自步学习的理论内涵

毕业设计 答辩

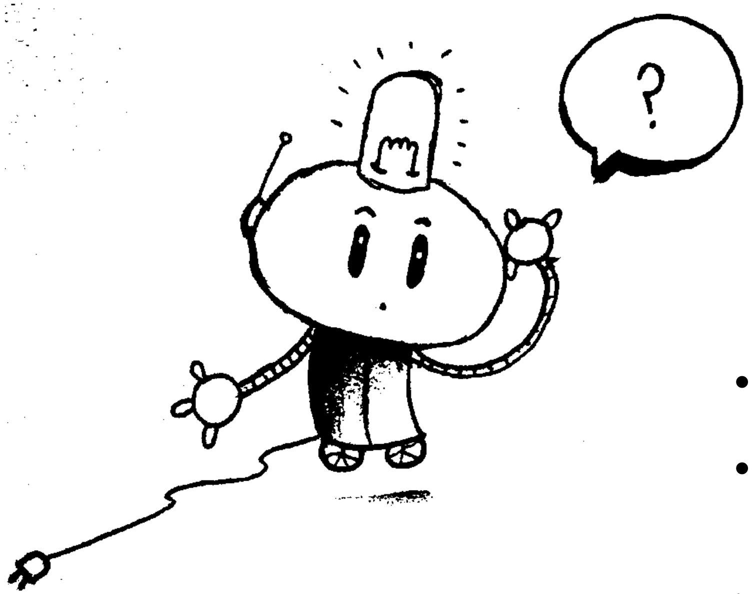
导师：孟德宇

报告人：刘仕琪

专业：数学与应用数学试验班

课题背景

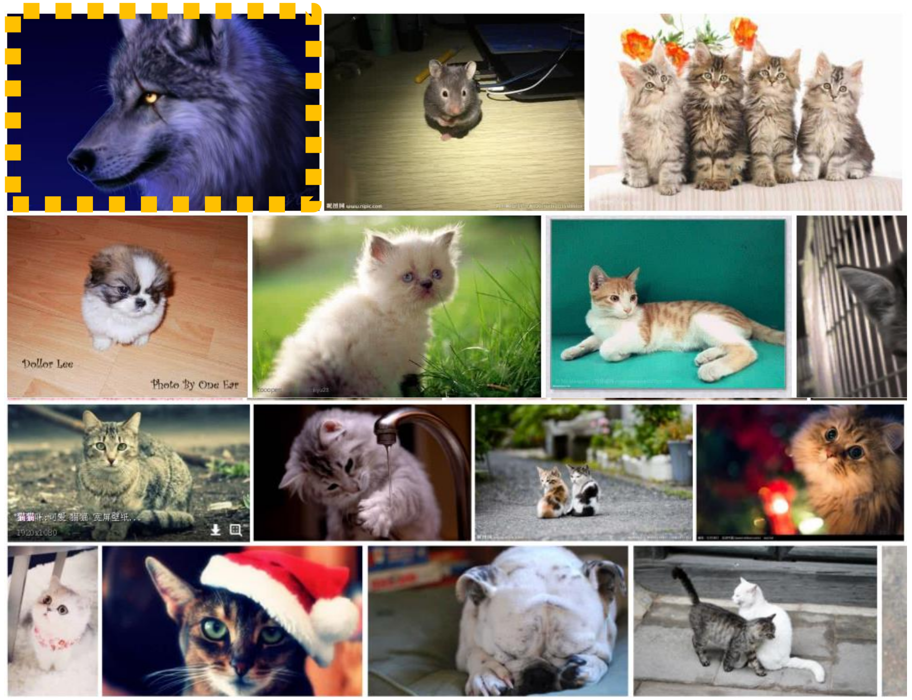
大数据时代下的人工智能



To learn or not to learn

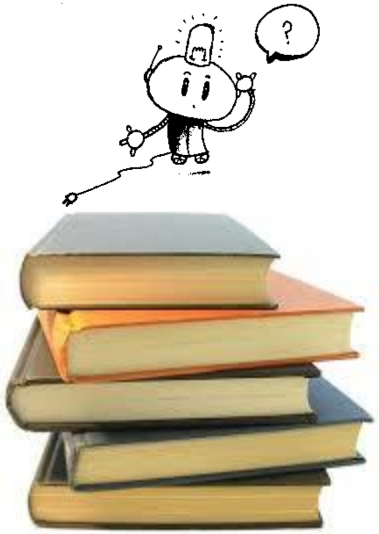
- 内容复杂
- 噪音大

百度六月十一日图片搜索**猫**得到的部分结果



课程学习（CURRICULUM LEARNING）

Bengio



结构性课程

人工成本高

课程设定难度大

Bengio Y, Louradour J, Collobert R, et al. Curriculum learning[C] //International Conference on Machine Learning. 2009: 41–48.

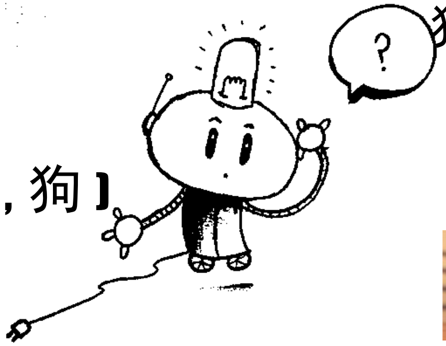
自步学习（SELF-PACED LEARNING）

设立难易程度的标准

简单定义为：预测损失小的样本

【  ，猫 】

【  ，狗 】



狗 简单

猫 ？？

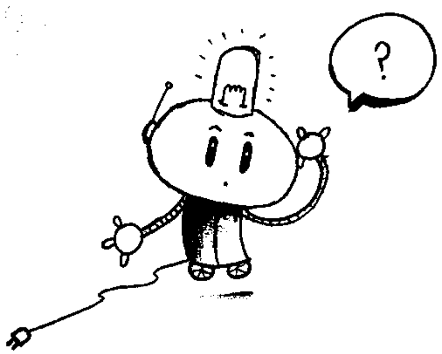
自适应的从易到难的课程学习模式

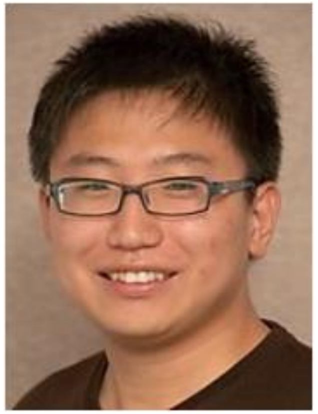
Kumar M P, Packer B, Koller D. Self-paced learning for latent variable models[C] //Advances in Neural Information Processing Systems. 2010: 1189–1197.

自步课程学习（SELF-PACED CURRICULUM LEARNING）

老师驱动和学生驱动结合

自适应从易到难的学习





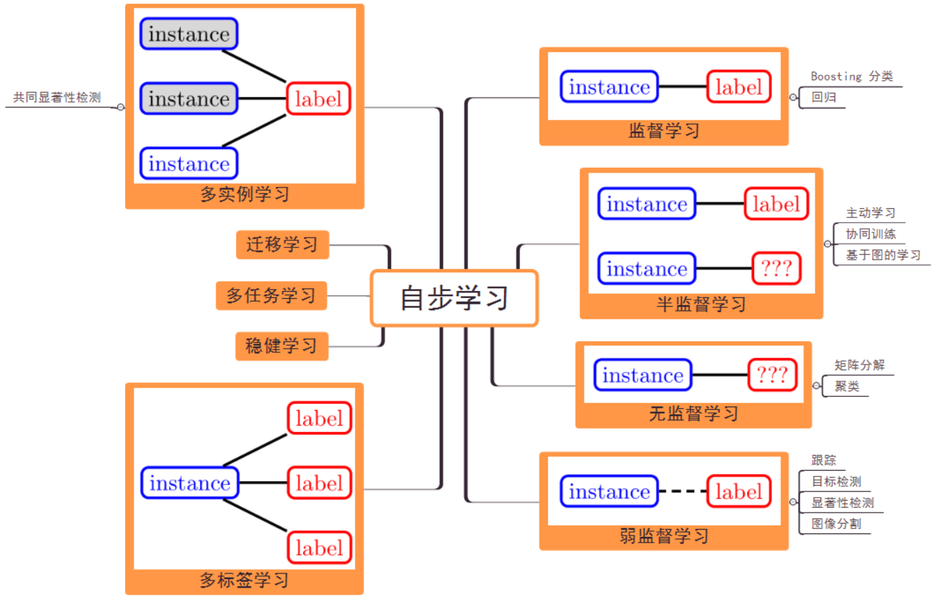
Jiang

外部先验和主观能动性结合

Jiang L, Meng D, Zhao Q, et al. Self-Paced Curriculum Learning.[C] //AAAI: Vol 2. 2015: 6.

自步学习现状

自步 稳健 重要性



多媒体领域

计算机视觉领域

最新水平

多图显著性检测、视频目标重标注、动作识别、行为识别

Lu Jiang, Deyu Meng et al. ACM MM, 2014

Dingwen Zhang, Deyu Meng et al. IJCAI, 2016

动机

自步学习的理论内涵是什么？

- 自步学习在优化什么？
- 难易程度的物理意义是什么？
- 自步学习为什么有效果？

主要理解

第一步　优化的角度：自步学习凹共轭性　+路径算法

第二步　回归的角度：自步学习的贝叶斯网

第三步　概率的角度：自上而下的概率分布学习

优化理解

自步学习模型

数据集 $D=\{\boldsymbol{x}^i, y^i\}_{i=1}^n$　　决策函数 $f(\boldsymbol{x},\boldsymbol{w})$　　模型参数 $\boldsymbol{w}$

各样本上误差 $L^i(y^i, f(\boldsymbol{x},\boldsymbol{w}))$　　样本误差向量 $\boldsymbol{l}=(L^1,\cdots,L^n)^T$　　数据样本权重向量 $\boldsymbol{v}$

$$\inf_{\boldsymbol{w},\boldsymbol{v}\in[0,1]^n} E(\boldsymbol{w},\boldsymbol{v};\lambda)=\inf_{\boldsymbol{w},\boldsymbol{v}\in[0,1]^n}\{\langle\boldsymbol{v},\boldsymbol{l}(\boldsymbol{w})\rangle+\mathrm{R}_{SP}(\boldsymbol{v},\lambda)+R(\boldsymbol{w})\}$$

年龄参数
（路径算法变量）

加权误差

自步
正则项

模型参数
正则项

前人工作

$$v^*(\lambda, l) = \arg \inf_{v \in [0,1]} \{vl + R_{SP}(v, \lambda)\}$$

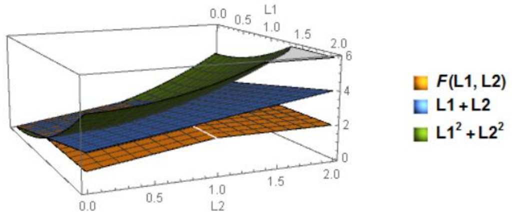
Meng 等人， 从优化算法的角度研究了 $v^*(\lambda, l)$与优化目标的关系。

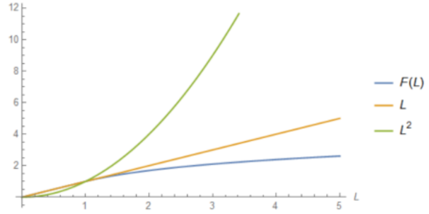
（交替迭代下降方法和优化最小化方法优化）

发现了隐式目标函数 $F_\lambda(\ell) = \int_0^\ell v^*(\lambda; l)dl$

稳健的损失函数

非凸惩罚正则（NCPR）





Fan 等人， 利用半二次优化研究了 $R_{SP}(v, \lambda)$ 与 $v^*(\lambda, l)$关系 提出了隐式正则项

Meng D, Zhao Q. What Objective Does Self-paced Learning Indeed Optimize?[J]. Computer Science,, 2015.

Fan Y, He R, Liang J, et al. Self-Paced Learning: an Implicit Regularization Perspective[J]. arXiv preprint arXiv:1606.00128, 2016.

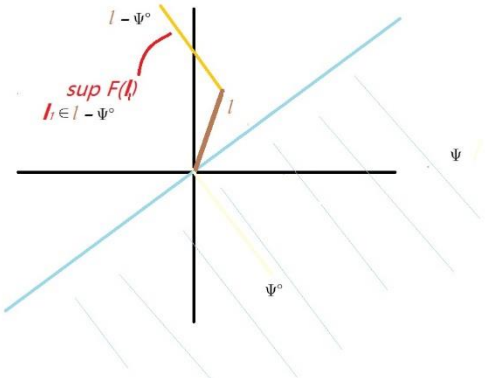
我的工作：自步学习凹共轭性

目标函数 $$\inf_{\boldsymbol{w},\boldsymbol{v}\in[0,1]^n} E(\boldsymbol{w},\boldsymbol{v};\lambda) = \inf_{\boldsymbol{w},\boldsymbol{v}\in[0,1]^n}\{\langle\boldsymbol{v},\boldsymbol{l}(\boldsymbol{w})\rangle + \mathrm{R}_{SP}(\boldsymbol{v},\lambda) + R(\boldsymbol{w})\}$$

$$= \inf_{\boldsymbol{w}}\{R(\boldsymbol{w}) + g^*(\boldsymbol{l}(\boldsymbol{w}))\} = \inf_{\boldsymbol{w}}\{R(\boldsymbol{w}) + F_\lambda(\boldsymbol{l}(\boldsymbol{w}))\}$$

统一简约的理论分析方法

课程函数 $$F^{new}(\boldsymbol{l}) = \inf_{\boldsymbol{v}\in\mathrm{R}^n}\langle\boldsymbol{v},\boldsymbol{l}\rangle + R_{SP}(\boldsymbol{v}) - C(\boldsymbol{v}) = \left(g(\boldsymbol{v}) + \boldsymbol{C}(\boldsymbol{v})\right)^* = F \oplus C^*(\boldsymbol{l})$$



课程区域 $$F^{new}(\boldsymbol{l}) = \inf_{\boldsymbol{v}\in\Psi}\langle\boldsymbol{v},\boldsymbol{l}(\boldsymbol{w})\rangle + R_{SP}(\boldsymbol{v}) = \inf_{\boldsymbol{v}\in[0,1]^n}\langle\boldsymbol{v},\boldsymbol{l}(\boldsymbol{w})\rangle + R_{SP}(\boldsymbol{v}) - \delta(\boldsymbol{v}|\boldsymbol{\Psi}) = F \oplus \delta^*(\cdot\,|\boldsymbol{\Psi})(\boldsymbol{l})$$

优化角度的总结

- ❑ 凹共轭理论 自步学习 建立联系
- ❑ 权重函数 隐藏目标函数 自步正则项 直接关系
- ❑ 自步学习的先验（正则项） 优化理解
- ❑ 对自步课程学习先验优化作用分析有指导
- ❑ 完成一篇论文：自步学习凹共轭性（修改中）
- ❑ 此外与马子璐完成自步学习收敛性论文投稿至中国科学

Theorem 6 (Model Equivalence). *In one dimension case of v, if $R_{SP}(v,\lambda)$ satisfy the assumption 1 and be strictly convex, then*

$$F_\lambda(l) = \int_0^l v(\lambda, j)\, dj + C(\lambda)$$

where $C(\lambda)$ is a function in λ.

Theorem 7 (Relations). *If $R_{SP}(v,\lambda)$ satisfy the assumption 1, then*

$$\begin{aligned}
l_\lambda(v) &= \partial_v(-R_{SP}(v,\lambda)) && (9)\\
v(\lambda,l) &= l_\lambda^{-1}(l) && (10)\\
v(\lambda,l) &= \partial F_\lambda(l) && (11)\\
F_\lambda(l) &= \langle v(\lambda,l), l\rangle + R_{SP}(v(\lambda,l),\lambda) && (12)\\
R_{SP}(v,\lambda) &= \langle v, l_\lambda(v)\rangle - R_{SP}(v,\lambda)(l_\lambda(v)) && (13)
\end{aligned}$$

If $R_{SP}(v,\lambda)$ and $F_\lambda(l)$ is srtictly convex in v and l respectively and one dimension situation is considered, we can further obtain

$$\begin{aligned}
F_\lambda(l) &= \int_0^l v(\lambda,j)\, dj + C(\lambda) && (14)\\
R_{SP}(v,\lambda) &= -\int_0^v l_\lambda(j)\, dj + C(\lambda) && (15)
\end{aligned}$$

Theorem 10 (Action of Linear Homogeneous Curriculum). *Suppose $R_{SP}(v)$ is essential strictly convex and satisfies assumption 1. Suppose we have knowledge of the weight variable v, denoted by $v^T k \geq 0$ corresponding to $\Psi = \{v | v^T k \geq 0\}$. If Ψ satisfies assumption 2, then $\nabla F^{new}(l)^T k = v^{new}(l)^T k \geq 0$ and*

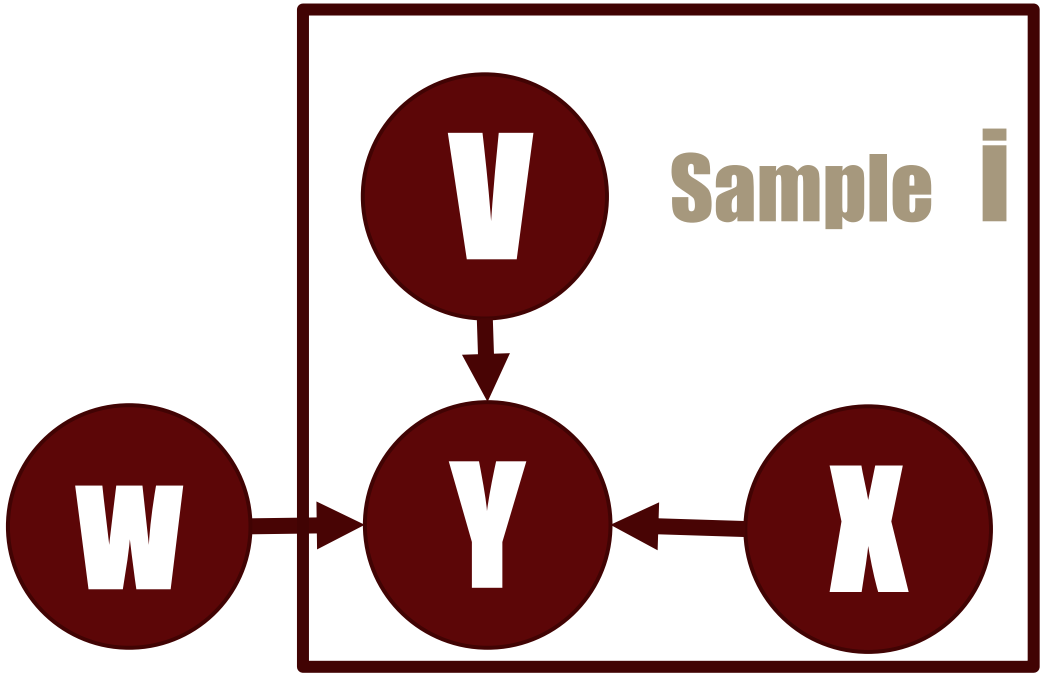
$$F^{new}(l) = \begin{cases} F(l) & l \in \partial(-R_{SP})(k^\perp) - ray_k \\ \sup\limits_{l' \in l + line_k} F(l')(\geq F(l)) & l \in \partial(-R_{SP})(k^\perp) + ray_k \end{cases}$$

.

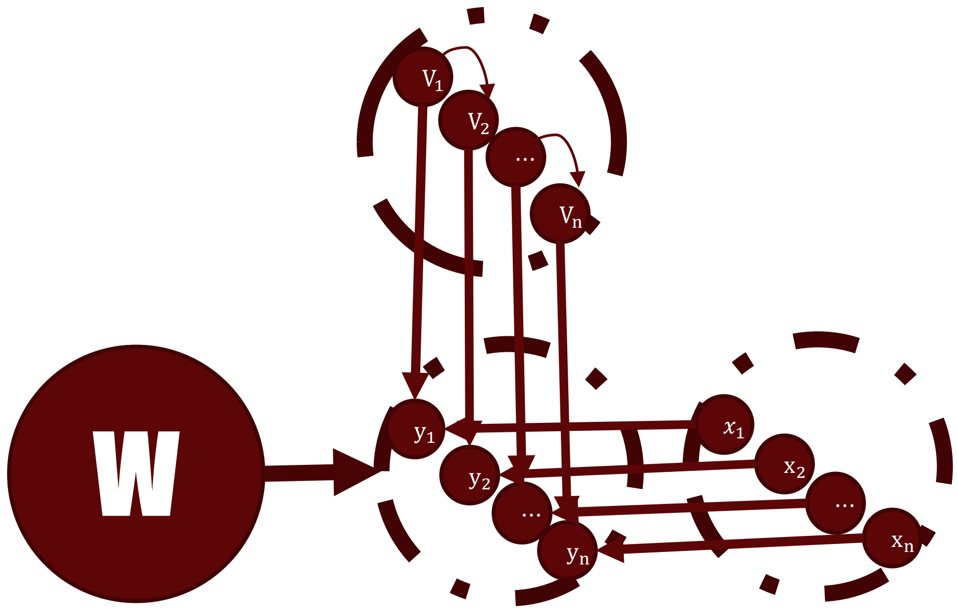
回归理解

自步学习的贝叶斯网

连续 $Y|(X,W,V)\sim\mathcal{N}(f(X,W),\frac{1}{2V})$

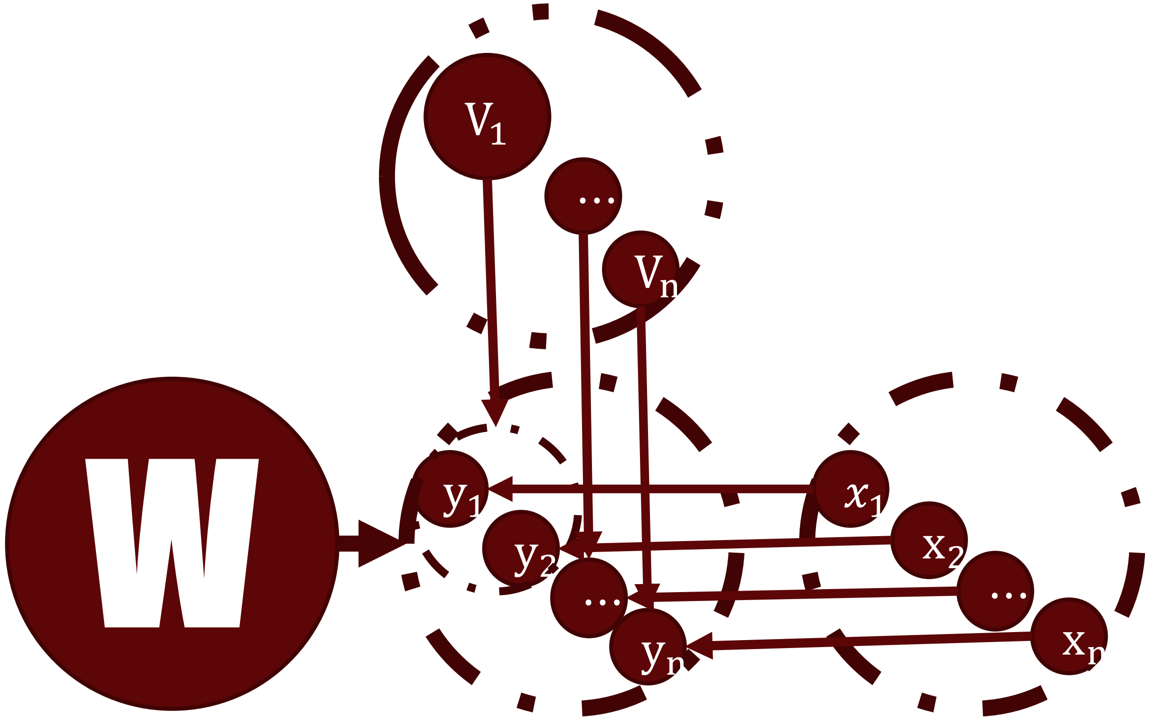


- 非IID情况下，赋予权重变量精度的物理意义
- 帮助先验信息的嵌入



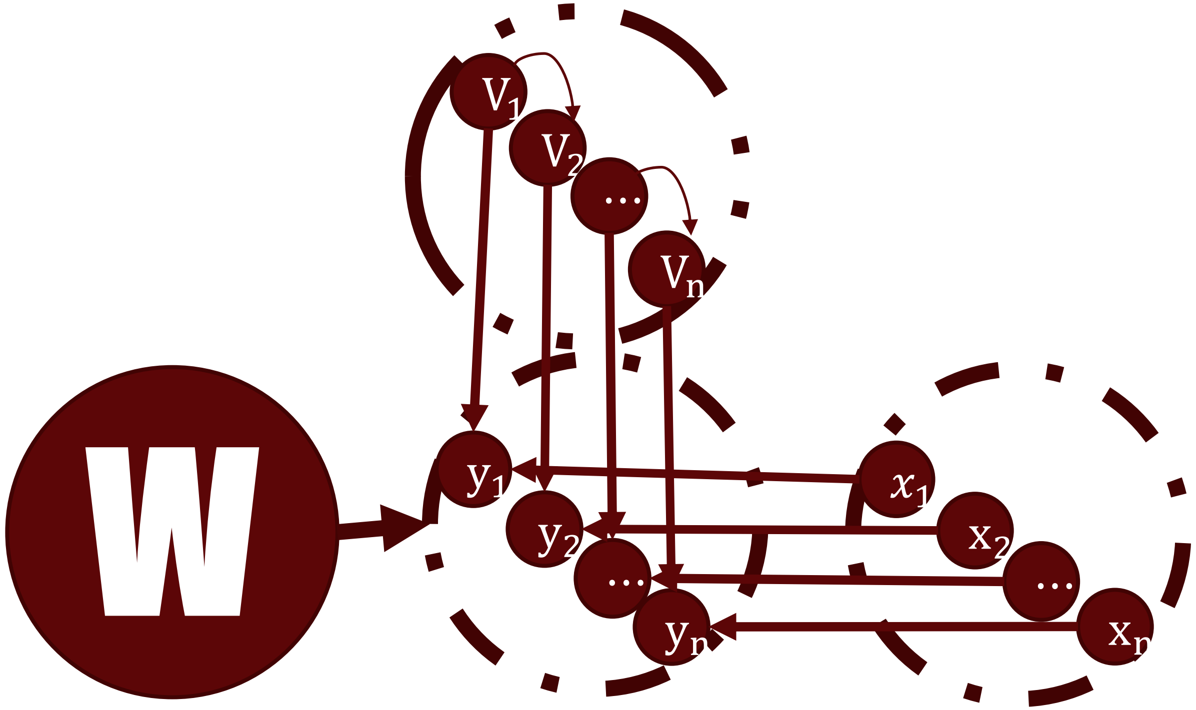
总体先验

V1=V2



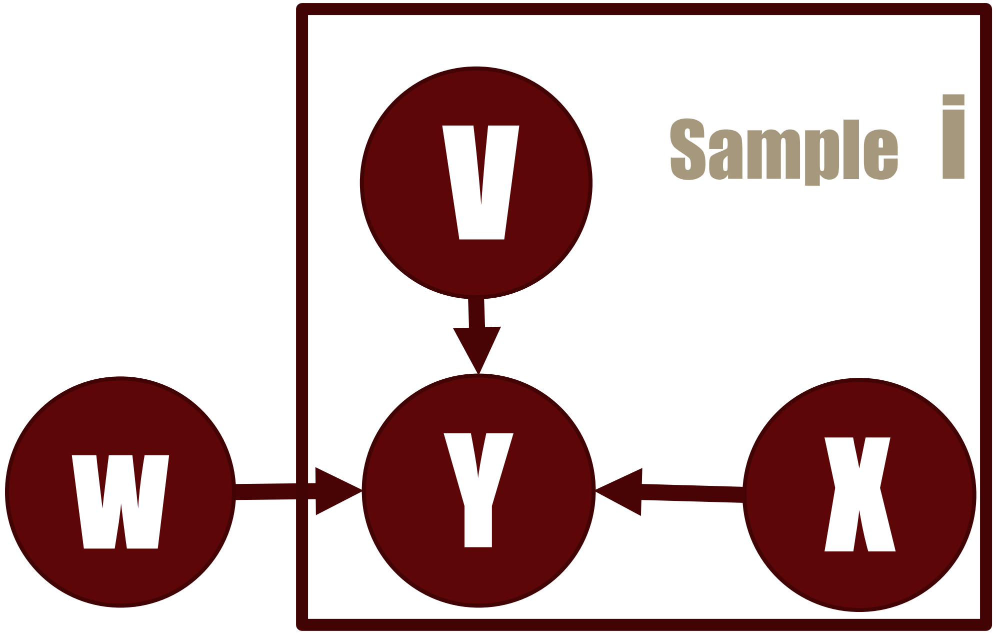
偏序先验

V1>V2, V3>V4



贝叶斯网　自步学习模型 实验设计

连续　$Y|(X,W,V) \sim \mathcal{N}(f(X,W),\frac{1}{2V})$



$$Y = A\,Signal(X-T)+\varepsilon$$

$$\varepsilon|v \sim \sqrt{\frac{v}{\pi}}e^{-vl^2}$$

$$v \sim U[0,1] \sim e^{\delta[v|[0,1]]}$$

$$\inf_{T,\boldsymbol{v}\in[0,1]^n} E(\mathrm{T},\boldsymbol{v})$$

$$= \inf_{T,\boldsymbol{v}\in[0,1]^n} \sum_{i=1}^{N}\left(v_i l_i(T) - \frac{1}{2}\log v_i\right)$$

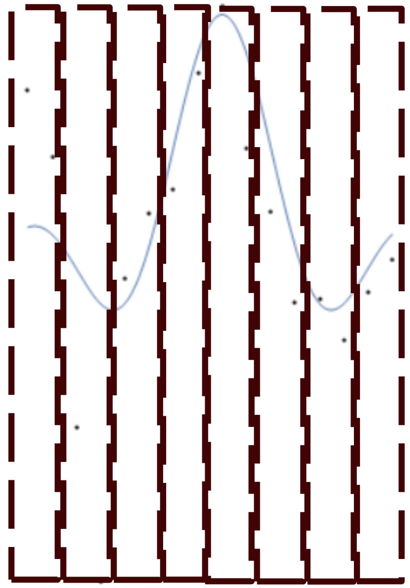
$$= \inf_{T} F(\boldsymbol{l}(T))$$

各样本上误差 $l^i = \left(y^i - ASignal(x-T)\right)^2$

实验效果

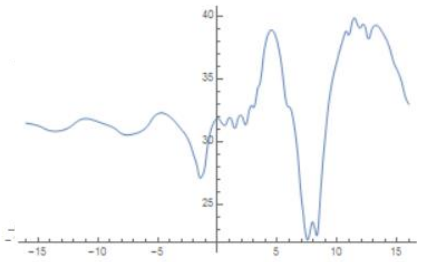
$Y = 10\,Signal(X - \mathrm{T}) + \varepsilon$ 回归出参数值为$T = 8$

八测量仪器



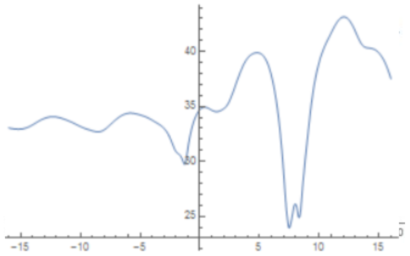
ε^1 ε^8

$Var\,\varepsilon^1 > \cdots > Var\,\varepsilon^8$

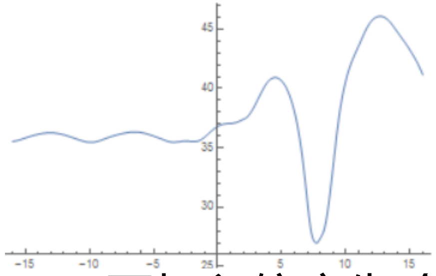


不加入先验

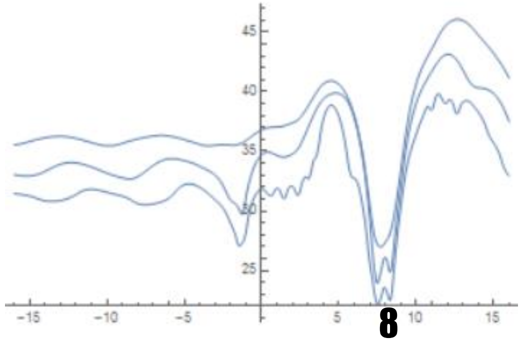
$F(l(T))$



加入总体先验



再加入偏序先验



总对比图

回归角度的总结

- 图模型所描述的独立性关系，在连续的情况有对应的**物理意义**。
- 由于各样本有**自己的精度参数**，属于小样本参数估计的情况，须利用贝叶斯最大后验方法加入较强先验。
- 图模型结构描述的分布可以是**混合高斯分布的推广**
- **总体先验和偏序先验**在图模型的意义下被提出。也许可以导出更多先验。
- 实验角度，先验的加入，从经验上使得了目标函数**光滑、局部极小数量减少**

概率理解

难易程度标准确定：损失概率化

自步学习 描述 难易程度

可靠性 **不确定性**

损失	分布	名称	应用
$(y-f(x))^2$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(y-f(x))^2}{2\sigma^2}}$	Gauss	连续回归
$\|y-f(x)\|$	$\frac{1}{2b}e^{\frac{-\|y-f(x)\|}{b}}$	Laplace	连续回归
$\\|y-f(x)\\|_{L^p}$	$\frac{1}{\Gamma(1+\frac{1}{p})b^{\frac{1}{p}}}e^{\frac{-\\|y-f(x)\\|_{L^p}^p}{b}}$	L^p	连续回归
$\log(1+e^{-\frac{yf(x)}{b}})$	$\frac{e^{\frac{yf(x)}{2b}}}{\sum_{i\in\{-1,1\}} e^{\frac{if(x)}{2b}}}$	Logistic	二分类
$\log(\frac{e^{\langle\beta_y,f(x)\rangle}}{\sum_{i\in\mathcal{Y}} e^{\langle\beta_i,f(x)\rangle}})$	$\frac{e^{\langle\beta_y,f(x)\rangle}}{\sum_{i\in\mathcal{Y}} e^{\langle\beta_i,f(x)\rangle}}$	Multinomial Logistic	多分类

生成模型下损失和联合概率分布的对应关系： $p_{model\,XY}(x,y)=e^{-\alpha L(f,(x,y))+\beta}$

判别模型下损失和条件概率分布的对应关系： $p_{model\,Y|X}(y|x)=e^{-\alpha L(f,(x,y))+\beta}$

难易程度标准

难易程度： 智能对于随机变量 $Y|X=x$ 的描述长度， $H(p_{model\,Y|X=x})$

可靠性： 智能对于事件 $(Y=y|X=x)$ 或 $(X=x,Y=y)$ 的发生概率的预测值 $p_{model\,XY}(x,y)$，或者 $p_{model\,Y|X}(y|x)$ 。

自步学习的最优表现的上界

半监督设定

$$\bar{Y} = decision(X, X_{labelled}, Y_{labelled}, Agent_{initial}, Prior)$$

$$P_{error} = Pr(Y_{unlabelled} \neq \bar{Y})$$

利用法诺不等式可以得到，

$$H(P_{error}) + P_{error} \log|\mathcal{Y}| \geq H(Y_{true}|\bar{Y}) \geq H(Y_{unlabelled}|X, X_{labelled}, Y_{labelled}, Agent_{initial}, Prior)$$

$$P_{error} \geq \frac{H(Y_{unlabelled}|X, X_{labelled}, Y_{labelled}, Agent_{initial}, Prior) - 1}{\log|\mathcal{Y}|}$$

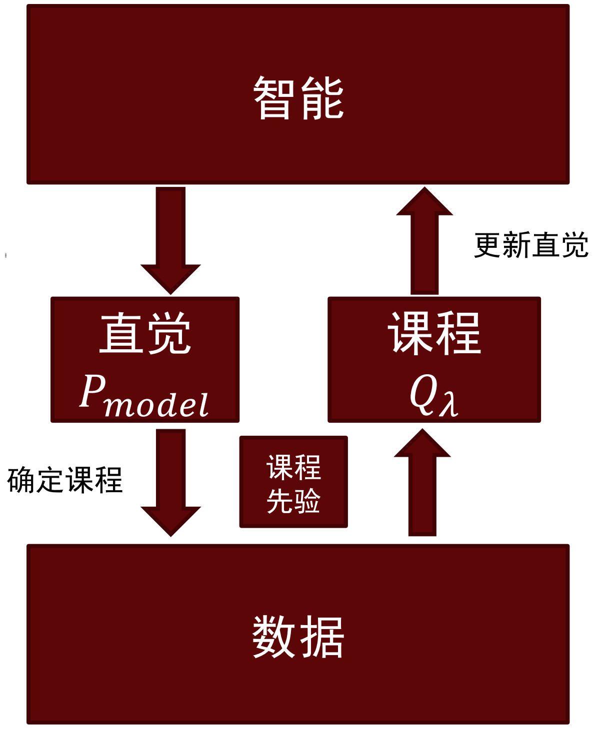
影响因子：**智能初始的经验，外部先验知识，数据本身的性质**

自上而下概率分布学习

通用自步学习概率模型

$$\min_{Q_\lambda \in \mathcal{Q}_\lambda, P_{model} \in \mathcal{P}_{model}} D(Q_\lambda || P_{model}) + R_{\mathcal{Q}_\lambda}(Q_\lambda) + R_{\mathcal{P}}(P_{model})$$

- 对数据经验分布重新建模
- 帮助先验信息的嵌入
- 与复杂化模型分布相反的道路
- 路径算法带来模型稳健性



λ控制课程难易程度

课程设定

课程设定：指定学习样本的数量

$$\min_{Q_\lambda, W\in\mathcal{W}} \quad E_{x\sim Q_{\lambda X}} D(Q_{\lambda Y|X=x}||p_{model Y|X=x;W}) - H(Q_\lambda) + R_{\mathcal{W}}(W)$$

$$\begin{aligned} s.t. \quad & v_i \geq 0, \quad \forall i=1,\cdots,n \\ & \textstyle\sum_{i=1}^n v_i = 1 \\ & \|V\|_0 = \lambda \quad 。\end{aligned}$$

λ 学习样本的数量

课程设定：指定学习样本的可靠性阈值

$$\min_{v\in[0,1]^n, W\in\mathcal{W}} \quad \sum_{i=1}^{n} -v_i \log p_{model Y|X;W}(y_i|x_i) + R_{\mathcal{W}}(W) + (\log\lambda)\|V\|_1 。$$

λ 学习样本可靠度/条件熵阈值

课程设定：指定学习样本的数量和等可靠性

$$\min_{Q_\lambda, W\in\mathcal{W}} \quad E_{x\sim Q_{\lambda X}} D(Q_{\lambda Y|X=x}||p_{model Y|X=x;W}) - H(Q_\lambda) + R_{\mathcal{W}}(W)$$

$$\begin{aligned} s.t. \quad & v_i \geq 0, \quad \forall i=1,\cdots,n \\ & \textstyle\sum_{i=1}^n v_i = 1 \\ & v_i = 0 \ or \ v_i = \tfrac{1}{\lambda} \quad \forall i=1,\cdots,n。\end{aligned}$$

课程设定：指定学习样本的难易程度阈值

$$\min_{v\in[0,1]^{n_{unlabelled}}, W\in\mathcal{W}} \quad E_{(x,y)\sim p_{label XY}} \log\frac{1}{p_{model Y|X;W}(y|x)} + R_{\mathcal{W}}(W)$$

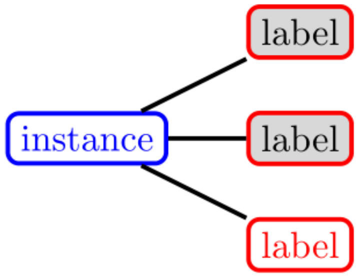
$$+\eta\Big(\sum_{i=1}^{n_{unlabelled}} v_i H(p_{model Y|X=x_i;W}) - \lambda\|V\|_1\Big)。$$

概率角度的意义总结

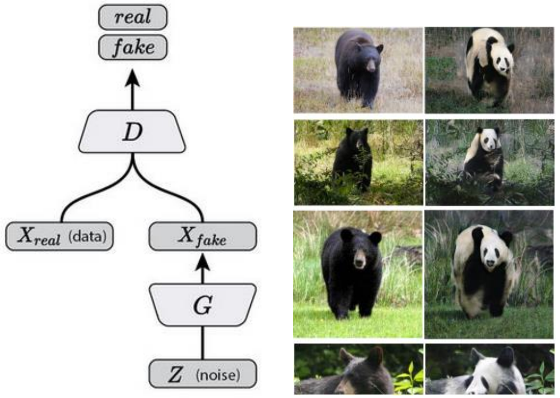
- 难易程度$\propto$**不确定性，可靠性**，为自步学习在学习材料有不确定因素任务（半监督、弱监督和稳健学习）取得好效果提供了解释，说明难易程度标准一定程度刻画任务的本质信息
- 概率框架下，自步学习充分利用 **智能直觉** $\boldsymbol{p_{model}}\,(\mathbf{y}|\mathbf{x})$、**学习材料**、**先验知识**
- 显著作用:**降低**对于**学习材料**和课程设置的需求，并附带**稳健**的样本选择（**课程生成**）准则
- 提出了自步学习通用的概率框架，能够解释并退化到并原有自步学习模型
- 解释部分**外部先验知识**的概率含义，提出了一种新的类别成分先验，并对先验嵌入有指导作用

展望

自步学习未来的一些潜在方向



自步学习与歧义学习



自步学习与预测学习

自步学习与好奇心

自步学习与主动学习

自步学习与迁移学习

样本标记更正

训练老师

课程探索

部分图像资源来自https://github.com/tatsuyah/CycleGAN-Models　Timothee Cour et al, JMLR 12 (2011) 1501-1536

Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven Exploration by Self-supervised Prediction[J]. arXiv preprint arXiv:1705.05363, 2017.

致谢

感谢机器学习小组

谢谢各位老师和同学的倾听和指教！

辅助材料

主要方法论

函数$g(\boldsymbol{v})$的凹共轭变换定义为

$$g^*(\boldsymbol{l}) = \inf_{\boldsymbol{v}\in R^n}\{\langle \boldsymbol{v}, l\rangle - g(\boldsymbol{v})\}$$

最大卷积 $$\mathrm{p} \oplus q(\boldsymbol{l}) = \sup_{l^1+l^2=l}\{p(l^1) + q(l^2)\}$$

凹共轭下运算对偶 加法⇔ 最大卷积

$$(p+q)^* = p^* \oplus q^* \qquad (\mathrm{p} \oplus q)^* = p^* + q^*$$

误差先验的影响|课程学习

研究方法　　凹共轭下运算对偶 加法⟺ 最大卷积　　最大卷积$p \oplus q(\boldsymbol{l}) = \sup_{l^1+l^2=l} \{p(l^1) + q(l^2)\}$

$$(p+q)^* = p^* \oplus q^* \qquad (\mathrm{p} \oplus q)^* = p^* + q^*$$

课程函数

$$F^{new}(\boldsymbol{l}) = \inf_{\boldsymbol{v} \in \mathrm{R}^n} \langle \boldsymbol{v}, \boldsymbol{l} \rangle + f(\boldsymbol{v}) - C(\boldsymbol{v}) = \left(g(\boldsymbol{v}) + \boldsymbol{C}(\boldsymbol{v})\right)^* = \boldsymbol{F} \oplus \boldsymbol{C}^*(\boldsymbol{l})$$

课程函数
（误差先验）

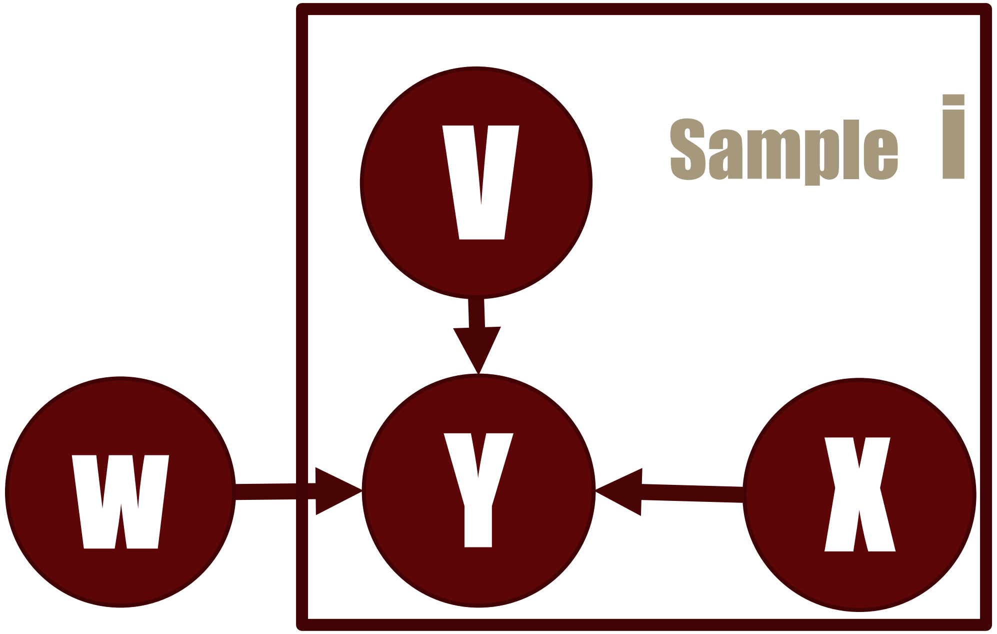
课程区域

$$F^{new}(\boldsymbol{l}) = \inf_{\boldsymbol{v} \in \Psi} \langle \boldsymbol{v}, \boldsymbol{l}(\boldsymbol{w}) \rangle + f(\boldsymbol{v}) = \inf_{\boldsymbol{v} \in [0,1]^n} \langle \boldsymbol{v}, \boldsymbol{l}(\boldsymbol{w}) \rangle + f(\boldsymbol{v}) - \delta(\boldsymbol{v}|\boldsymbol{\Psi}) = \boldsymbol{F} \oplus \boldsymbol{\delta}^*(\cdot\,|\boldsymbol{\Psi})(\boldsymbol{l})$$

课程区域
（误差先验）

贝叶斯网　自步学习模型 实验设计

连续　$Y|(X,W,V)\sim\mathcal{N}\left(f(X,W),\frac{1}{2V}\right)$



$$Y = A\,Signal(X-T)+\varepsilon$$

$$\varepsilon|v \sim \sqrt{\frac{v}{\pi}}e^{-vl^2}$$

$$v \sim U[0,1] \sim e^{\delta[v|[0,1]]}$$

$$\inf_{T,\boldsymbol{v}\in[0,1]^n} E(\mathrm{T},\boldsymbol{v})$$

$$= \inf_{T,\boldsymbol{v}\in[0,1]^n} \sum_{i=1}^{N}\left(v_i l_i(T) - \frac{1}{2}\log v_i\right)$$

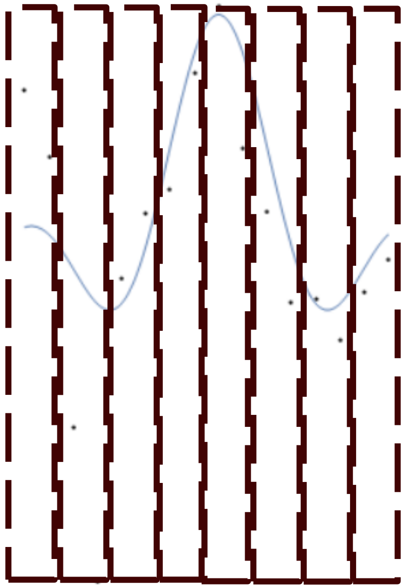
$$= \inf_{T} F(\boldsymbol{l}(T))$$

各样本上误差 $l^i = \left(y^i - ASignal(x-T)\right)^2$

实验效果

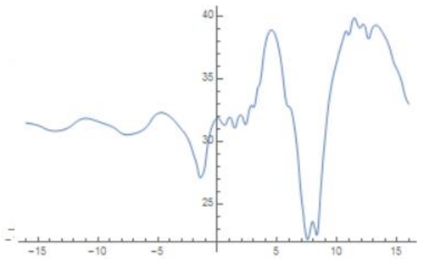
$Y = 10\, Signal(X - \mathrm{T}) + \varepsilon$ 回归出参数值为$T = 8$

八测量仪器



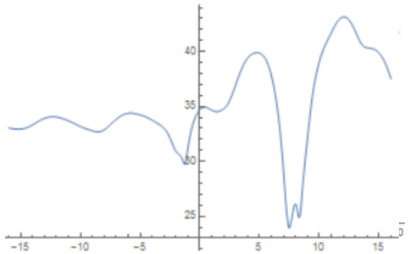
ε^1 ε^8

$Var\, \varepsilon^1 > \cdots > Var\, \varepsilon^8$

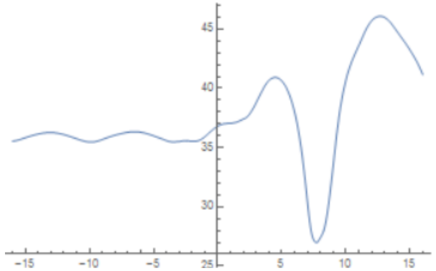


不加入先验

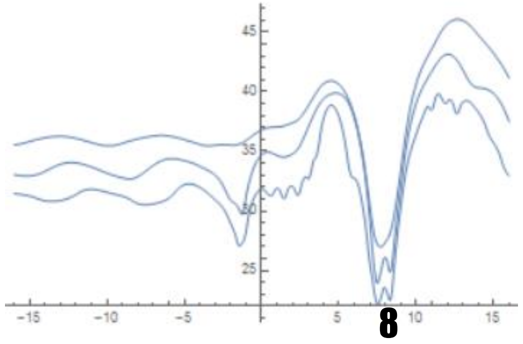
$F(l(T))$



加入总体先验



再加入偏序先验



总对比图

课程先验

多样性先验：

利用概率框架的一种类别成分先验 $D(p_{true_Y} || Q_{\lambda_Y})$

导出了诱导多样性的正则项 $-\sum_{i=1}^{b} p_i \log \|v^{(i)}\|_{L^1}$

关联性先验：

$$p_{mix}(y|x) = \sum_{i=1}^{K} \pi_i p_{agent_i}(y|x)$$