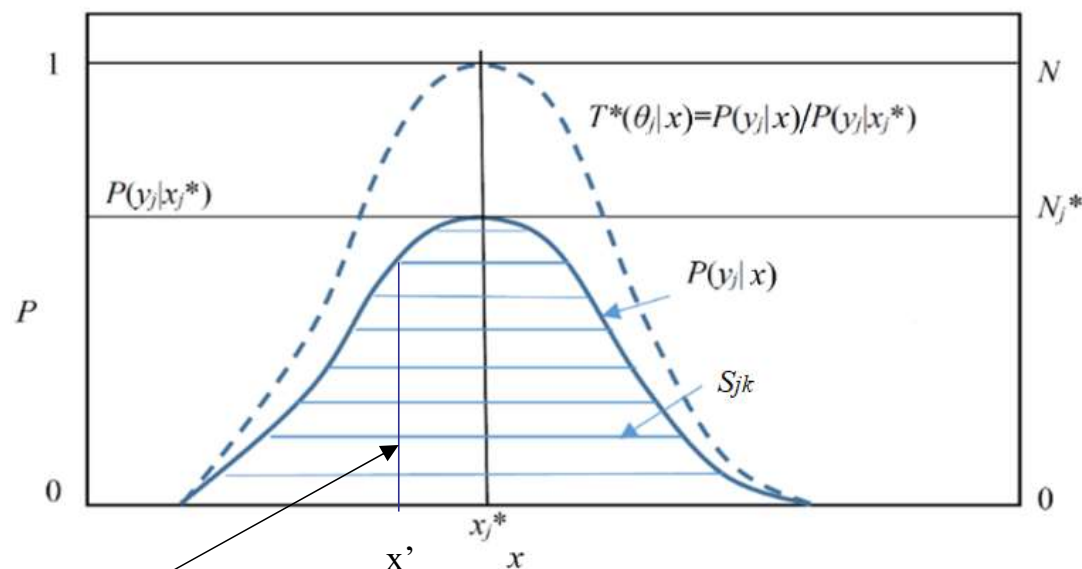
或 $P(x | \theta_j) = P(x | y_j)$

时语义信息量最大, 等于KL信息或香农互信息。

上式兼容汪培庄教授的随机集落影理论。

从随机集落影到随机点落影

- 随机集落影：隶属函数可看作一个随机集合 S_j 的无数取值 s_{jk} , $k=1,2$ 投影在论域上产生的：



x' 在模糊集 θ_j 中的隶属度，就是 x' 落在随机集 S_j 中的概率：

N_j^* ：落影最厚的地方

$$\begin{aligned}
 m_{\theta_j}(x) &= P(x \in S_j) = \frac{1}{N_j^*} \sum_k P(x \in S_{jk}) \\
 &= \left[\frac{1}{N} \sum_k P(x \in S_{jk}) \right] / \frac{N_j^*}{N} = P(y_j | x) / \max_x P(y_j | x) \\
 &= T(\theta_j | x)
 \end{aligned}$$

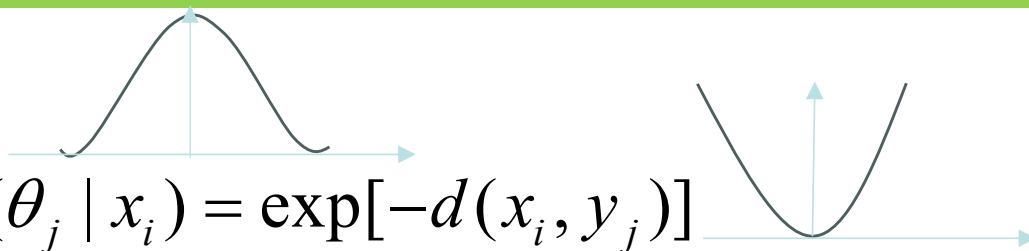
想象：用若干点凑一个集值

语义互信息和平均失真之间的关系

真值函数和失真函数之间的关系：

$$d(x_i, y_j) = -\log T(\theta_j | x_i),$$

$$T(\theta_j | x_i) = \exp[-d(x_i, y_j)]$$



一个损失函数也确定了一个语义信道。

比如对于GPS或北斗定位器指针的语义或真值函数是：

$$T(\theta_j | x) = \exp[-(x - x_j)^2 / \sigma^2 / 2]$$

相对失真函数就是

$$d(x, y_j) = (x - x_j)^2 / \sigma^2 / 2$$

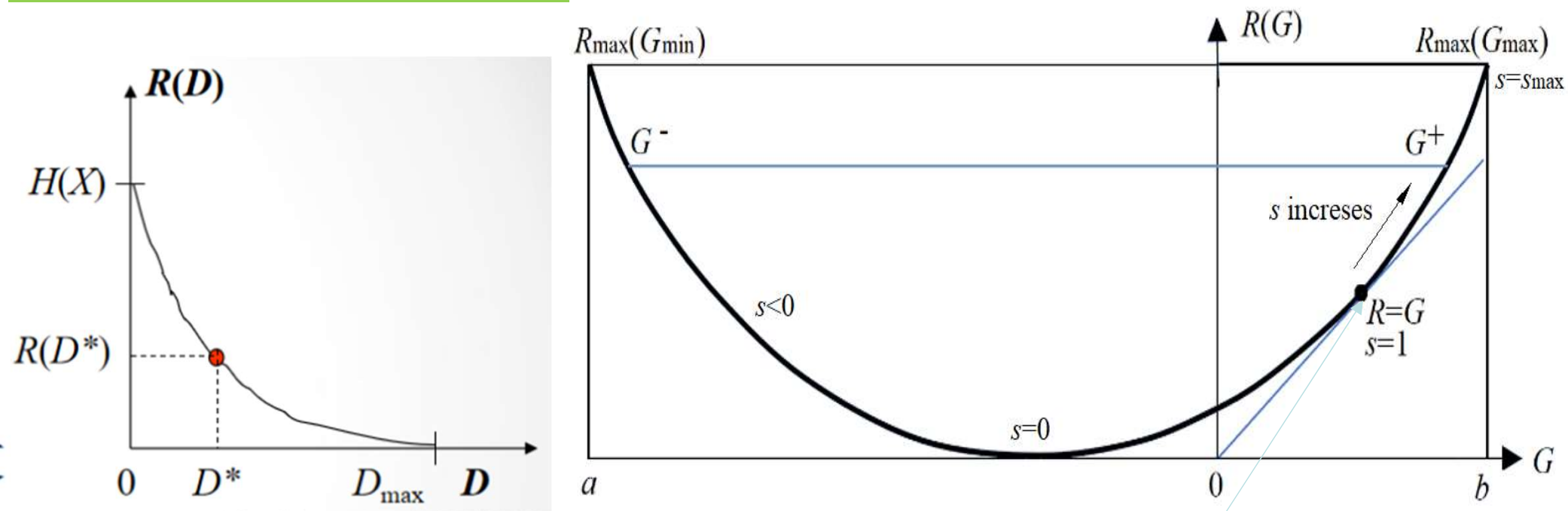
所以语义互信息可以写成：

$$I(X; Y_\theta) = H(Y_\theta) - \sum_j \sum_i P(x_i, y_j) d(x_i - y_j)$$

这就是负的正则化误差，可见最大语义互信息准则兼容**RLS**(正则化最小误差平方)准则。

语义互信息和香农互信息的匹配关系

信息率失真 $R(D)$ 函数的推广： $R(G)$ 函数



$$T(\theta_j | x) \propto P(y_j | x), \text{ 或 } P(x | \theta_j) = P(x | y_j)$$

信息率失真 $R(D)$ 函数
给定 D 的最小互信息 R

信息率保真度函数 $R(G)$ ——给定 G 的最小互信息
 $G(R)$ 表示给定 R 时 G 的上下限。

总有 $I(X;Y) \geq I(X;Y_\theta)$, 所有可能信息对只能出现在碗中。任何 $R(G)$ 函数都是碗状的。其中有一个 $R(G)=G$ 的点 ($s=1$)——两个信道相互匹配点。信息效率 $G/R=1$,达最大。

$R(G)$ 函数的参数解:

$$G(s) = \sum_i \sum_j I_{ij} P(x_i) P(y_j) 2^{s I_{ij}} / \lambda_i = \sum_i \sum_j I_{ij} P(x_i) P(y_j) m_{ij}^s / \lambda_i,$$

$$R(s) = sG(s) - \sum_i P(x_i) \log \lambda_i, \quad m_{ij} = T(\theta_j | x_i) / T(\theta_j), \quad \lambda_i = \sum_j P(y_j) m_{ij}^s,$$

有监督学习：多标签学习和分类

- 流行的方法是Binary Relevance (Zhang, M.L.; Li, Y.K.; Liu, X.Y., Geng, X. Binary relevance for multi-label learning: an overview), 它把一个多标签学习任务转换成多个两标签学习任务。原因是参数化 $P(y|x, \theta)$ 的归一化要求:

$$P(\theta_1|x) + P(\theta_2|x) + \dots = 1, \text{ 对于每个 } x$$

很难满足。而用真值函数没有归一化要求, 不需要转换。如果样本足够大, 我们直接得到(随机点落影):

$$T(\theta_j | x) = P(y_j | x) / \max_x (P(y_j | x)) = \frac{P(x | y_j)}{P(x)} \bigg/ \max_x \left(\frac{P(x | y_j)}{P(x)} \right)$$

- 如果样本不够大, 我们用语义KL公式:

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

- 分类时, 我们用最大语义信息分类器

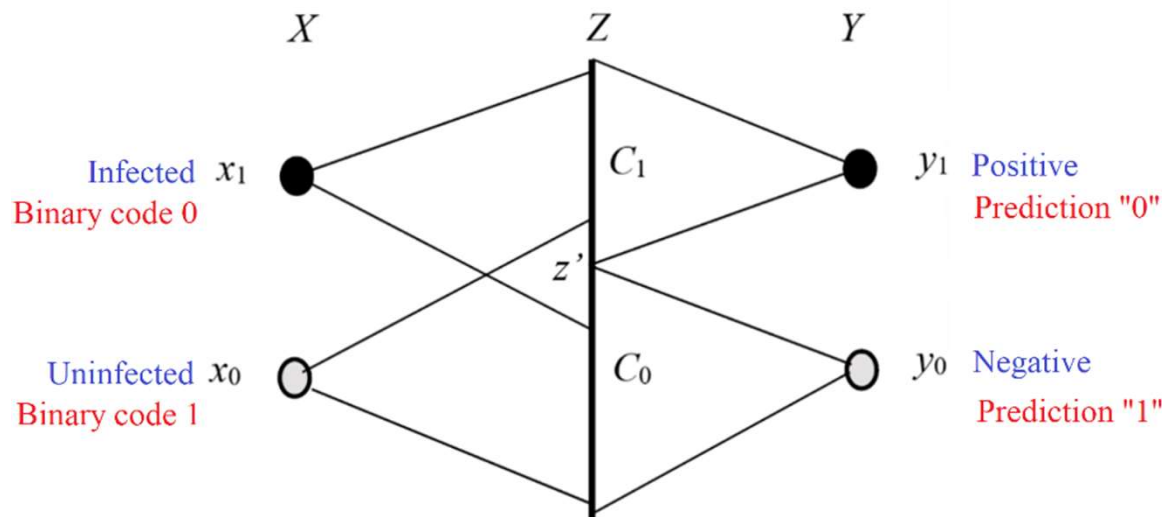
$$y_j^* = h(x) = \arg \max_{y_j} I(x; \theta_j) = \arg \max_{y_j} \log \frac{T(\theta_j | x)}{T(\theta_j)}$$

- 或最大真值分类器: $y_j^* = h(x) = \arg \max_{y_j} T(\theta_j | x) = \arg \min_{y_j} d(x, y_j)$

- 用 $P(\theta_j|x)$ 或 $P(y_j|x)$ 分类的问题: 考虑 $x=17$ 岁, $y_j=$ “ x 是未成年年轻人”, 虽然 $P(y_j|x)$ 很小, 但是 $T(\theta_j|x)$ 很大。

半监督学习：不可见实例的最大互信息分类

- 一个简单例子——说明医学检验和信号检测：



也就是最大似然比检验

- 给定 $P(x)$ 和 $P(z|x)$ ，任务是发现带来最大互信息的划分 S^* (如 z' 所示)：

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

- 我们先假设一个划分, 然后做下面迭代。
- 匹配1**（语义信道匹配香农信道）：首先我们根据给定的 S 获得香农信道：

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

最大互信息分类举例

- 然后从 $P(y|x)$ 能得到 $T(\theta_j|x)$, $T(\theta_j)$ 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:

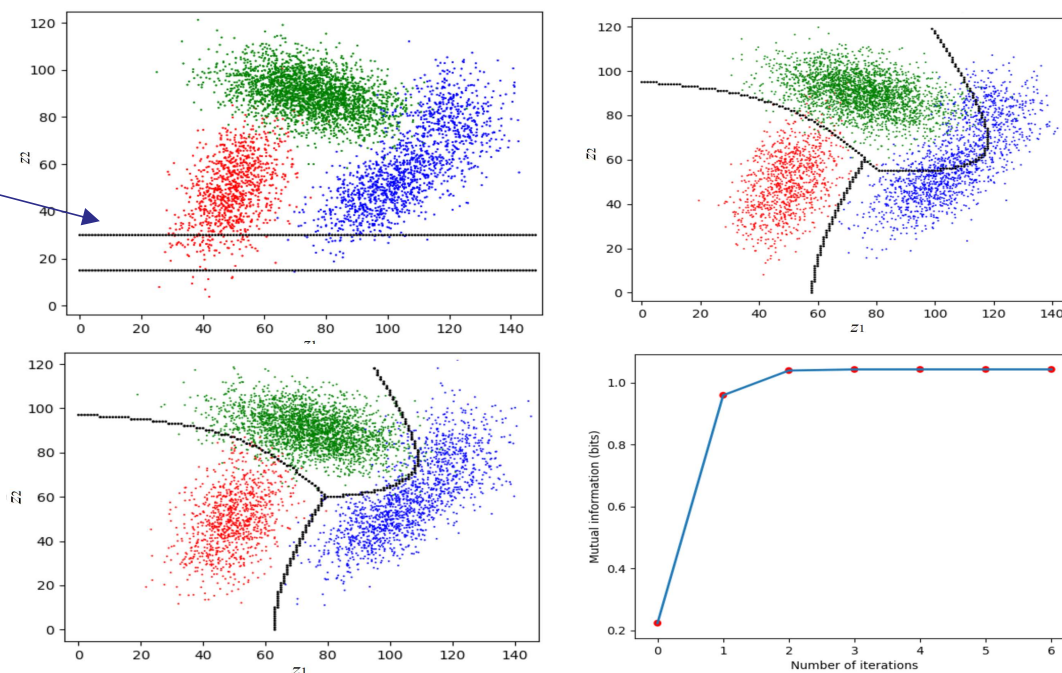
$$I(X; \theta_j | z) = \sum_i P(x_i | z) I(x_i; \theta_j) \quad , j=0,1,\dots,n.$$

- 匹配2** (香农信道匹配语义信道): 通过下面分类器得到信道划分:

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

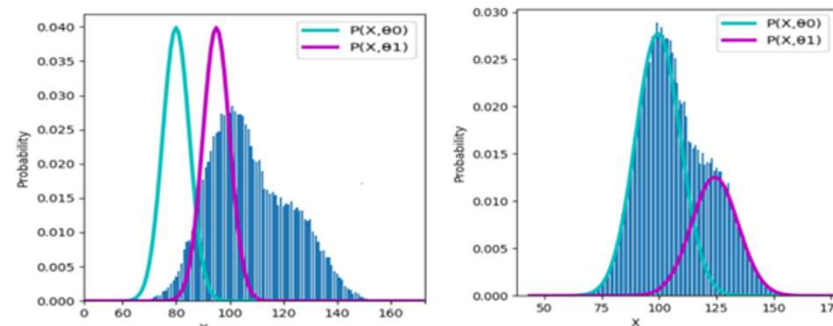
- 重复匹配1和匹配2直至 S 不再变化。收敛的 S 就是我们要找的 S^* 。
- 一个例子:

初始划分很糟糕



无监督学习：证明混合模型收敛

- 样本分布 $P(x)=\sum_j P(y_j)P(x|y_j)$ ，右边混合不可见。建立混合模型 $P_\theta(x)=\sum_j P(y_j)P(x|\theta_j)$ 逼近 $P(x)$ ，使得相对熵 $H(P\|P_\theta)$ 接近0。



EnM算法（类似于EM算法）：先假设一个 $P(y)$

匹配1：令香农信道匹配语义信道，通过重复下面两个公式 n 或更少次：

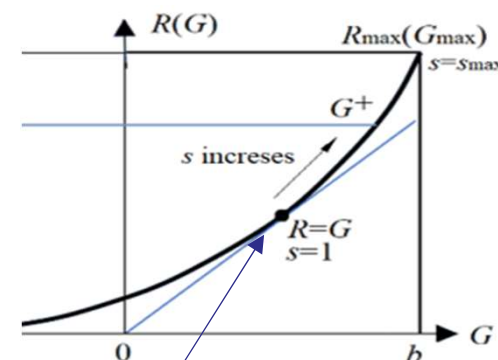
E步 $P(y_j | x) = P(y_j)P(x | \theta_j) / P_\theta(x)$, $P_\theta(x) = \sum_j P(y_j)P(x | \theta_j)$,

这一步是从M步移来的 $P^{+1}(y_j) = \sum_x P(x_i)P(y_j | x_i)$.

匹配2：令语义信道匹配香农信道，通过

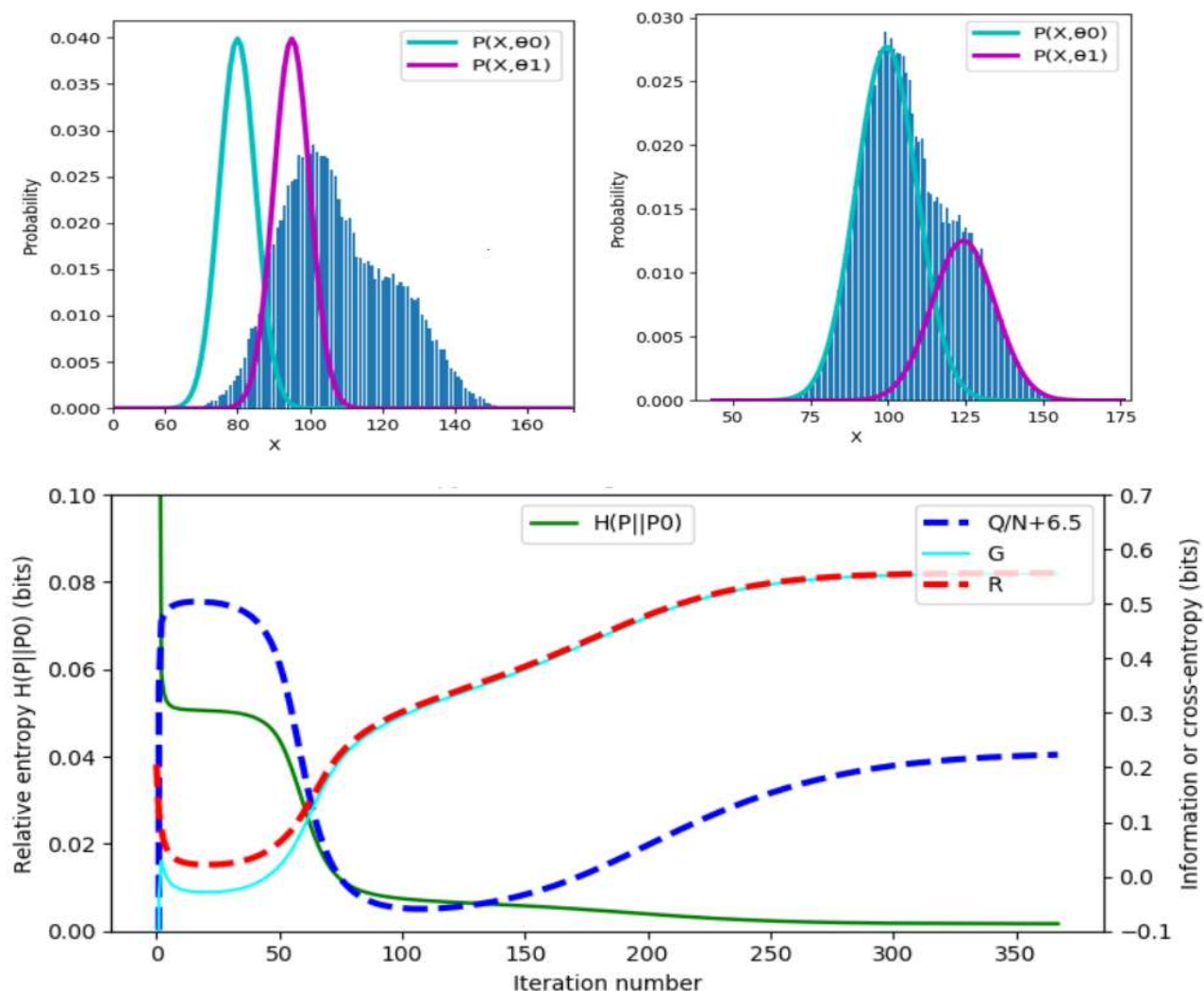
$$P(x | \theta_j^{+1}) = P(x)P(y_j | x) / P^{+1}(y_j), \quad P_\theta(x) = \sum_j P(y_j)P(x | \theta_j)$$

- 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
- $H(P \| P_\theta) = R - G + H(P^{+1}(y) \| P(y)) \rightarrow 0$
- 因为匹配2最大化 G ，且匹配1最小化 R 和 $H(P^{+1}(y) \| P(y))$ ，所以 $H(P \| P_\theta)$ 能接近0。
- 这个公式非常重要，能消除最大信息和最小自由能之间的矛盾——即最大语义互信息和最小香农互信息 R 或最大后验熵之间，并不矛盾。



证明混合模型收敛证明（续）

一个高斯混合模型的例子，用以比较EM和E3M算法，并解释收敛原因。



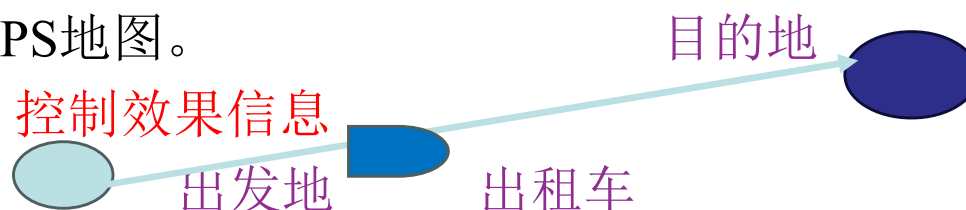
EM算法需要大约340次迭代，而E3M算法需要大约240次迭代。收敛过程中，完整数据对数似然度 Q 并不是单调增加的。 $H(P||P_\theta)$ 随 $R - G$ 减小。

强化学习：合目的信息作为奖惩函数

考虑乘客和出租车司机观察车上的GPS地图。

乘客信息：合目的信息；司机信息：控制效果信息

即合目的信息等于控制效果信息。



如果起始位置或无控制位置、目标位置都是不确定的，特别是：目标是一个模糊范围，比如下面情况，如何度量合目的信息？

- 人口死亡年龄最好超过80岁，用 $T(\theta_j|x)$ 表示；当前死亡年龄分布是 $P(x)$ 。
- 动车到达的时间误差最好在1分钟之内，用 $T(\theta_j|x)$ 表示。目前分布是 $P(X)$ 。

同样可用语义信息测度度量合目的信息：

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

而香农信息或KL信息反映控制成本：

$$I(X; a_j) = \sum_i P(x_i | a_j) \log \frac{P(x_i | a_j)}{P(x_i)}.$$

优化是改变 $P(x|a_j)$ ，让它向 $P(x|\theta_j)$ 靠拢，最小化 $R-G=I(X;a_j)-I(X;a_j/T)$ ，提高通信效率 G/R 。有了相互匹配的 $P(x|a_j)=P(x|\theta_j)$ ，再考虑最大化 G 和 R 。

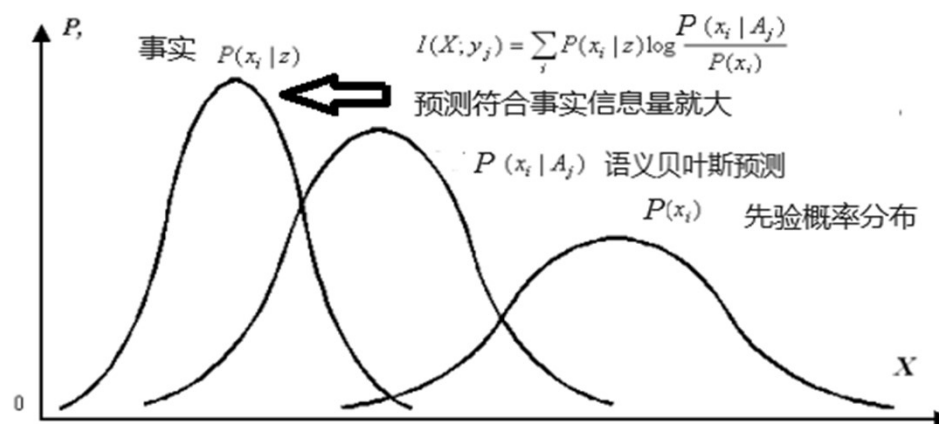
陈述句和祈使句：两种语义信息的异同

都用语义KL公式：

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

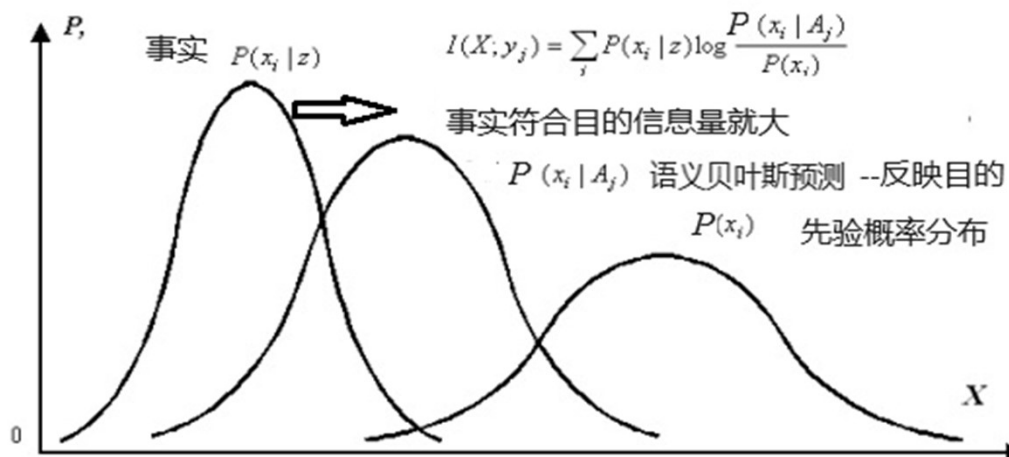
陈述句（天气预报）的语义信息：

希望预测符合事实



祈使句（指令）的语义信息（合目的信息）：

希望事实符合目的

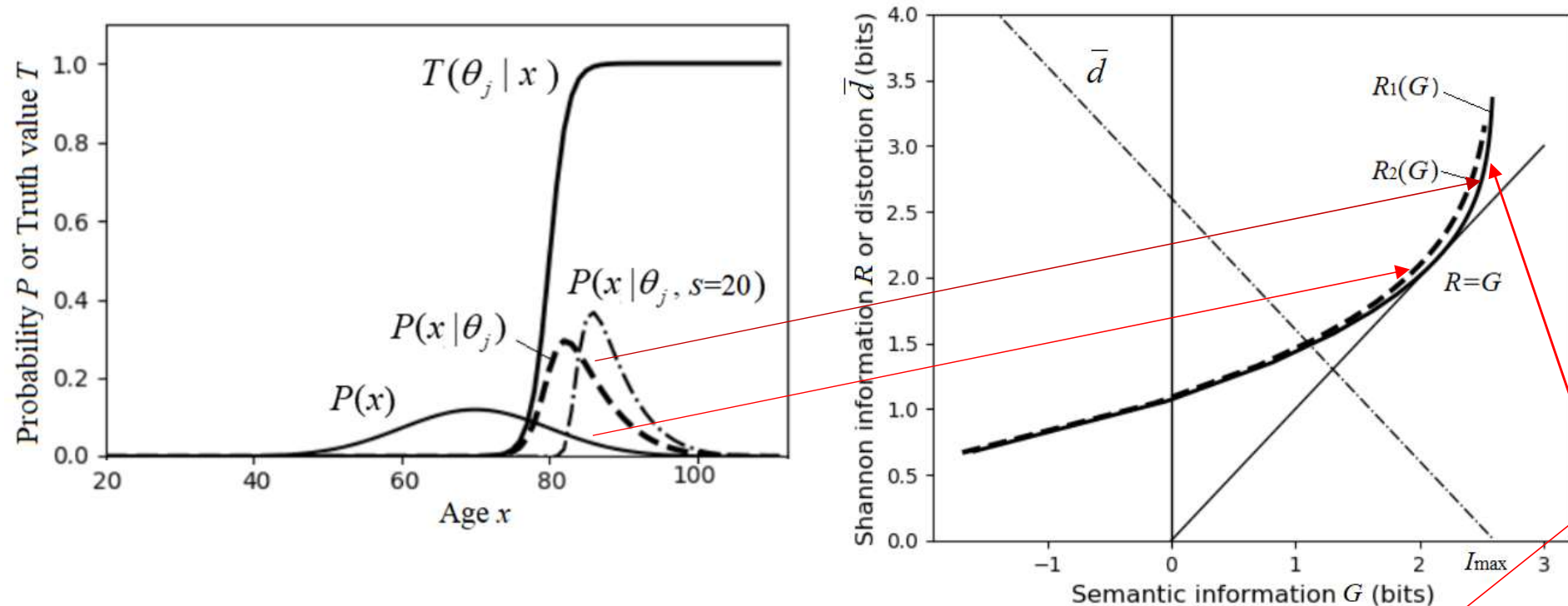


合目的信息作为奖惩函数 (续)

- $R(G)$ 函数的参数解提供了权衡合目的信息 G 和效率 G/R 的方法。仿照 $R(G)$ 函数中 $P(y_j|x)$ 的参数表达式, 我们得到最大化合目的信息且最小化KL信息的 $P(x|a_j)$:

$$P(x_i | a_j) = P(x | \theta_j, s) = P(x_i) T(\theta_j | x_i)^s / \sum_k P(x_k) T(\theta_j | x_k)^s.$$

比如: 控制目标是改变医疗条件使平均死亡年龄接近或超过80岁。



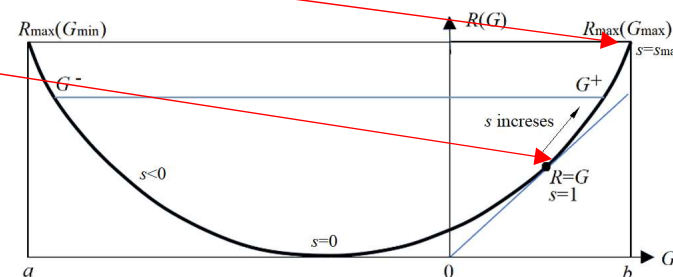
两个 $R(G)$ 函数: $R_1(G)$ 使用了 $P(x|a_j)=P(x|\theta_j, s)$; $R_2(G)$ 使用了 $P(x|a_j) = P(x|\beta_j, s)$ ——它是正态分布. 说明 s 太大 R 太大时, 效率 G/R 较低, 选在 s 在5-10之间合适。 $P(x|a_j)$ 的期望在82岁, G 接近最大了。如果是多目标, 则要用语义互信息, $P(a)$ 也要优化

深度学习新潮流——使用相似函数和估计互信息

- 不久，深度学习(自监督学习)领域出现了以InfoNCE(信息噪声对比估计)和MINE(最大神经信息估计)为代表的使用相似函数和估计互信息的潮流。笔者已经在一篇评论文章中说明：**相似函数是真值函数的特例，估计互信息就是语义互信息的特例。**
- 求解隶属函数的公式也能用来求解相似函数：

$$T(\theta_j | x) = P(y_j | x) / \max_x (P(y_j | x)) = \frac{P(x | y_j)}{P(x)} \bigg/ \max_x \left(\frac{P(x | y_j)}{P(x)} \right)$$

- 流行的信息瓶颈解释忽视了估计互信息和香农互信息的明显区别。用 $R(G)$ 函数解释深度学习更加简单：预训练阶段最大化信息效率，如求混合模型，使 G 逼近 R ， $s \approx 1$ ；微调阶段就是增大 s ，提高分类的正确率。



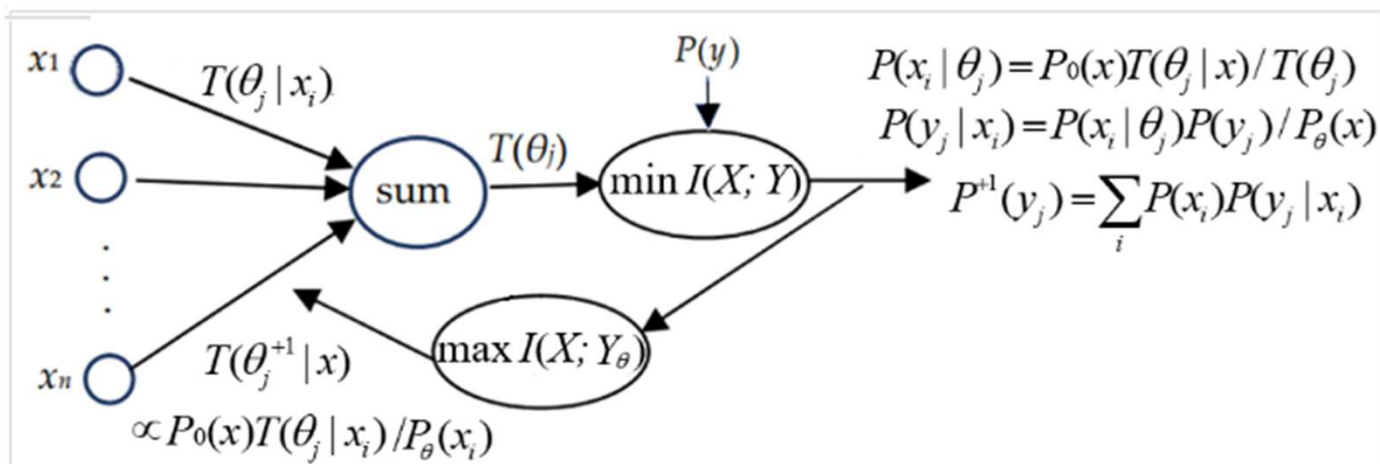
- 参看：
- [回顾学习函数和语义信息测度的进化——进而理解深度学习](#) [Reviewing Evolution of Learning Functions and Semantic Information Measures for Understanding Deep Learning \(survivor99.com\)](#)

值得探讨方向1——用信道混合模型机代替有限波尔茨曼机

混合模型和RBM类似，用于无监督学习即聚类。但是混合模型不需要考虑梯度和反向传播。现在我们可以用RBM用的负指数函数表示相似函数或真值函数 $T(\theta_j|x)$ ，和 $P(x)$ 一起产生似然函数，然后用混合模型的EM或EnM算法优化模型参数：

$$T(\theta_j^{+1} | x) = \exp \left[\frac{-(x - \mu_j)^2}{2\sigma_j^2} \right] \propto \frac{P(y_j | x)}{P(y_j)} = \frac{P_0(x)T(\theta_j | x)}{T(\theta_j)P_\theta(x)} \propto \frac{P_0(x)T(\theta_j | x)}{P_\theta(x)}$$

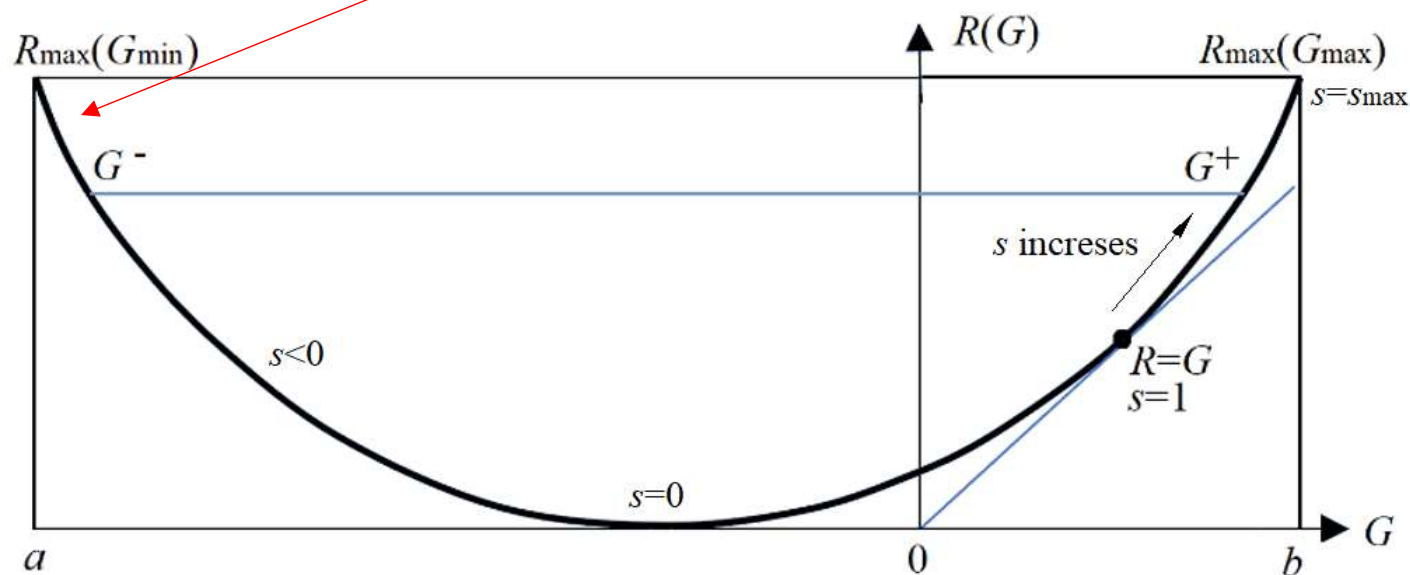
即用右边的期望和标准偏差作为左边的期望和标准偏差。
这样得到信道混合模型机可以取代RBM，无需考虑梯度。



值得探讨方向2——语义加密造成负信息

流行的加密系统以解码可读性作为成功解码标准。但是，我们也可以设想一种语义加密——解码得到的可读明文有多种，只有一种是对的——提供正的信息，其他都是错的——提供负的信息，如蒋干盗书。语义信息测度和 $R(G)$ 函数提供了最小化语义信息方法，有助于产生最大负信息的加密和误解密。这样的系统有待研究。

西方有人研究：<https://mp.weixin.qq.com/s/xD3amg5gRfuWbNzmm5usYw>
但是没用语义信息论指导。



结束语

- 理论要点:

- 1) P-T概率框架融合了逻辑概率、统计概率和模糊集合，表达力更强；
- 2) 把似然函数和真值函数带进信息量公式，语义信息测度因而可用于机器学习；
- 3) $R(G)$ 函数描述了语义互信息和香农互信息之间的匹配关系；
- 4) 通过语义信道和香农信道相互匹配，我们能解释和改进机器学习。

- 需要改进:

- 1) 为解释和改进深度学习，我们还需要结合深度神经网络中许多有效的特征抽取方法做进一步研究。
- 2) 本讲稿为范畴论和AI讨论组准备。我不太懂范畴论，但是本讨论组关心人工智能和语义信息论。对于把范畴论用到语义信息论和机器学习的努力，上述理论和方法或许能起到桥梁作用。如何架桥有待努力。

谢谢各位听讲！欢迎批评和交流！

Thanks for your listening! Welcome to exchange ideas.

鲁晨光的语义信息论和机器学习网页 <http://survivor99.com/>

<http://www.survivor99.com/lcg/CM/Recent.html>