SciSci研讨会讲稿，2025-10-5
本PPT已经以PDF格式放在鲁晨光主页 http://survivor99.com

语义信息G理论、应用和最新进展

G 理论即推广香农理论得到语义信息论。
该理论继承了Shannon，Popper， Zadeh 和 Fisher的思想，
它能用于机器学习，约束控制，投资组合，贝叶斯确证…，
最新进展是把Friston的最小自由能原理改进为最大信息效率原理。

鲁晨光
Email: Survival99@gmail.com
Homepage: http://survivor99.com

信息科学vs信息论——区分内容信息和形式信息

1.　区分内容信息和形式信息：

“早上7点”提供信息，包含两种：

- “7点”的物理学含义——太阳升起，小孩起床上学。这是内容信息。
- “现在几点”，“7点！”区别于其他时间。提供形式信息。

再比如说大人手指苹果教小孩说“这是苹果”，这是提供内容信息。

“那是什么？”“是苹果！”这是提供形式信息。

苹果的感觉表象对于符号“苹果”来说是内容，对于物理的苹果来说又是符号或其组合。

怎样区分：内容信息——学习知识——发送编码定义——不可量化——一辈子只学一次。

形式信息——知道什么发生——发送编码——可量化——一辈子无数次。

信息定义：两者都是告诉我们以前不知道的东西，或特殊性。

2. 信息论和语义信息论研究的是形式信息，　不是内容信息。

简单说来，用了log就是信息论，不用log就不是信息论。

我建议：**把信息论理解为关于形式信息的信息量的理论。**

用信息科学表示更一般的关于信息的理论（钟义信和李宗荣教授这样做了）。

3. 信息科学包含：

——信息哲学；

——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：信息论的应用；获取信息内容的理论（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；…）以及关于信息传播的理论（包括人文社科）。

语义信息VS香农信息

- 什么是日常生活中的信息或语义信息，我的回答和Floridi类似：

日常语言，天气预报，时钟，电子地图上的GPS指针，温度表，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。

- 香农信息和语义信息，两者相同的是：都是消除不确定性,或被反映的特殊性。
- 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（比如：“大雨”，“老年人”）。

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。

GPS 定位提供信息。定位有误差，定位中心并不是最大可能位置

各种语义信息论

自从Weaver提出信息的三个层次，出现了许多语义信息论。ChatGPT回答：

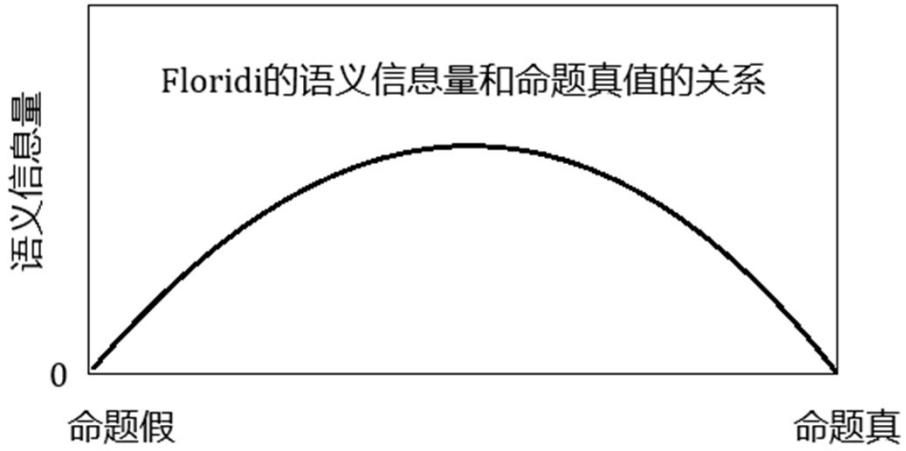
1. Carnap和Bar-Hillel：

语义信息量= log(1/逻辑概率)

贯彻了波普尔的思想：命题函数的逻辑概率越小， 信息量越大。

2. Dretske: 关于信息和知识关系的理论；没有量化。
3. Floridi的强语义信息理论， 强调**真**对信息量的贡献，公式比较复杂；信息量图解：

问题：没有区分命题的真值和命题函数的逻辑概率。



4. 鲁晨光的语义信息G理论。

同时使用逻辑概率和真值。

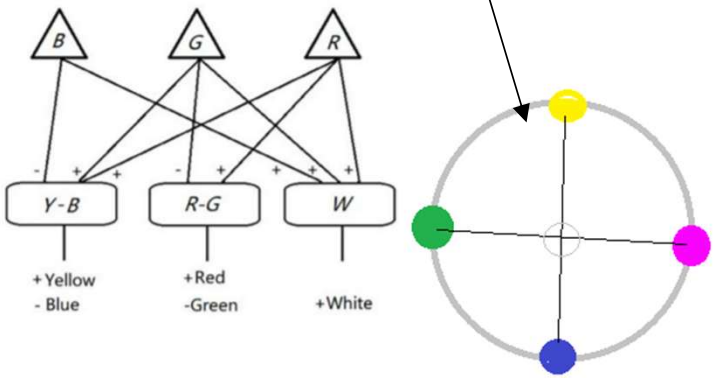
5.中国其他人的语义信息论：

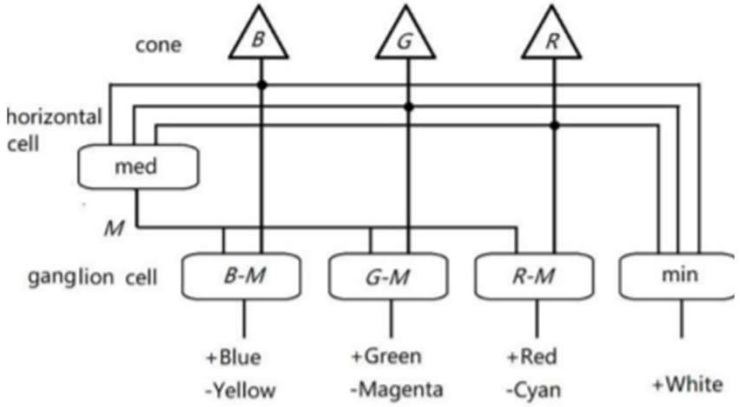
钟义信的全信息理论只用真值，我有文章讨论：《钟义信的全信息理论存在的问题》，http://www.hgacademy.cn/index.php/View/478.html

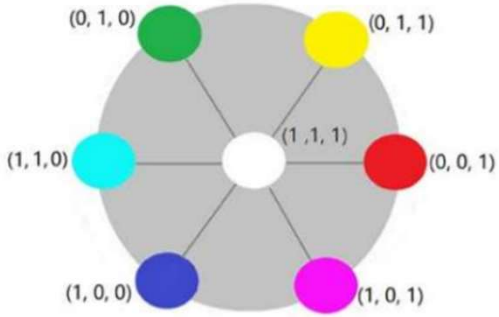
牛凯和张平的基于同义映射的语义信息理论。去年新提出，问题正在被逐渐发现。

我的研究起因：色觉机制3-8译码模型

- 经典的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
- 流行的阶段模型Zone Model　　我的3-8译码模型

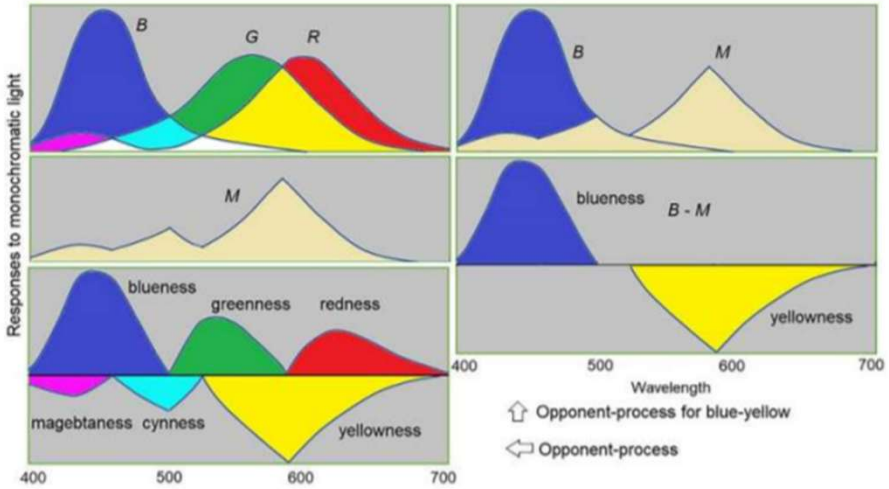




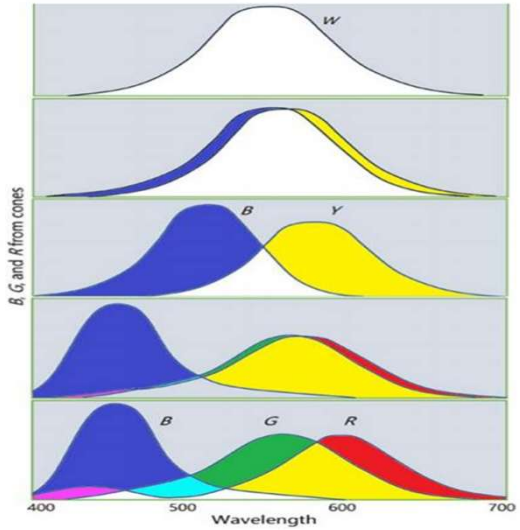


- 我提出3-8译码模型，它是对称的，解释互补处理，色盲和色觉进化更好，更适合电脑图像的颜色转换（RGB和HSL系统之间）。文章于1989《光学学报》发表

互补处理



色觉进化



怎样度量色觉信息——说明分辨率高信息多？

- 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），如GPS指针的含义

感觉信息量=log[后验概率/先验概率]

(使用语义贝叶斯公式，后面讲到）

=log[Q(j|i)/分子平均]

=Log(Softmax函数)

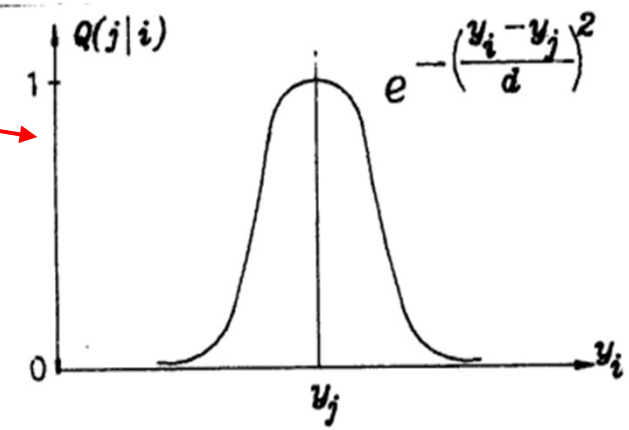


Fig.1 Confusion probability function of y_j with $y_i = y(x_i)$.

- 后来发现，把分辨率函数换成隶属函数或真值函数，分子的平均就是逻辑概率，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年发表在一个不正规的模糊数学期刊BUSEFUL上。网上有：https://www.univ-smb.fr/listic/production-scientifique/revue-busefal/version-electronique/ebusefal-44/

用真值还是逻辑概率？不同语义信息公式图解

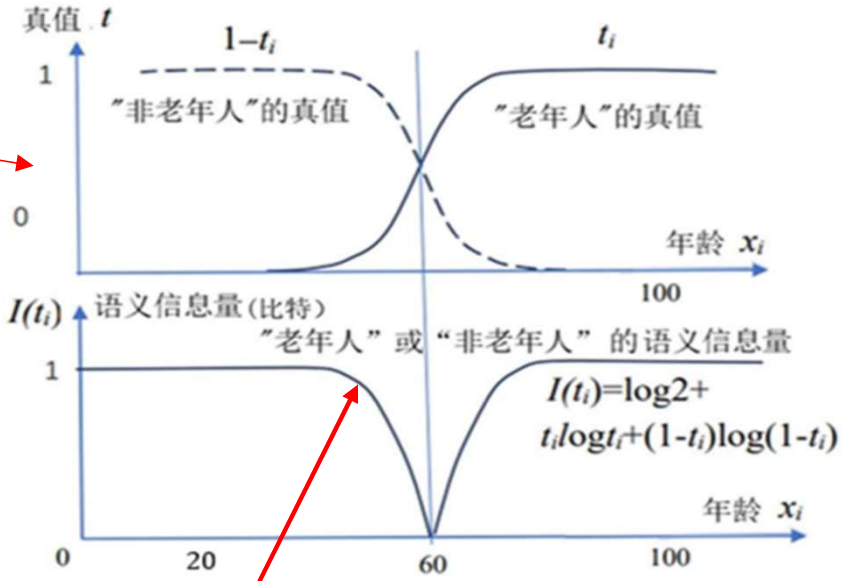
例：x是年龄，语句y是“老年人”。计算y关于x的语义信息量，它随x变化。

- 1) 用逻辑概率，Carnap和Bar-Hillel的语义信息公式（和年龄无关）：

I_{CB}=log(1/逻辑概率）

- 2）用真值和模糊熵，如钟义信：

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



- 3）同时用两者，我的语义信息公式：

比较我的语义信息公式：

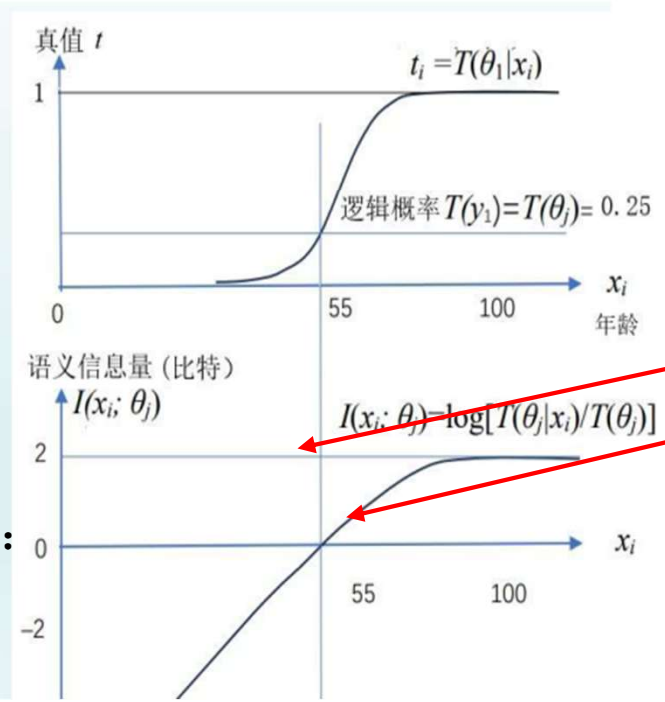
$$I(x_i;\theta_j) = \log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}$$

$$= \log[1/T(\theta_j)] - d(x_i, y_j)$$

$$= I_{CB} - \text{失真}$$

利用了真值和失真的关系：

T(θj |x)=exp(-失真)

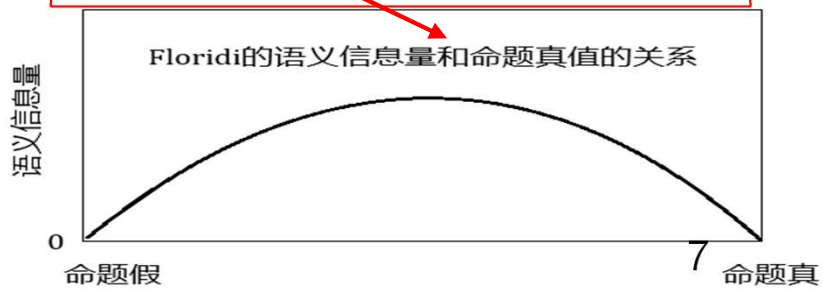


钟义信的语义信息量

Carnap和Bar-Hillel的

我的

Floridi的



- 4）Floridi的公式

如果求平均信息？用不用统计概率？

- 鲁晨光：区分并同时使用真值和逻辑概率，并加上统计概率

求平均用　　似然函数　　真值

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

统计概率　　逻辑概率

- 其他所有语义信息论求平均仍然用log后面的函数
- 比如模糊熵：

$$F(X)=-\left[t_n\log t_n+(1-t_n)\log(1-t_n)\right]$$

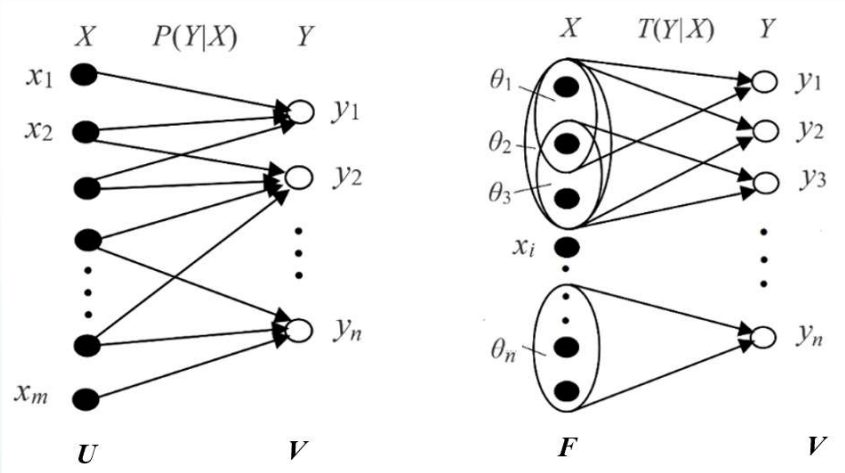
- 语义熵
- 别人的语义熵（本质上还是香农熵）：$H(Y_\theta)=-\sum_i T(\theta_j)\log T(\theta_j)$
- 我的语义熵（交叉熵，不归一化熵）：

$$H(Y_\theta)=-\sum_i P(y_j)\log T(\theta_j)$$

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。

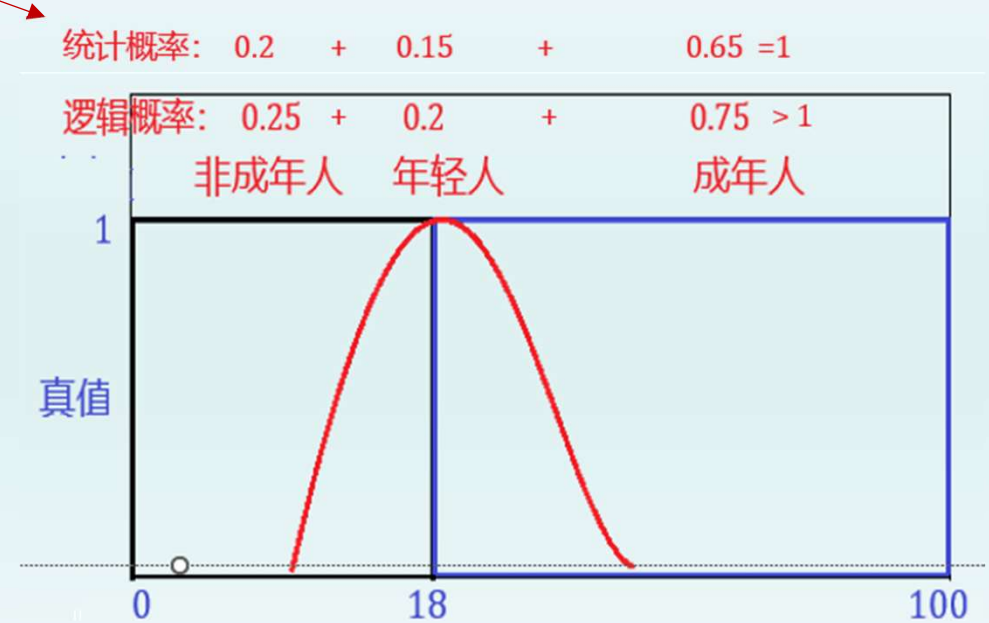
- 一个语句（命题函数或谓词）的逻辑概率是它被判断为真的概率，
- 而它的统计概率是它被选择的概率。
- 统计概率是归一化的，是用=定义的，
- 例：P(yj)=P(Y=yj);
- 逻辑概率是用“∈”定义的，
- 比如T(θj)=P(X∈θj);
- 年龄例子：逻辑概率之和大于1



香农信道P(y|x)　　语义信道T(y|x)

极端例子：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.

条件逻辑概率就是真值T(θj|xi)，其最大值是1（纵向归一化）.



语义Bayes公式

- 用真值函数表示形式语义，符合Davidson的真值条件语义学。
  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。
- 命题函数或谓词yj(x)=“x在θj中”的真值函数就是x在模糊集θj上的隶属函数：
- T(θj |x) = Mθj (x)
- 逻辑概率即真值的平均，或Zadeh的模糊事件概率：

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

- 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, 能推导出似然函数和真值函数之间的的关系：

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

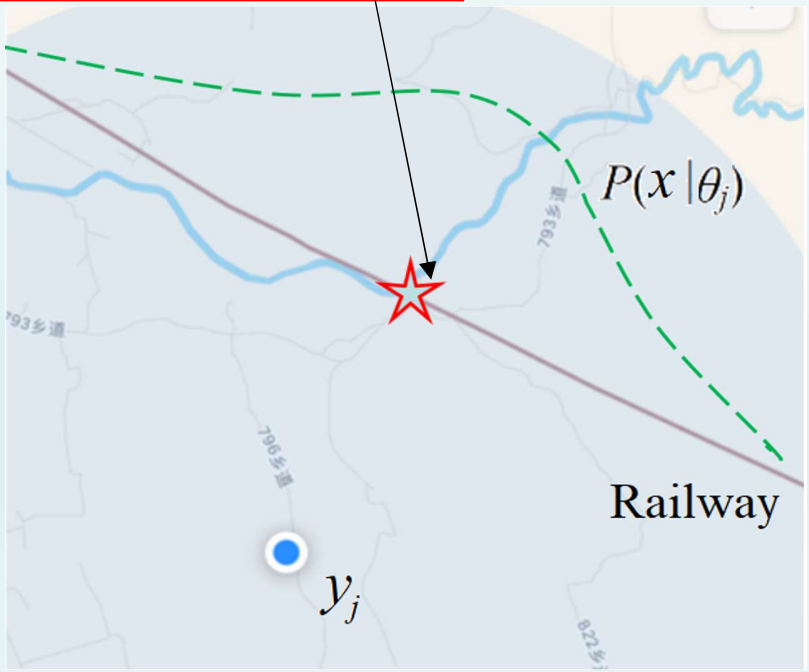
从GPS或北斗定位指针看语义贝叶斯公式

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

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

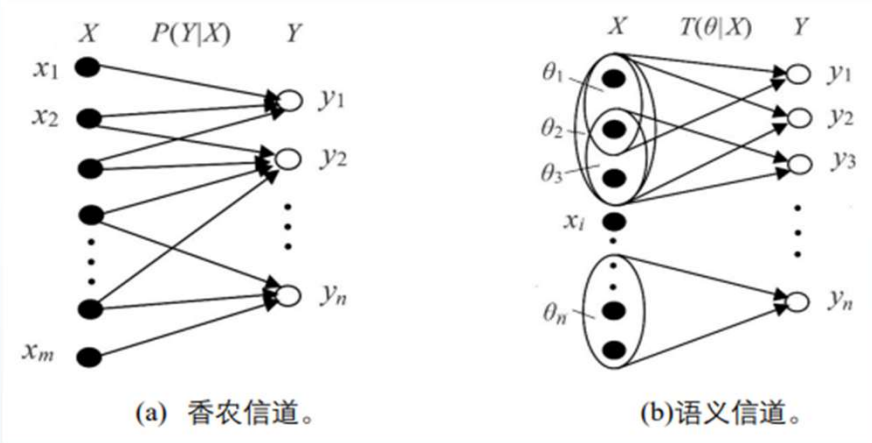
已有的数学方法不行

先验：在动车上。
最大可能位置？



优化语义通信和机器学习的基本思想：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∈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

$T(\theta j|x) \propto P(yj|x)$，或 $P(x|\theta j)=P(x|yj)$，$j=1,2,\cdots$

语义信息量达最大。反过来，则信息效率(语义信息/香农信息=G/R)达最大。

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：

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

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

Kolmogorov概率——
逻辑概率，一个论域$\mathcal{B}$(U)

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

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

频率论Mises概率，
——统计概率，香农用，两个论域U， V

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

类似地，也可以对称地求出 $P(y_j|x)$.

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

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

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

两种概率+模糊逻辑
两个论域

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

语义信息测度——来自香农信息测度的推广

香农互信息的核心（香农信息量公式）：

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

我的语义信息测度：

带进似然函数

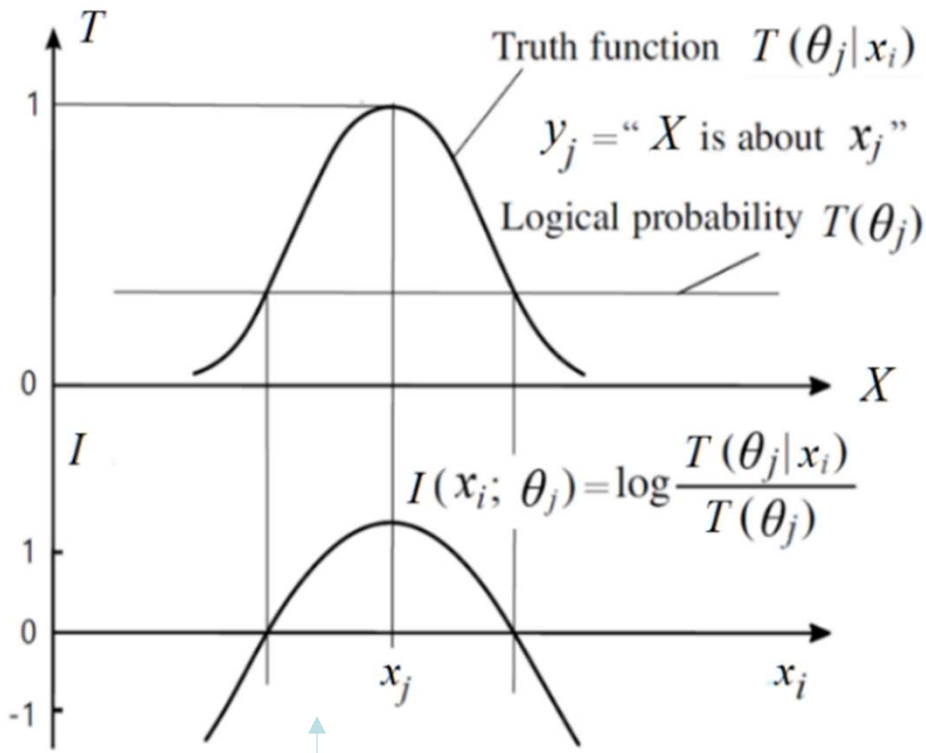
$$I(x;\theta_j)=\log\frac{P(x\mid\theta_j)}{P(x)}$$

$$=\log\frac{T(\theta_j\mid x)}{T(\theta_j)}=\log\frac{\text{真值}}{\text{逻辑概率}}$$

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。

体现了Popper思想。



实际位置

语义KL（Kullback-Leibler）信息和最大似然准则

给定yj为真，对语义信息量求平均，就得到语义KL信息（广义KL公式）：

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

Log似然度和交叉熵之间的关系：

$$\log P(\mathbf{X}_j\mid\theta_j)=\log P(x(1),x(2),\ldots,x(N)\mid\theta_j)=\log\prod_i P(x_i\mid\theta_j)^{N_{ji}}$$

$$=N_j\sum_i P(x_i\mid y_j)\log P(x_i\mid\theta_j)=-N_jH(X\mid\theta_j).$$

语义KL信息和似然度的关系：

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

所以最大语义信息准则等价于最小交叉熵准则和最大似然准则。

语义互信息和变分自由能，以及平均失真的关系

再对y求平均，得到语义互信息：

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

$= H(X)-H(X|Y\theta) = H(Y\theta)-H(Y\theta|X)$

= 信源 − 变分自由能 = 语义熵 – 模糊熵（即平均失真）

变分自由能：

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

语义熵（正则化项）：

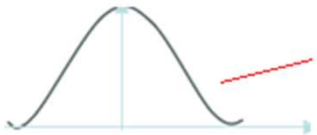
$$H(Y_\theta)=-\sum_j P(y_j)\log T(\theta_j)$$

模糊熵即平均失真（多语句）：

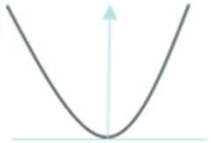
$$H(Y_\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i)$$

因为真值函数和失真函数之间的关系：

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



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



所以模糊熵等于平均失真，于是

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

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.

语义信道匹配香农信道和随机集落影理论改进

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

当P(x|θj)=P(x|yj) 或T(θj |x)∝P(yj|x)时，语义信息到批评香农信道，语义互信息达到最大值，等于香农互信息。令真值函数最大值说1， 可以解出：

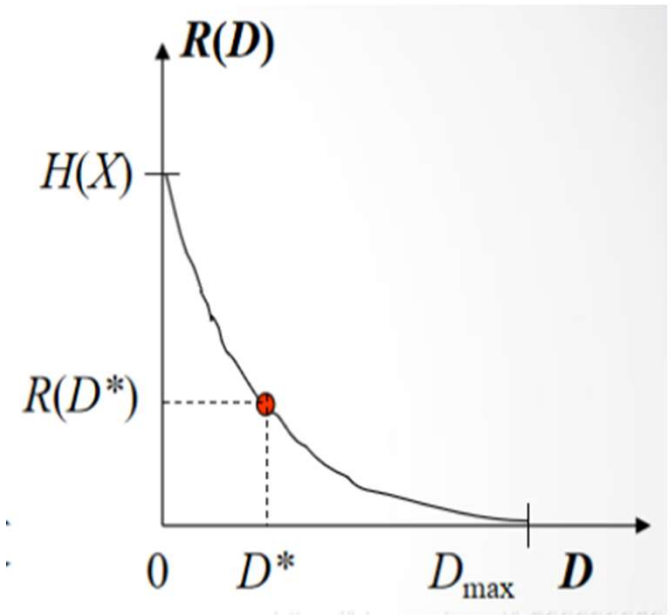
$$T^*(\theta_j|x)=\frac{P(y_j|x)}{\max\limits_x(P(y_j|x))}=\frac{P(x|y_j)}{P(x)}\Big/\max_x\left[\frac{P(x|y_j)}{P(x)}\right]\propto P(y_j|x),$$

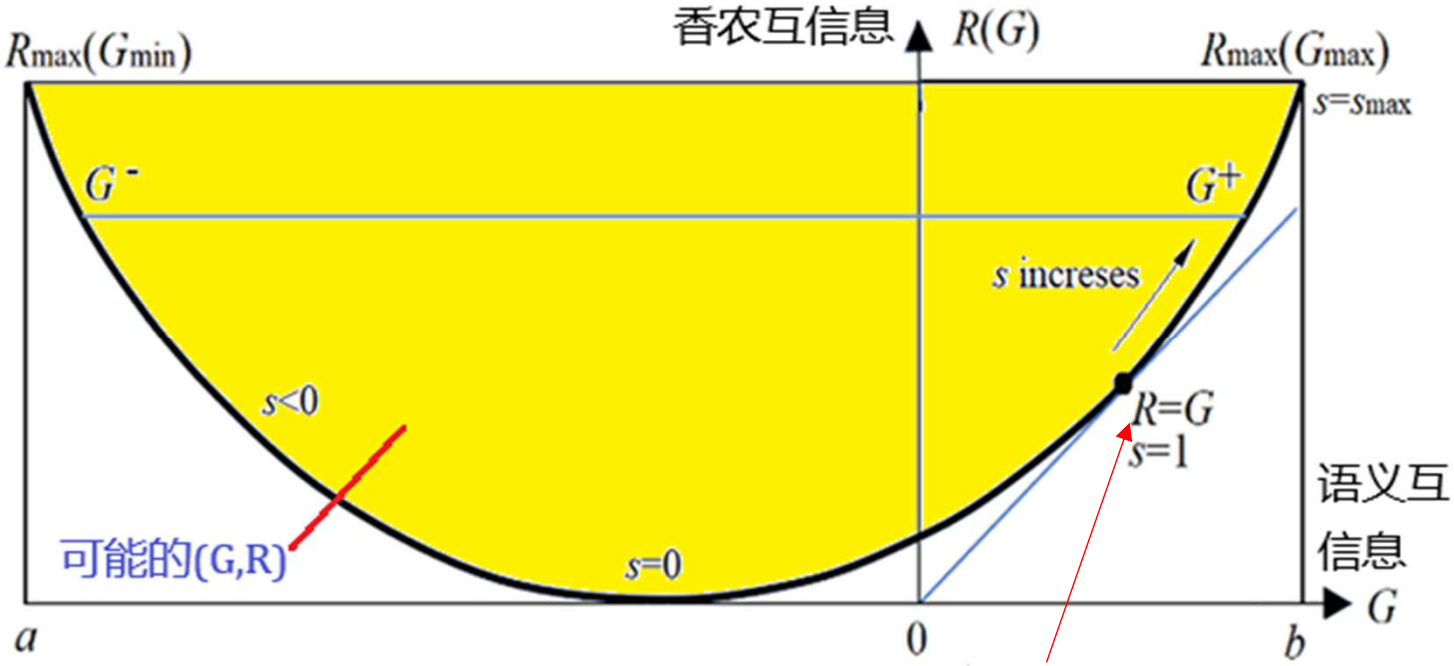
样本不够大时，用 $T^*(\theta_j|x)=\arg\max\limits_{\theta_j}\sum\limits_i P(x_i|y_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}$

上式兼容汪培庄教授的随机集落影理论——我称之为随机点落影方法。

1991年，汪培庄教授给了我在北师大做访问学者的机会。今年五月一篇合作文章发表于《智能系统学报》，他看了非常欣慰，因为他的理论终于落地——可以付诸应用。他6月去世。我在科学网博文《40年师生缘——回忆汪培庄先生》中写道：这篇文章似乎就是我做汪先生学生的使命所在。汪先生的隶属函数统计思想指导了我的论研究，我的研究反过来又发展了他的统计思想。

语义互信息和香农互信息的匹配关系
信息率失真函数*R(D)*的推广：信息率逼真函数*R(G)*





$$T(\theta_j \mid x) \propto P(y_j \mid x),\ 或\ P(x \mid \theta_j = P(x \mid y_j)$$

信息率失真*R(D)*函数
给给定D的最小互信息R

信息率逼真函数*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$I(X;Y) \geqslant I(X;Y_\theta)$，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。

求解$R(G)$函数需要的香农信道

$$P^*(y \mid x_i) = P(y)\left[\frac{T(\theta_j \mid x_i)}{T(\theta_j)}\right]^s \Big/ Z_i,\ \ Z_i = \sum_k P(y_k)\left[\frac{T(\theta_k \mid x_i)}{T(\theta_k)}\right]^s,$$

用的最小互信息迭代公式：

$$P^*(y_j) = \sum_i P(x_i)P(y_j \mid x_i).$$

我的信息价值公式

增值熵

$$H_g = \log R_g = \sum_{i=1}^{W} P(x_i)\log R_i = \sum_{i=1}^{W} P(x_i)\log \sum_{k=0}^{N} q_k R_{ik} .$$

设不同收益的先验概率分布是 $P(x)$，收益矩阵是(R_{ik})，则预计的增值熵是 $H_g(X)$. 据此可以求出最优投资比例矢量 $\boldsymbol{q}^*$，它是 $P(x)$和(R_{ik})的函数，即 $\boldsymbol{q}^*=\boldsymbol{q}^*(P(x),\ (R_{ik}))$。当预测的收益的概率分布变为 $P(x|\theta_j)$时，增值熵变为

$$H_g{}^*(X\,|\,\theta_j) = \sum_i P(x_i\,|\,\theta_j)\log R_i(\boldsymbol{q}^*),\quad R_i(\boldsymbol{q}^*) = \sum_k q_k{}^* R_{ik} . \qquad (52)$$

最优投资比例变为 $\boldsymbol{q}^{**}=\boldsymbol{q}^{**}(P(x|\theta_j),\ (R_{ik}))$。我们把语义 KL 信息 $I(X;\ \theta_j)$提供后得到的增值熵的增量定义为信息价值(即平均信息价值)：

$$V(X;\theta_j) = \sum_i P(x_i\,|\,y_j)\log \frac{R_i(\mathbf{q}^{**})}{R_i(\mathbf{q}^*)} . \qquad (53)$$

可见 $V(X;\theta_j)$ 和 $I(X;\theta_j)$ 有类似结构。对于上式，当 x_i 确定发生时，y_j 的信息价值变为

$$v_{ij} = v(x_i;\theta_j) = \log \frac{R_i(\mathbf{q}^{**})}{R_i(\mathbf{q}^*)} . \qquad (54)$$

它和单个语句提供的信息量公式形式类似。信息价值也需要事实检验，错误预测可能带来负的信息价值。

Arrow和Cover的信息价值公式

Arrow的信息价值公式：

比较我的Ri= $\log \sum_{k=0}^{N} q_k R_{ik}$

Ri=

Arrow 定义的效用函数是【65】：

$$U=\sum_i P_i U(q_i R_i)=\sum_i P_i \log(q_i R_i)=\sum_i P_i \log q_i+\sum_i P_i \log R_i\,, \quad (55)$$

其中 $U(q_iR_i)=\log(q_iR_i)$是第 i 种收益发生时，投资人获得的效用。

在 $\sum_i q_i=1$ 的限制下，当 $q_i=P_i(i=1,2,\ldots)$时，U 达最大，为

$$U^*=\sum P_i \log P_i+\sum_i P_i \log R_i\,. \quad (56)$$

而当投资人有了信息后，准确知道哪一种收益将要发生，从而将资金全部投到它上面时，有

$$U^{**}=\sum_i P_i \log R_i\,. \quad (57)$$

信息价值被定义为有无信息时投资的效用差，它正好是 Shannon 熵，即

$$V=U^{**}-U^*=-\sum_i P_i \log P_i=H(X)\,. \quad (58)$$

结论漂亮， 但是荒唐！

Cover有类似结论：信息等于翻倍率。Cover和Thomas是《信息论基础》作者

这些结论是错的（参看我的《…语义推广…》），太乐观。如果赔率很小， 赚10%都很难，更不要说翻倍了。

最优下注比例和投资组合理论

我的信息价值公式和Kelley的打赌下注公司兼容。

Kelley公式： $q^*=P-(1-P)/r2$

其中P：赢的概率；r2：赢的赔率；假设亏的赔率说1.

我推广的下注公式： $q^*=E/(r1r2)$.

比如：掷硬币打赌，正反面等概率，出正面赚两倍，出反面亏一倍， 最优比例：

$$q^*= 0.5(2-1)/(1X2)=0.25$$

我还得到分散投资极限定理和投资容量公式——和信道容量公式类似。

E说期望收益，Rr是风险

$$H_C{}^* \approx \frac{1}{2}\log(1+\frac{E^2}{R_r{}^2}) \cdot$$

参看短篇系列：《漫谈投资组合的几何增值理论》

http://www.survivor99.com/lcg/mantan.htm

和专著《投资组合的熵理论和信息价值》，《股市幸存者如是说》

http://www.survivor99.com/

G理论可用于机器学习的主要原因

- **1.** 优化准则

最大语义信息准则等价于最大似然准则， 兼容正则化误差准则

- **2.** 语义信道和香农信道相互匹配——即模型和样本分布相互匹配

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

$$=\sum_j\sum_i P(x_i)P(y_j\mid x_i)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}$$

- **3.** 解决 语义约束问题

从失真函数，相似函数，真值函数， 隶属函数…都可以得到似然函数，以前能用似然函数作为约束的地方， 现在也能用其他函数作为约束。

有监督学习：多标签学习和分类

- **根据样本分布P(x,y)得到学习函数再进行分类**。流行的方法是Binary Relevance，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,θ)（学习函数）的归一化要求：P(θ1|x)+P(θ2|x)+…=1， 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，我们直接得到学习函数(随机点落影）：

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

- 如果样本不够大，我们用语义KL公式：

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

- 分类时，我们用最大语义信息分类器

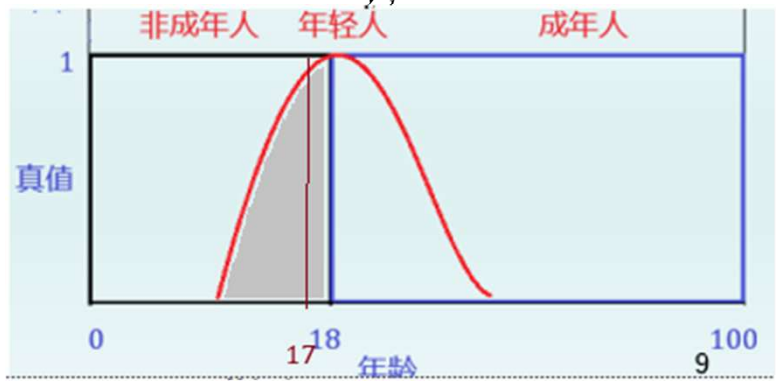
$$y_j^* = h(x) = \arg\max_{y_j} I(x;\theta_j) = \arg\max_{y_j} \log\frac{T(\theta_j \mid x)}{T(\theta_j)}$$

- 或最大真值分类器：

$$y_j^* = h(x) = \arg\max_{y_j} T(\theta_j \mid x) = \arg\min_{y_j} d(x, y_j)$$

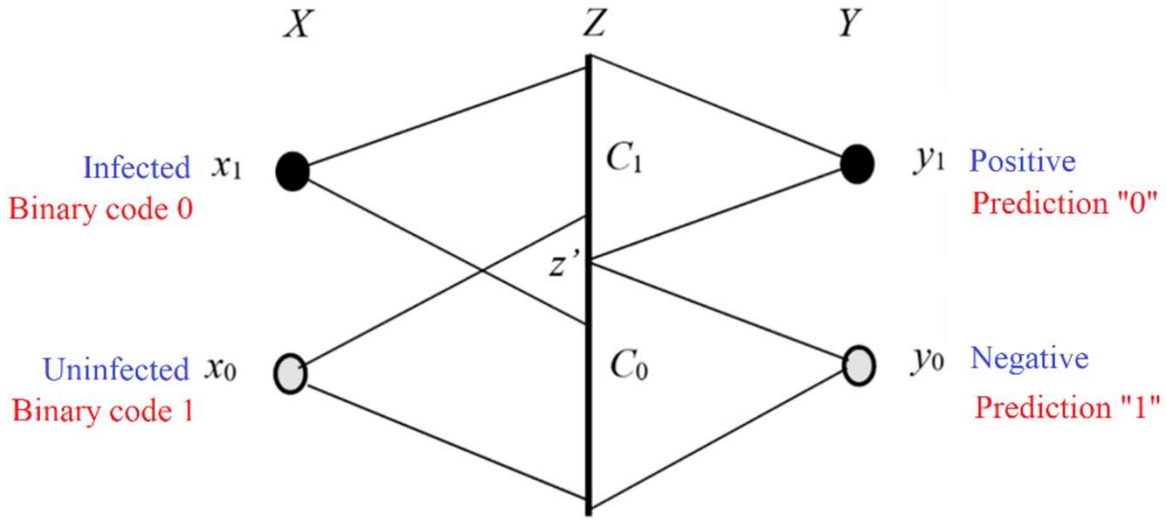
一个例子说明用真值函数和语义信息准则更好：

考虑x=17岁，用标签yj=“x是未成年年轻人”语义信息量最大。但是用P(θj|x)或P(yj|x)分类是不会选yj, 因为P(yj|x)和P(θj|x)很小。



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：

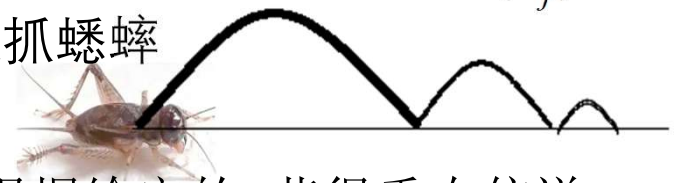


也就是最大似然比检验

- 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。 就像抓蟋蟀



•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：

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

最大互信息分类（续）

- 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：

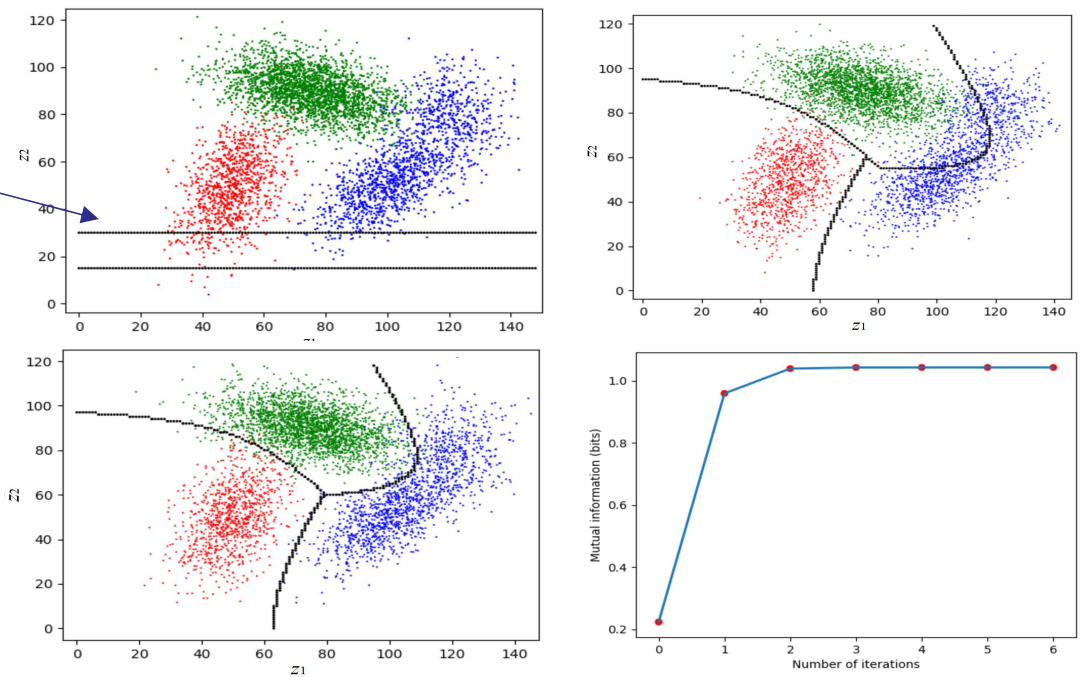
$$I(X;\theta_j|z)=\sum_i P(x_i\,|\,z)I(x_i;\theta_j) \quad , j=0,1,\cdots,n.$$

- **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：

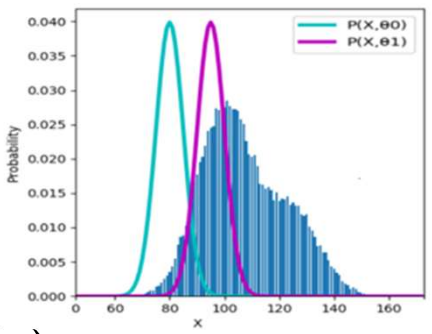
$$y_j{}^*=f(z)=\arg\max_{y_j} I(X;\theta_j|z) \quad , j=0,1,\cdots,n.$$

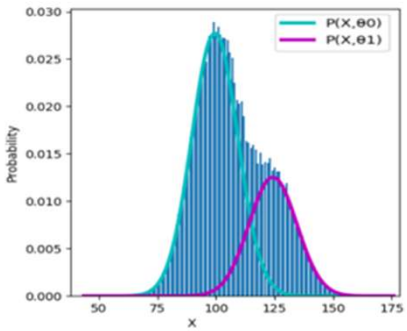
- 重复匹配1和匹配2直至S不再变化。收敛的S就是我们要找的S^*。
- 一个例子：

初始划分很糟糕



无监督学习：证明混合模型收敛





- 样本分布$P(x)=\sum_j P(y_j)P(x|y_j)$，右边混合比例不可见。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。

我把EM算法改进为EnM算法：先假设一个P(y)

匹配1：令香农信道匹配语义信道, 通过最小互信息迭代公式（s=1）：

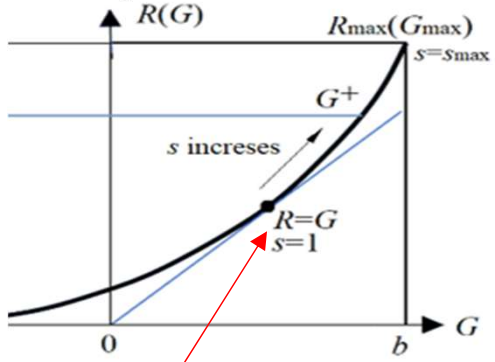
E步 $$P(y_j|x)=P(y_j)P(x|\theta_j)/P_\theta(x),\ P_\theta(x)=\sum_j P(y_j)P(x|\theta_j),$$

$$P^{+1}(y_j)=\sum_x P(x_i)P(y_j|x_i).$$

重复直至P(y)不变，

匹配2：令语义信道匹配香农信道, 通过

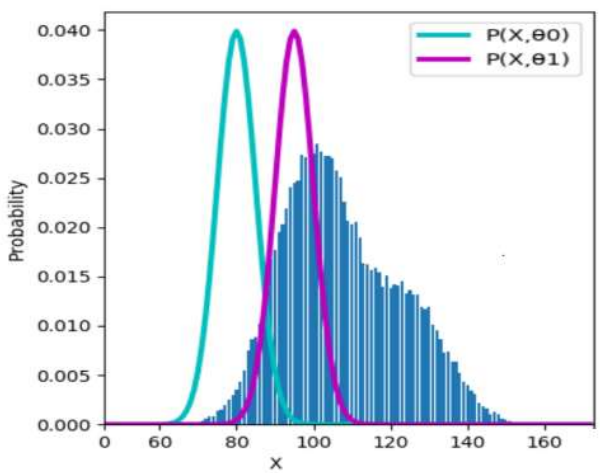
$$P(x|\theta_j^{+1})=P(x)P(x|\theta_j)/P_\theta(x),$$

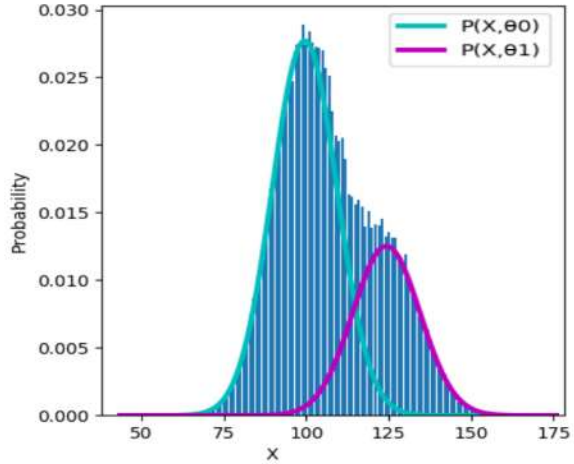


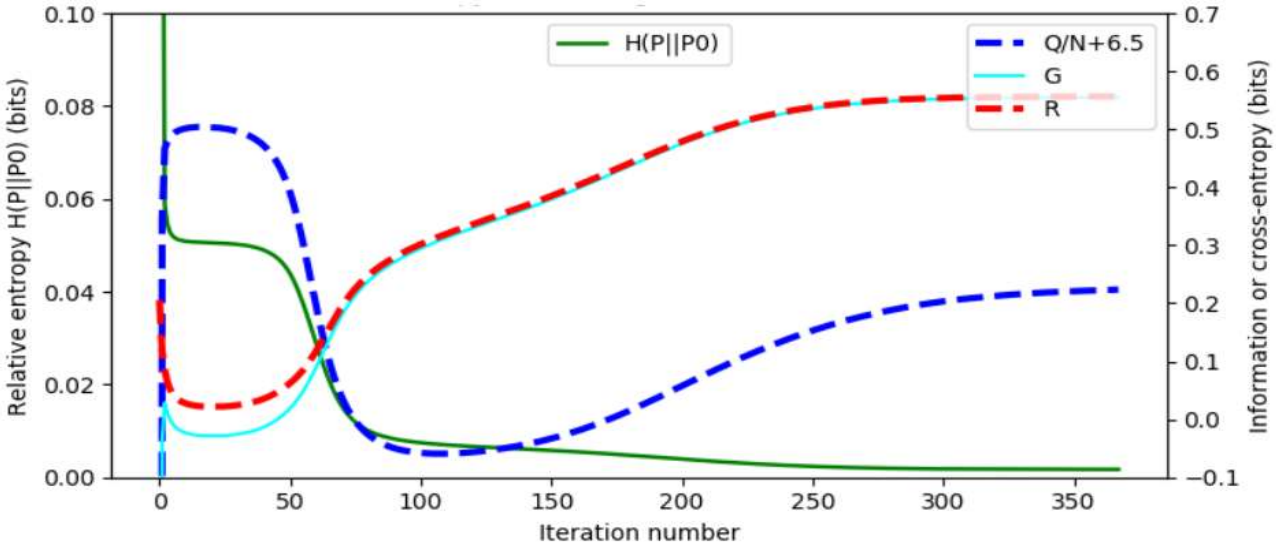
- 直至相对熵或θ不能改进为止。为了证明迭代收敛, 我们能推导出
- $$H(P\|P_\theta)=R-G+H(P^{+1}(y)\|P(y)) \rightarrow 0$$
- 因为匹配2最大化G, 且匹配1最小化R和$H(P^{+1}(y)\|P(y))$, 所以$H(P\|P_\theta)$能接近0。
- 这个公式非常重要，涉及深度学习预训练和最小变分自由能原理。

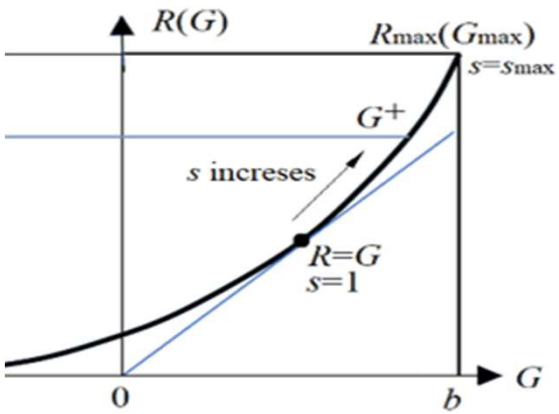
EnM算法和EM算法比较

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。我发现收敛过程中，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。

用于主动推断和强化学习

陈述句和祈使句：两种语义信息的异同

都用语义KL公式：

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

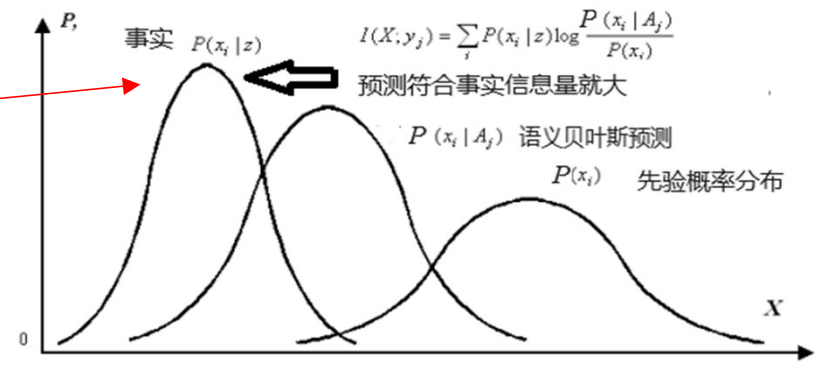
陈述句（天气预报）的语义信息：

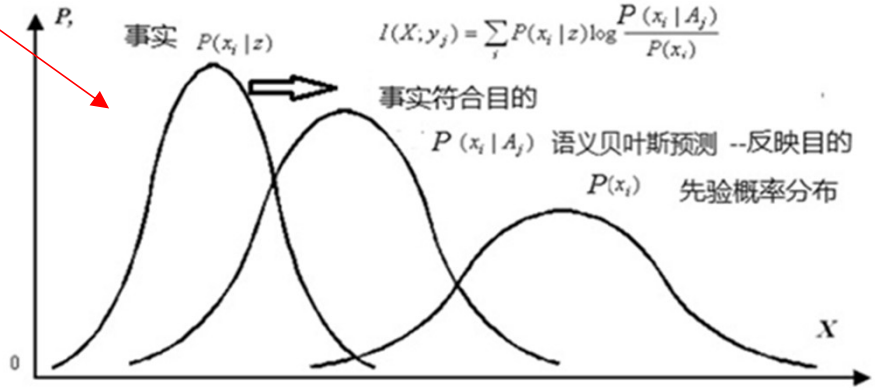
希望预测符合事实，语义信息量大

祈使句（指令）的语义信息

（合目的信息）：

希望事实符合目的，信息效率高



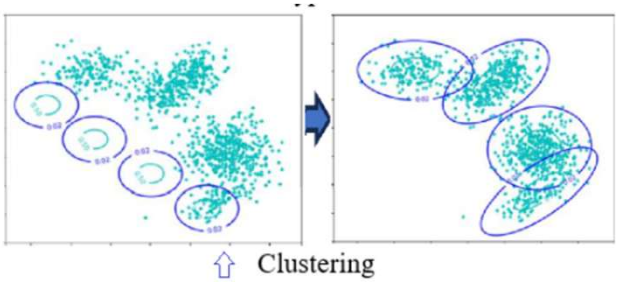


主动推断的两个例子

概率预测和

混合模型：

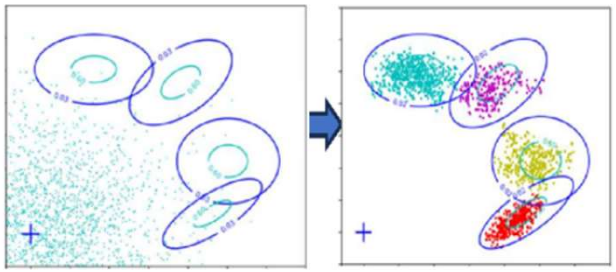
如套圈圈



祈使句和

主动推断：

如赶羊进牧场

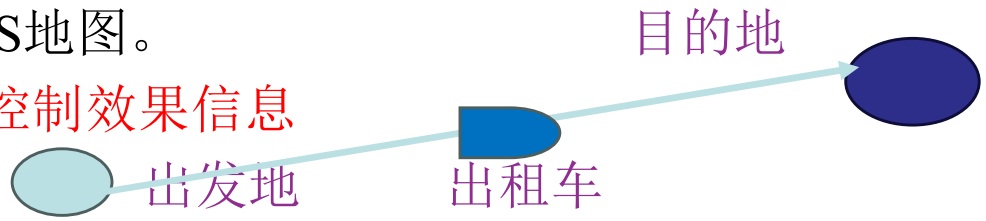


合目的信息作为奖惩函数

考虑乘客和出租车司机观察车上的GPS地图。

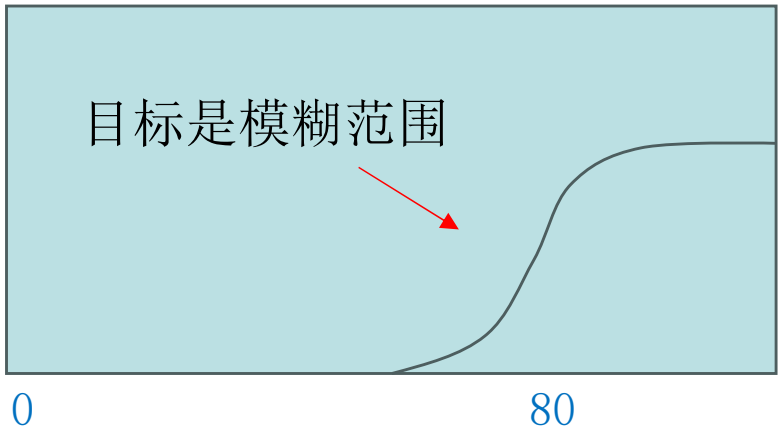
乘客信息：合目的信息；司机信息：控制效果信息

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，如何度量和优化合目的信息？

- 人口死亡年龄最好超过80岁，用T(θj|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(θj|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：

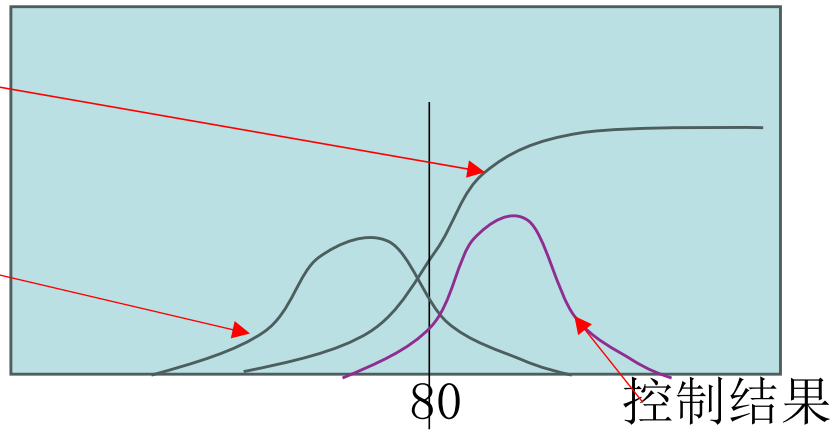
$$I(X;a_j/T)=\sum_i P(x_i\mid a_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)},$$

而香农信息或KL信息反映控制成本：

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

合目的信息优化——最大化信息效率

- 人口死亡年龄最好超过80岁，用T(θj|x)表示；当前死亡年龄分布是P(x)。
- 要在成本和效用之间权衡



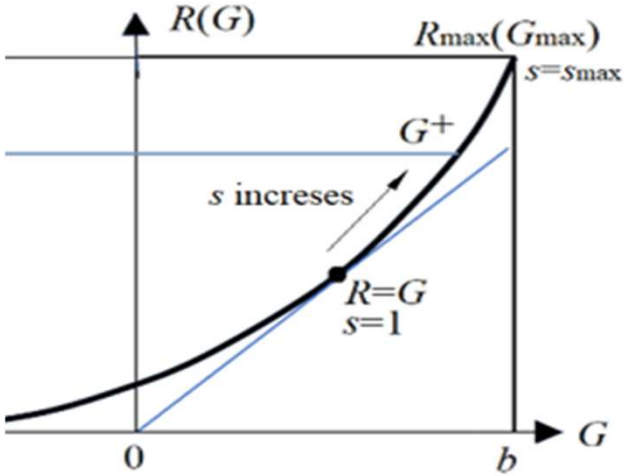
优化是改变$P(x|aj)$,让它向$P(x|\theta j)$靠拢，最小化$R-G=I(X;aj)-I(X;aj/T)$，提高通信效率G/R。有了相互匹配的$P(x|aj)=P(x|\theta j)$，再考虑最大化G和R——通过增大s.

$$P(a_j \mid x) = P(a_j) m_{ij}^s / \lambda_i,\ \lambda_i = \sum_j P(y_j) m_{ij}^s,$$

$$P(a_j) = \sum_i P(x_i) P(a_j \mid x_i).$$

其中：

$$m_{ij}=P(x_i|\theta_j)/P(x_i)=T(\theta_j|x_i)/T(\theta_j)$$



我们可以适当选择s(比如选s=5)，在最大化G和最大化G/R之间权衡。这也是语义变分贝叶斯的基本思想和主动推断要改进的地方。

如果有多任务，I(X; aj/T)就变成语义互信息I(X; A/T). 需要求隐含变量分布P(y)。

改进变分贝叶斯(VB)为语义变分贝叶斯(SVB)

- 给定 P(x)和约束 P(x|θj), j=1,2,… 求解产生 P(x)=∑j P(yj)P(x|θj)的 P(y). P(y)被称 之为隐含变量(实际指隐含变量的概率分布)。流行的是变分贝叶斯方法(简称 VB)。起源于Hinton等人的研究。

$$F=\sum_{y} g(y)\log\frac{g(y)}{P(x,y\mid\theta)}=-\sum_{y} g(y)\log P(x\mid y,\theta)+KL(g(y)\,\|\,P(y))$$

- 最小化F的g(y)就是所求隐含变量。
- 用语义信息论方法，F应该是：

$$F=\sum_{i} P(x_i)\sum_{i} P(y_j\mid x_i)\log\frac{P^{+1}(y_j)}{P(x_i,y_j\mid\theta)}$$

$$=\sum_{j} P^{+1}(y_j)\sum_{i} P(x_i\mid y_j)=\log\frac{P^{+1}(y_j)}{P(x_i\mid\theta_j)P(y_j)}=H(X\mid Y_\theta)+KL(P_Y^{+1}\,\|\,P_Y)$$

- 右边相对熵趋于0，所以可以认为F就是交叉条件熵*H(X|Yθ)*。
- 我提出语义变分贝叶斯SVB，优化似然函数P(x| θj)和VB一样， 优化P(y|x)使用最小互信息迭代。结果和VB类似， 但是算法更简单。
- VB实际求解P(y)的方法是先用平均域近似方法求P(Y|X)，实际上也是用P(y|x)代替P(y)作为变分，结果类似， 但是算法复杂。

比较变分贝叶斯(VB)和语义变分贝叶斯(SVB)

- **比较：**

准则: VB 的定义，它名义上采用最小自由能(即最小语义后验熵)准则； 而在求 P(y)的时候，它使用 P(y|x)作为变分，实际上使用的是使混合模型收敛的最大似然准则；而 SVB 采用最大信息效率准则，它等于最大似 然准则(优化模型参数)加上最大熵准则（即最小互信息准则）。

变分方法：VB 只用 P(y)或 P(y|x)作为变分，而 SVB 轮流使用 P(y|x)和 P(y)作为变分。

计算复杂性：VB 求 P(y|x)用到对数和指数函数，计算比较复杂； SVB 中的 P(y|x)计算比较简单(对于同样任务，即 s=1 时)。

$$P(a_j \mid x) = P(a_j) m_{ij}^s / \lambda_i, \ \lambda_i = \sum_j P(y_j) m_{ij}^s,$$

$$P(a_j) = \sum_i P(x_i) P(a_j \mid x_i).$$

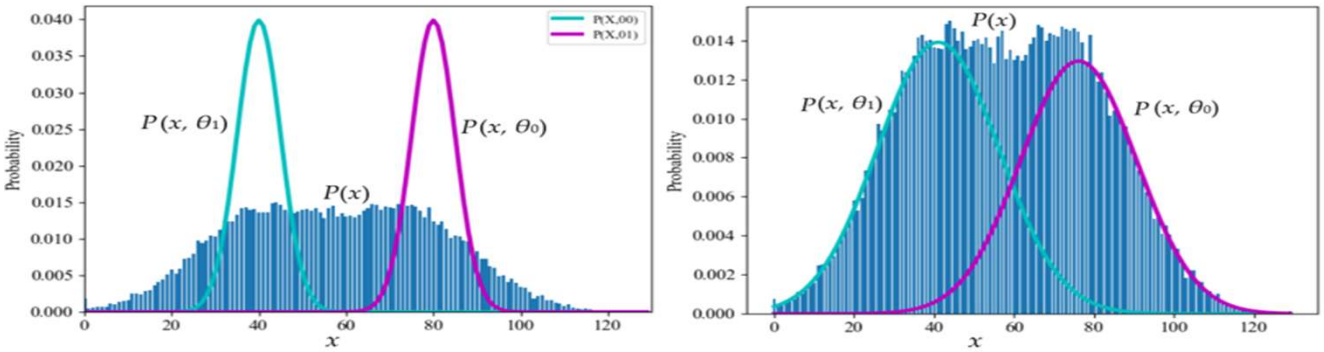
约束：VB 只用似然函数作为约束；SVB 允许使用各种学习函数(包括似 然、真值、隶属、相似、失真函数)作为约束，并且允许用参数 s 增强约 束。

一个典型反例——证伪VB和最小自由能原理（FEP）

Neal和Hinton曾用最小自由能准则优化混合模型，说在E-step和M-step都最大化负的自由能F'可以使混合模型收敛。但是存在反例：

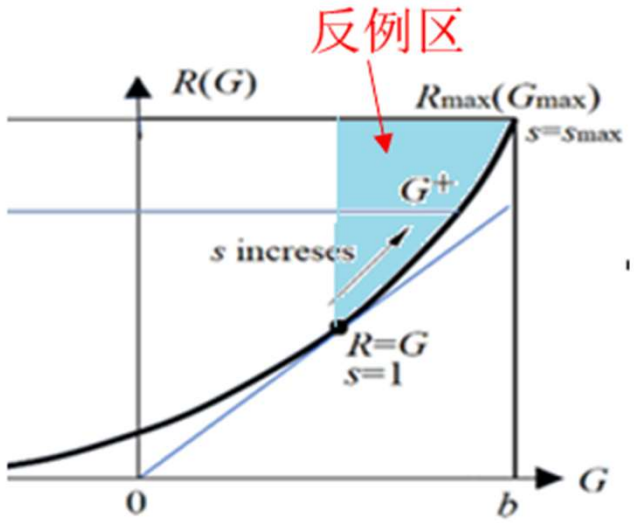
一个混合模型的 F' 和 Q 随迭代过程减小

	Real Parameters			Starting parameters		
	μ^*	σ^*	$P^*(Y)$	μ	σ	$P(Y)$
y_1	40	15	0.5	40	5	0.5
y_2	75	15	0.5	80	5	0.5



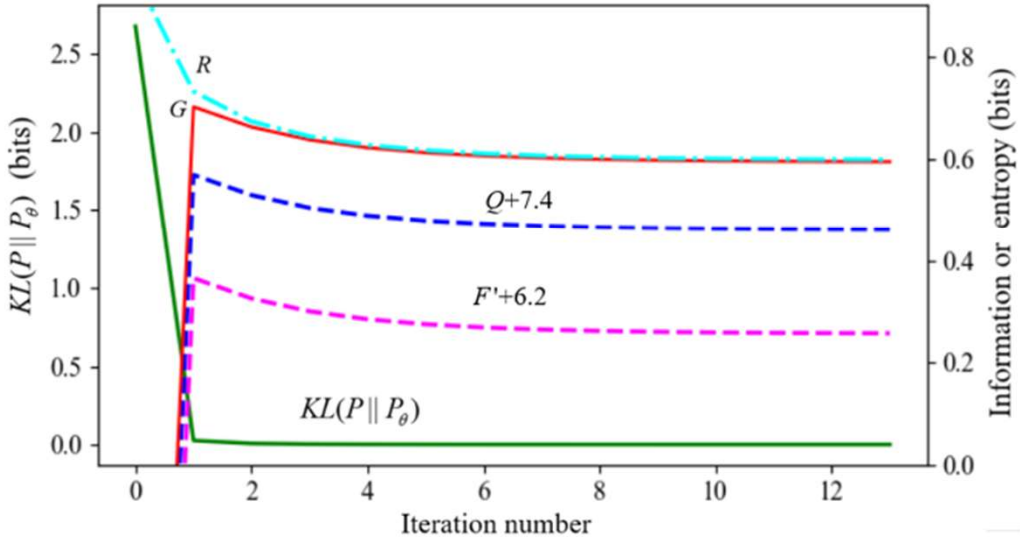
(a)迭代开始

G减小F增大的例子



迭代收敛过程中，负的自由能和完整数据对数似然度Q是减小的，即变分自由能是增大的。

原因是真模型的交叉条件熵 $H(X|Y\theta)$ =F较大。

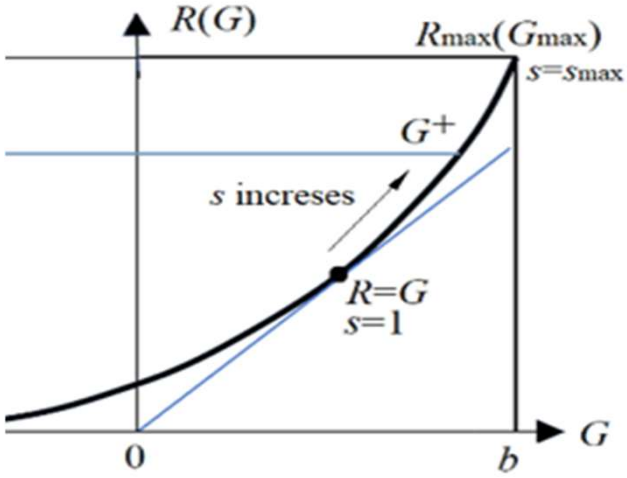


重点：改进最小自由能原理为最大信息效率原理

- Friston把最变分贝叶斯(VB)应用于感觉和脑科学，提出最小自由能原理（FEP）。
- 我发现FEP存在三个问题： http://www.survivor99.com/LCG/information/Improve/

1. VB存在反例：混合模型收敛时，变分自由能可能增大；
2. 变分自由能F和物理学自由能通常反相关，因为F=H(X|Yθ), 语义互信息I(X; Yθ) =H(X)-F; 信息对应自由能，而F对应某种热力学熵；
1. VB和FEP中约束函数只能用似然函数， 而使用SVB， 约束函数可以是多种函数（包括真值函数，隶属函数，相似函数，失真函数）。

- 为此，我把FEP改进为最大信息效率原理:



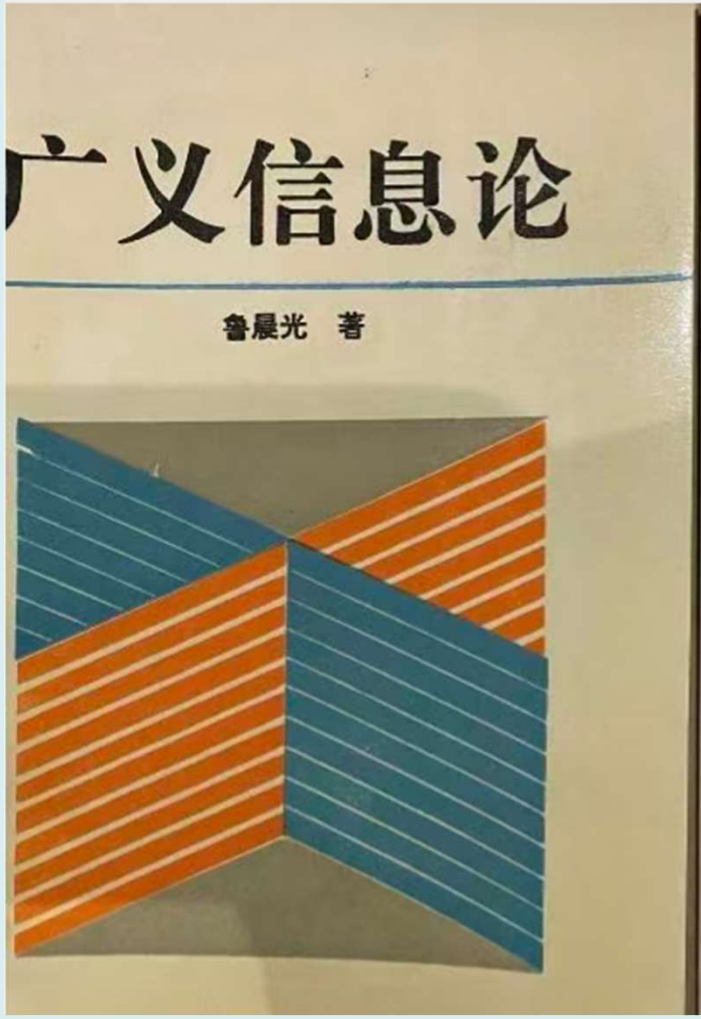
在机器学习和约束控制（或主动推断）中，我们要么追求最大信息效率G/R=1（比如在概率预测和混合模型任务中）；要么给定G最大化G/R(即最小化R）（最大熵方法）或在最大化G和最大化G/R之间权衡（比如在约束控制和主动推断任务中）。

我们可用这一原理解释生物自组织行为——生物使用它预测和适应（包括改造）环境。它兼容适者生存原理。

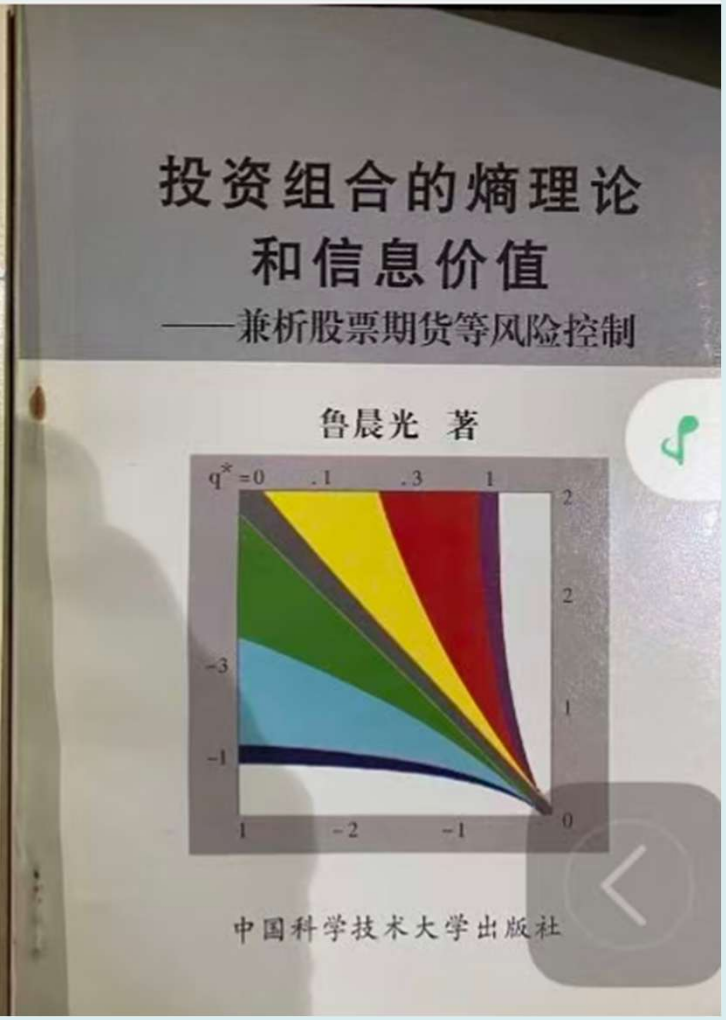
我已经告诉Friston我的研究结果， 他欢迎我参加他们的研讨会。

发表的信息论专著和文章

1993



1997



发表的信息论文章（部分）

- 1989: 色觉的译码模型及其验证, 光学学报,9,2(1989)，158—163
- 1990: Shannon equations reform and applications,BUSEFAL 44， 4 （1990）, 45—52
- 1994:《广义熵和广义互信息的编码意义》，《通信学报》， 5卷6期, 37-44.
- 1999: A generalization of Shannon‘s information theory“ , Int. J. of General Systems, 28: (6) 453-490 1999 这是《广义信息论》的缩写改进版。
- 2018： 语义信息G理论和逻辑贝叶斯推断用处机器学习. 英文版发表于Information 2019, 10, 261. https://www.mdpi.com/2078-2489/10/8/261

  中文版：http://www.survivor99.com/lcg/books/GIT/
- 2019： 信道确证和预测确证——从医学检验到乌鸦悖论，英文版： Entropy 2020, 22, 384. https://www.mdpi.com/1099-4300/22/4/384
- 中文版 ：，**http://www.survivor99.com/lcg/CM/Raven/index.html**
- 2023: 英文评论：回顾学习函数和语义信息测度的进化——进而理解深度学习，英文版： *Entropy* **2023**, *25*, 802. https://doi.org/10.3390/e25050802 中文版：（ ， http://survivor99.com/LCG/information/Review-Evolution/index.html
- 2025：英文综述文章:**香农信息论的语义推广(G理论)和应用，Entropy，https://www.mdpi.com/1099-4300/27/5/461 中文：http://www.survivor99.com/LCG/information/G-theory/index.html**
- **2025：改进最小自由能原理为最大信息效率原理，英文：Entropy，https://www.mdpi.com/1099-4300/27/7/684， 中文：http://www.survivor99.com/LCG/information/Improve/** 今后重点研究方向

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。

- **今后重点研究方向：**

1）用最大信息效率原理解释生物自组织和进化，改进主动推断；

2）用真值函数解释Word2Vec，用语义信息论解释和改进大模型——结合深度学习方法。个人能力有限，心有余而力不足；期待后人继续努力。

谢谢各位听讲！欢迎批评和交流！

鲁晨光主页：http://survivor99.com/

语义信息论和机器学习网页： **http://www.survivor99.com/lcg/books/GIT/**