贝叶斯概率模型

> 机器学习狭义上是指代统计机器学习，如图 1 所示，统计学习根据任务类型可以分为监督学习、半监督学习、无监督学习、增强学习等。

雷锋网按：本文出自美图数据研究院

什么是贝叶斯概率模型?

机器学习狭义上是指代统计机器学习，如图 1 所示，统计学习根据任务类型可以分为监督学习、半监督学习、无监督学习、增强学习等。

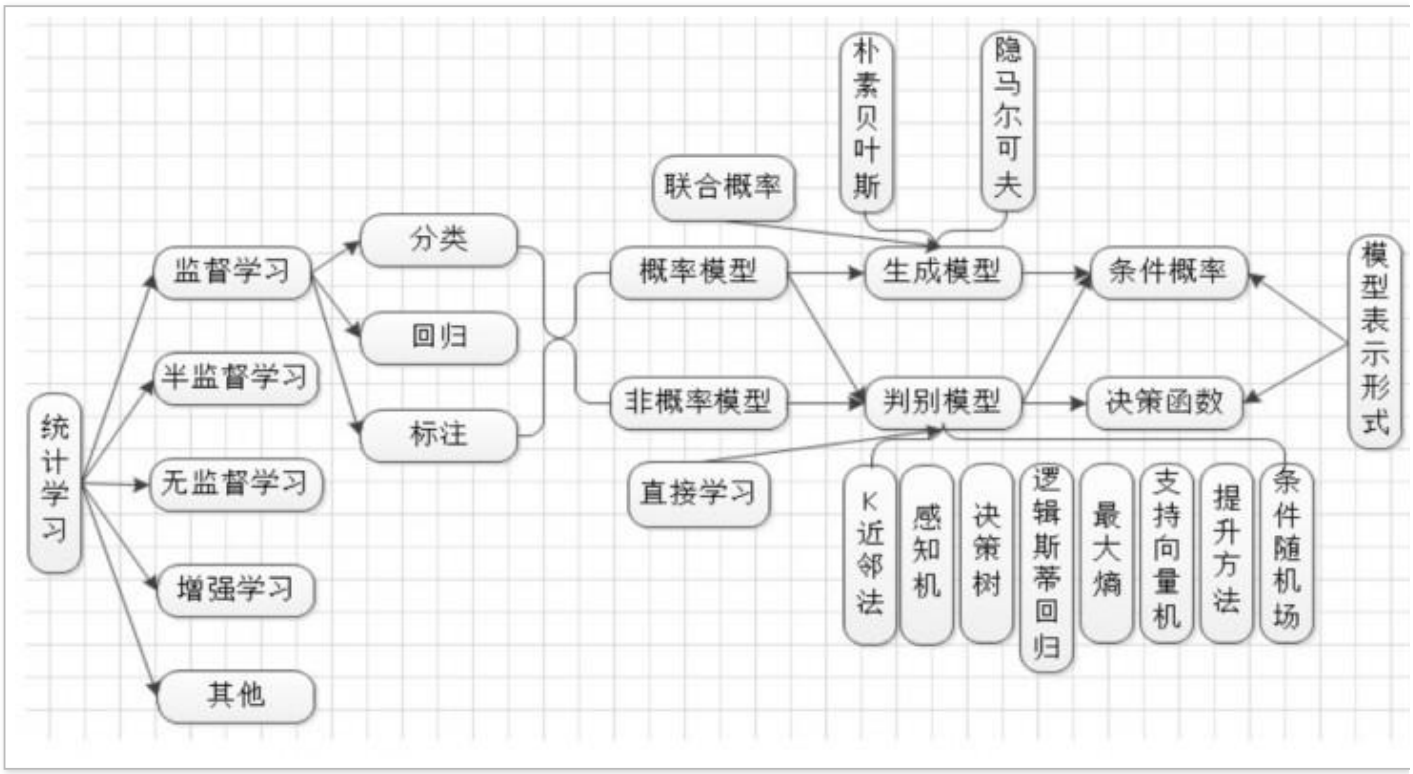


图 1

在每类任务中，又可以将各类模型归结为概率模型和非概率模型，以下以监督学习为例说明。

概率模型（生成模型）通过函数 F 来描述 X 和 Y 的联合概率或者条件概率分布，如 P(X|Y)；非概率模型（判别模型）通过函数 F 来直接描述 X 到 Y 的映射，如 Y=f（X)。判别模型的优化目标五花八门，但都符合人类认知；而在概率模型中，所有模型的优化目标是统一的，即最大化观测数据在概率模型中出现的概率。这两者在部分模型表现形式上又可以互相解释，如神经网络等。

贝叶斯概率模型的诞生

所有概率模型描述的都是在系统参数 w 下观测变量对 X,Y 的联合概率分布或条件概率分布，即 P(Y,X|w) 。设计好概率模型后，剩下的问题就是如何通过大量的观测数据来决定参数 w, 这时出现了贝叶斯理论。

频率学派主张大数定律，对参数的最佳选择是使观测变量概率最大的值；而贝叶斯学派提出了贝叶斯公式和主观概率，他们认为参数可以是一个分布，并且最初可以通过主观经验设置。频率学派的人对此是无法接受的，他们认为参数应该是一个确定的值不应该有随机性。

举个例子，有个检测太阳是否爆炸的探测器，它有 0.3 左右的概率撒谎。当探测器说出太阳爆炸时，两个学派的人答案是不一样的。

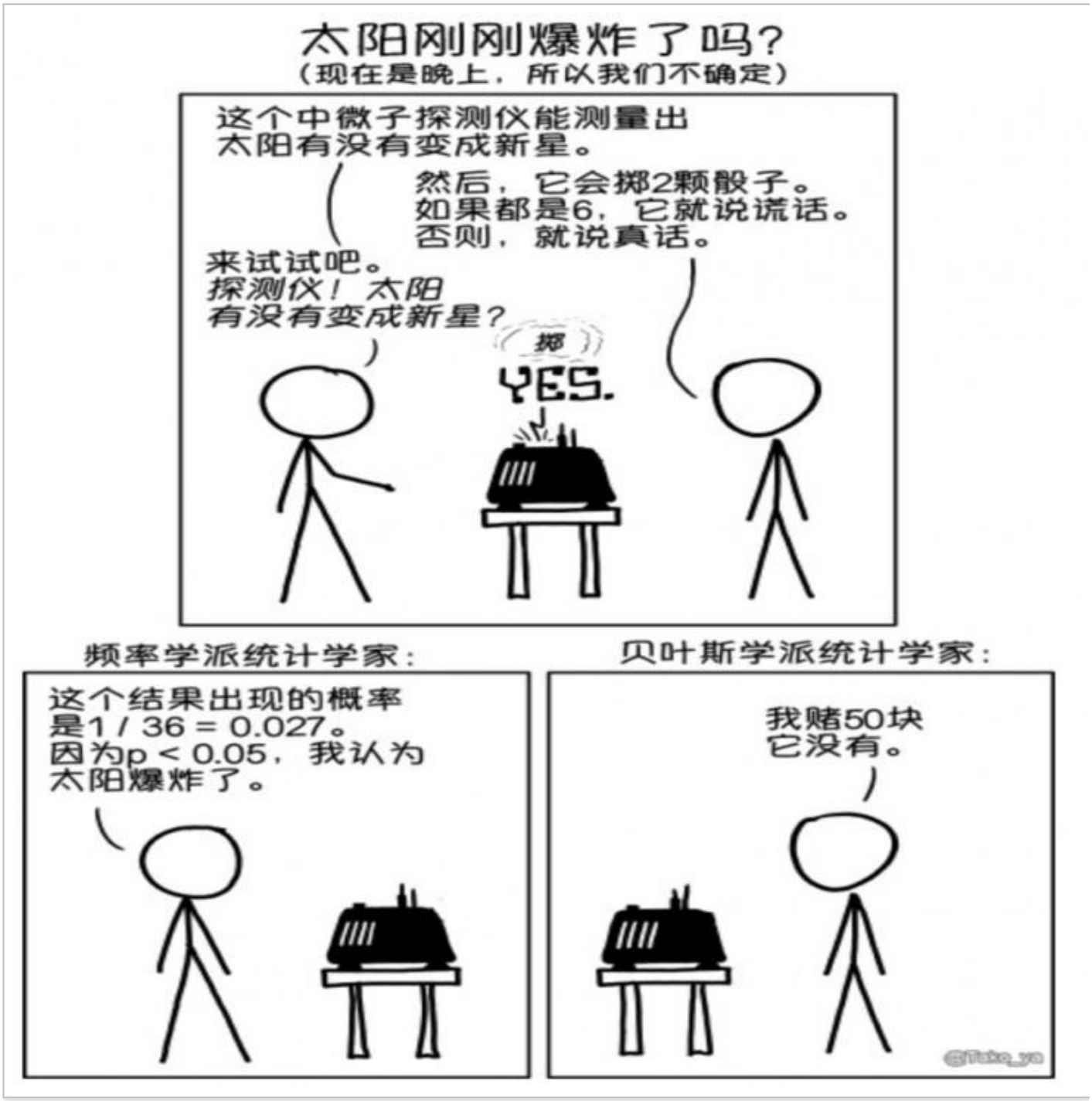


图 2

如图 3 所示这里太阳的状态是系统参数 w，探测器回答是观测变量 data。以频率学派的理论来讨论，如果参数只能是一个确定的值，那么应该选取出错概率最小的那个参数，那太阳应该是爆炸了；如果以贝叶斯学派来讨论，将参数视为分布，并根据我们的经验赋予先验，得到的后验认为太阳应该是没有爆炸的，只有当探测器多次回答「yes」，后验分布才会相信太阳爆炸了。

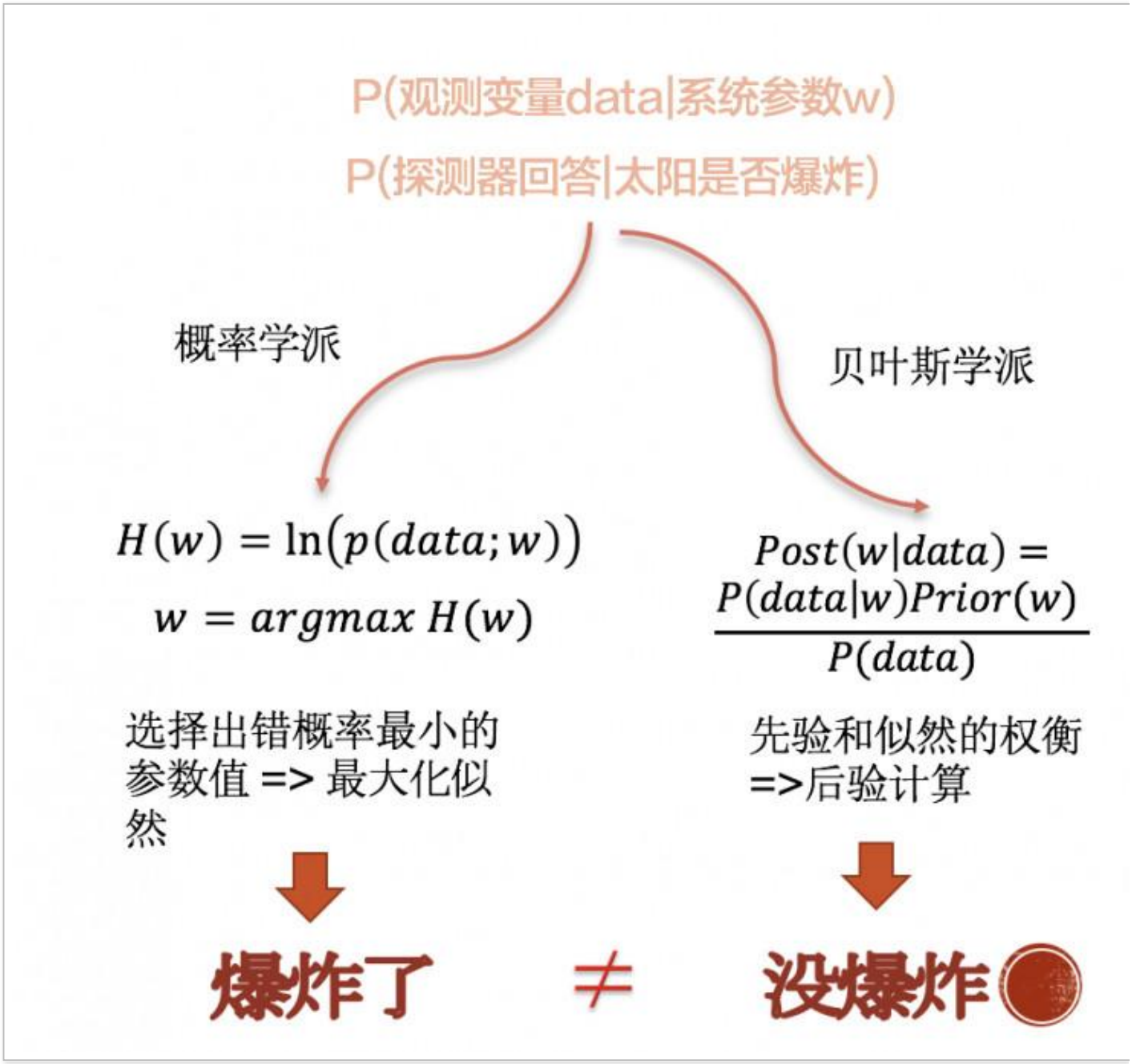


图 3

贝叶斯学派和概率学派在小数据量的场景下的推论结果常常是有一定区别的，因此它有存在的必要。

构建贝叶斯概率模型

接下里通过构建贝叶斯概率模型案例直观地感受贝叶斯概率模型的核心概念、构建思想和优势。

CKF （Collaborative Kalman filter）

	movie1	movie2	movie3	moive4
user1	1			
user2	2			3
user3		5	4	
user4	2			4

$$\textbf{MF:} \quad y_{ij} \sim N(< u_i, w_j >, \sigma^2)$$

图 4

如图 4 所示，这种基于频率学派模型存在两个比较大的缺陷：

- 无法增量训练。理论上每新增一条用户行为，模型就要重新估计一遍参数；
- 无法处理用户兴趣漂移。粗暴的做法是设置时间衰减，但是衰减的函数和力度都需要人工把握，模型对超参数很敏感，而且每个用户的兴趣漂移能力应该是不同的，这点无法建模。

根据以上提到的两大缺陷，通过贝叶斯将该模型进行改造。首先将参数都变成分布的，把用户向量 u 和物品向量 w 都赋予维纳过程：

$$w_{n+1} \mid w_n \sim N(w_n, \alpha I)$$

$$u_{n+1} \mid u_n \sim N(u_n, \alpha I)$$

给 u 和 w 赋予一个方差很大的先验分布。输入数据时计算后验。将后验通过维纳过程得到下一刻的先验：

每个用户的兴趣漂移能力不同：

* 这里通过维纳过程算下一刻先验，实质上在上一刻后验的基础上加一个方差 α。从而保证状态始终有一个漂移能力，如果这个 α 等于 0，就会出现随着推理的进行 u 的分布只会越来越集中，这样即便后面用户兴趣漂移了，由于先验分布集中，似然函数无法调整。所以这里的 α 控制的是兴趣漂移的能力。

再设一层概率分布，令 α 也为一个维纳过程，让每个用户的兴趣漂移能力可以自适应去调整和变化：

经过贝叶斯改造之后，CFK 模型有以下优势：

- 训练过程中是增量进行的；
- 无参化，数据越来越多时，后验方差会越来越小，分布越来越集中，实现先验与数据的自动权衡；
- 漂移参数自适应，当用户兴趣发生漂移时，状态会跟随着漂移。

Bayesian Neural Networks

Bayesian Neural Networks 是指通过后验推理扩展标准网络。通过优化标准神经网络训练（从概率学派的角度来看）等同于权重的最大似然估计。

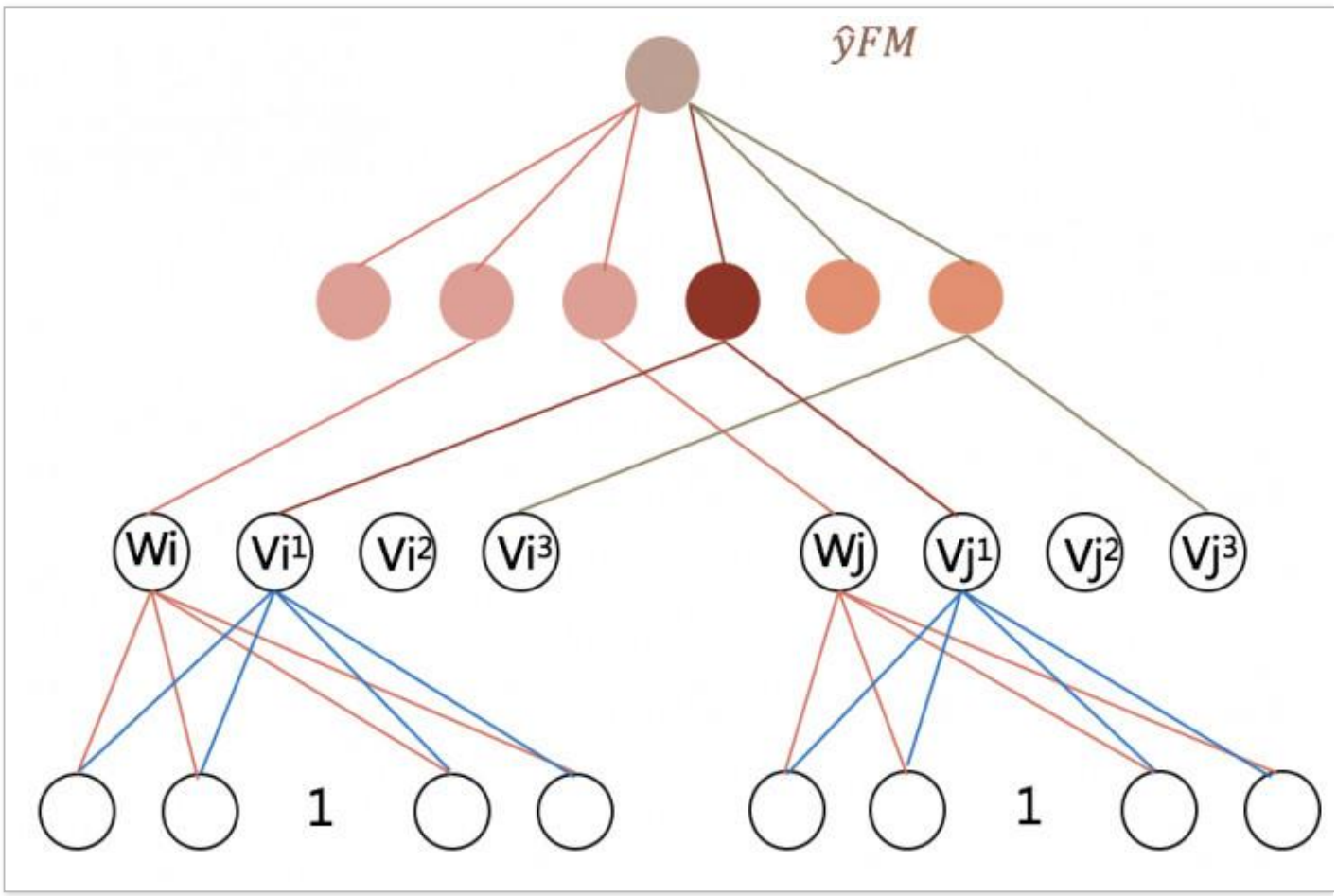


图 5

它存在以下三大缺陷：

- 无法增量训练；
- 网络结构等需要超参数设置；
- 无法衡量预测不确定性。

$$p(\mathbf{y} \mid \mathcal{W}, \mathbf{X}, \gamma) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid f(\mathbf{x}_n; \mathcal{W}), \gamma^{-1}).$$

$$p(\mathcal{W} \mid \lambda) = \prod_{l=1}^{L} \prod_{i=1}^{V_l} \prod_{j=1}^{V_{l-1}+1} \mathcal{N}(w_{ij,l} \mid 0, \lambda^{-1})$$

$$p(\lambda) = \mathrm{Gam}(\lambda \mid \alpha_0^{\lambda}, \beta_0^{\lambda})$$

$$p(\gamma) = \mathrm{Gam}(\gamma \mid \alpha_0^{\gamma}, \beta_0^{\gamma})$$

针对以上问题的解决方案是引入正则化，从贝叶斯学派的角度来看，这相当于在权重上引入先验。从概率学派的角度来看这不是正确的做法，尽管它在实践中确实很有效。改造后它有以下优势：

- 可以进行增量训练；
- 非参数模型，无参并非没有超参数，而是把超参数隐藏到更深层，以达到更弱的参数敏感性；
- 可以刻画预测的不确定性；
- 先验与数据自动权衡。

如何更新模型?

变分推理 Variational inference

问题描述：观测变量 X={x1,x2,……,xn}, 隐变量 Z={z1,z2,……zm}，已知 P（X，Z） 或 P（X | Z），求后验分布 P(Z | X) 。由于后验分布有时很难获得解析解，在受限制函数空间中搜索与后验分布函数近似的函数，这里需要一个函数相似性的度量（泛函）：

那么如何获得近似解 q(Z) 呢?

Step 1: 通过调整 q(Z), 最小化 q(Z) 与 后验 p(Z|X) 的 KL 散度 KL（q||p）

Step 2: 将最小优化散度 KL（q||p）的问题转化为最大化 L(q)

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q \,\|\, p)$$

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} \mathrm{d}\boldsymbol{Z}$$

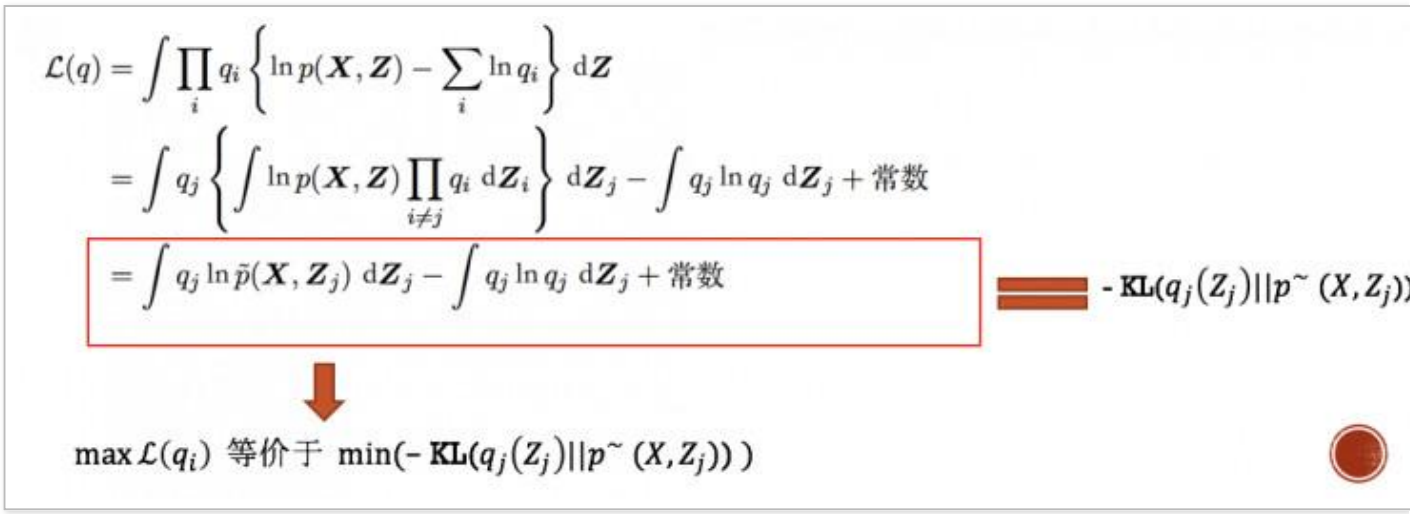
$$\mathrm{KL}(q \,\|\, p) = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z} \mid \boldsymbol{X})}{q(\boldsymbol{Z})} \right\} \mathrm{d}\boldsymbol{Z}$$

*MinKL(q||p) 等价于 MaxL(q)

考虑概率分布 q(Z) 是受限制的类别，我们的目标是充分限制 q(Z) 可以取得的概率分布的类别范围，使得这个范围中的所有概率分布都是可以处理的概率分布。同时还要使得这个范围充分大、充分灵活，从而它能够提供对真实后验概率分布的一个足够好的近似。

Step 3: 利用平均场理论限制函数空间，将 q(Z) 简化为互不相关的几个组：

Step 4: 将分组简化后的 q(Z) 代入以上公式，将其它组视为常来，轮流优化



所有模型的变分推理，都是在交替计算该公式。该公式与模型无关，当对 P(X,Z) 赋予具体形式，便可算出 q 的更新公式：

令： $\ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + 常数$

当： $\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + 常数$

$- \mathrm{KL}(q_j(Z_j) || \tilde{p}(X, Z_j))$ 取得最小值

在对 q(Z) 分组的原则及 q(Z) 函数族的选取原则有两个小建议：

1. 在概率模型中同一层次的隐变量分在一组，在算积分的时候可以使其它层次的对应的条件概率因为不含有改组内的变量而被当做常量，不需计算。

2.q(Z) 的函数族选取条件分布的共轭分布族，在计算期望的积分时需要建条件分布与 q(z) 相乘，选取条件分布的共轭分布族保证相乘完的形式还是原来的简单形式。

概率反向传播 Probabilistic Backpropagation

概率反向传播是贝叶斯神经网络的更新方式，已知：

$$p(\mathbf{y} \mid \mathcal{W}, \mathbf{X}, \gamma) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid f(\mathbf{x}_n; \mathcal{W}), \gamma^{-1})$$

$$p(\mathcal{W} \mid \lambda) = \prod_{l=1}^{L} \prod_{i=1}^{V_l} \prod_{j=1}^{V_{l-1}+1} \mathcal{N}(w_{ij,l} \mid 0, \lambda^{-1})$$

$$p(\lambda) = \mathrm{Gam}(\lambda \mid \alpha_0^{\lambda}, \beta_0^{\lambda})$$

$$p(\gamma) = \mathrm{Gam}(\gamma \mid \alpha_0^{\gamma}, \beta_0^{\gamma})$$

求后验分布 q（ω，ϒ，λ）。

Step 1：利用 KL 逼近 w 的后验

w 的后验分布可以写成

，其中 f(w) 是与 w 相关的似然，设待求后验为高斯分布

。

在算关于 w 的后验过程中，不含有 w 的函数部分都可以看成常数忽略掉。因为 w 与另外两个方差 γ 和 λ 在不同层次，所以 f(w) 中不含有这两个参数。

* 这里虽然是搜索最优函数，但因为限制了函数空间的形式，所以其实是在搜索最优参数 m 和 v

通过最小化 KL 散度 KL(q^new||s)，可以得到直接得到如下的最优值：

$$m^{\text{new}} = m + v \frac{\partial \log Z}{\partial m},$$

$$v^{\text{new}} = v - v^2 \left[\left(\frac{\partial \log Z}{\partial m} \right)^2 - 2 \frac{\partial \log Z}{\partial v} \right]$$

一般来讲这里需使用变分推理，是因为这里的 Z 比较难求（需要求这个积分）：

但是这里通过 Z 的近似形式来替代，它约等于最后一层神经元响应的分布：

Step 2：前向传播，得到 Z^L 的均值和方差

前向过程本质上是一个概率分布的传播过程，但是由于都是高斯分布，所以可以简化成分布参数的传播过程。

$$
\begin{aligned}
\boldsymbol{m}^{\boldsymbol{a}_l} &= \boldsymbol{m}^{\boldsymbol{\omega}_l}\boldsymbol{m}^{\boldsymbol{z}_{l-1}}/\sqrt{V_l+1} \\
\boldsymbol{v}^{\boldsymbol{a}_l} &= [\boldsymbol{v}^{\boldsymbol{\omega}_l}\boldsymbol{v}^{\boldsymbol{z}_{l-1}} + (\boldsymbol{m}^{\boldsymbol{\omega}_l}\circ\boldsymbol{m}^{\boldsymbol{\omega}_l})\boldsymbol{v}^{\boldsymbol{z}_{l-1}} \\
&\quad + \boldsymbol{v}^{\boldsymbol{\omega}_l}(\boldsymbol{m}^{\boldsymbol{z}_{l-1}}\circ\boldsymbol{m}^{\boldsymbol{z}_{l-1}})]/(V_l+1)
\end{aligned} \tag{13}
$$

Let $\boldsymbol{z}_l = max(0, \boldsymbol{a}_l)$, the mean and variance of the i-th element of $\boldsymbol{z}_l$ can be approximated as :

$$
\begin{aligned}
\boldsymbol{m}_i^{\boldsymbol{z}_l} &= \Phi(\alpha_i)v_i' \\
\boldsymbol{v}_i^{\boldsymbol{z}_l} &= \Phi(\alpha_i)\boldsymbol{v}_i^{\boldsymbol{z}_l}(1-\gamma_i(\gamma_i+\alpha_i)) \\
&\quad + \boldsymbol{m}_i^{\boldsymbol{z}_l}\Phi(-\alpha_i)v_i'
\end{aligned} \tag{14}
$$

where

$$
v_i' = \boldsymbol{m}_i^{\boldsymbol{a}_l} + \sqrt{\boldsymbol{v}_i^{\boldsymbol{a}_l}}\gamma_i, \quad \alpha_i = \frac{\boldsymbol{m}_i^{\boldsymbol{a}_l}}{\sqrt{\boldsymbol{v}_i^{\boldsymbol{a}_l}}}, \quad \gamma_i = \frac{\phi(-\alpha_i)}{\Phi(\alpha_i)} \tag{15}
$$

Finally, we can get the mean and variance of $\boldsymbol{z}_L$.

Step 3：利用该公式反向传播，更新参数

$$
m^{\text{new}} = m + v\frac{\partial \log Z}{\partial m},
$$

$$
v^{\text{new}} = v - v^2\left[\left(\frac{\partial \log Z}{\partial m}\right)^2 - 2\frac{\partial \log Z}{\partial v}\right]
$$

最后 Z 将变成含有各层参数的高斯函数，从而可以进行反向梯度计算并更新各层分布的参数，这样就解决了贝叶斯神经网络的模型更新问题。

本文主要介绍了机器学习中的概率模型及贝叶斯理论在概率模型中的应用，这也是人工智能目前比较活跃的方向，相信会有越来越多的工作在这方面进行探索，期待新的发展。我们也会把贝叶斯神经网络应用于实际的业务中，后续的文章中将会与各位交流一些实践经验。

全文完

本文由 简悦 SimpRead 转码，用以提升阅读体验，原文地址