高级机器学习

半监督学习

大纲

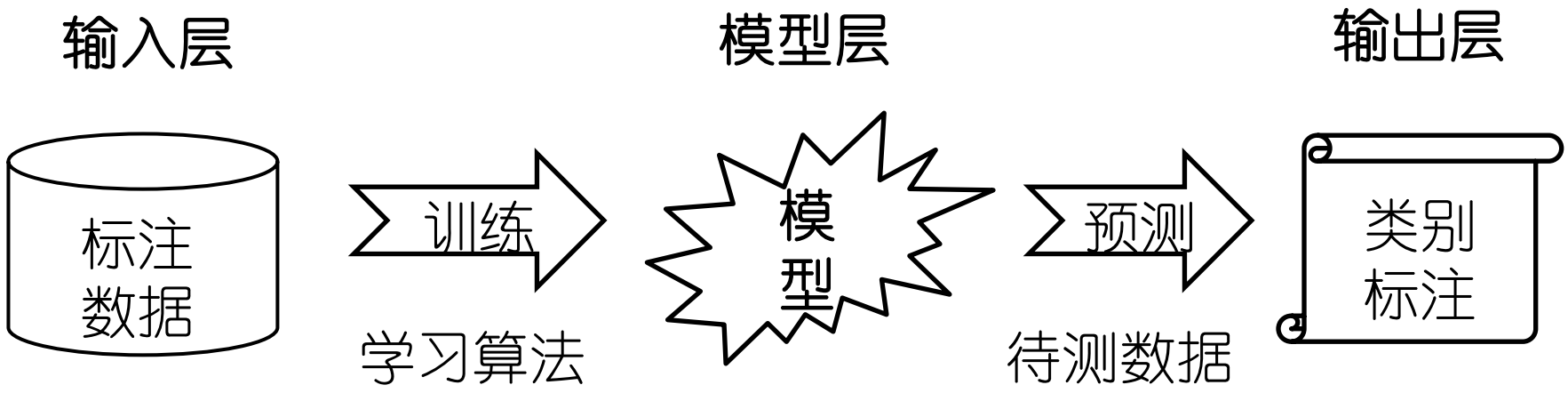
- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

典型机器学习过程

利用标注数据训练模型，在未见测试数据上输出正确的类别标注



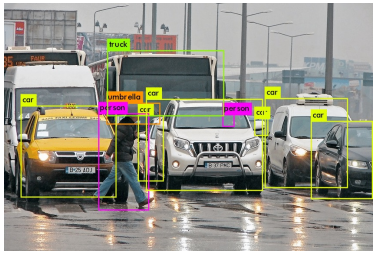
既有机器学习技术取得强泛化性能
依赖大量标注数据

标注获取困难

数据标注获取耗费大量的人力、物力、财力

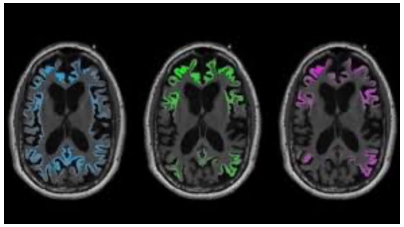
现实任务往往<u>无标注数据多，标注少</u>

目标检测



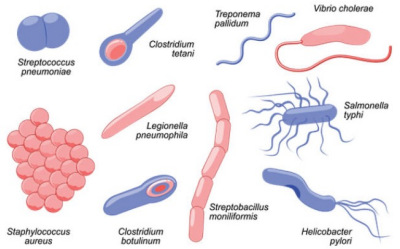
精细标注**难**

医疗诊断



专家知识**缺**

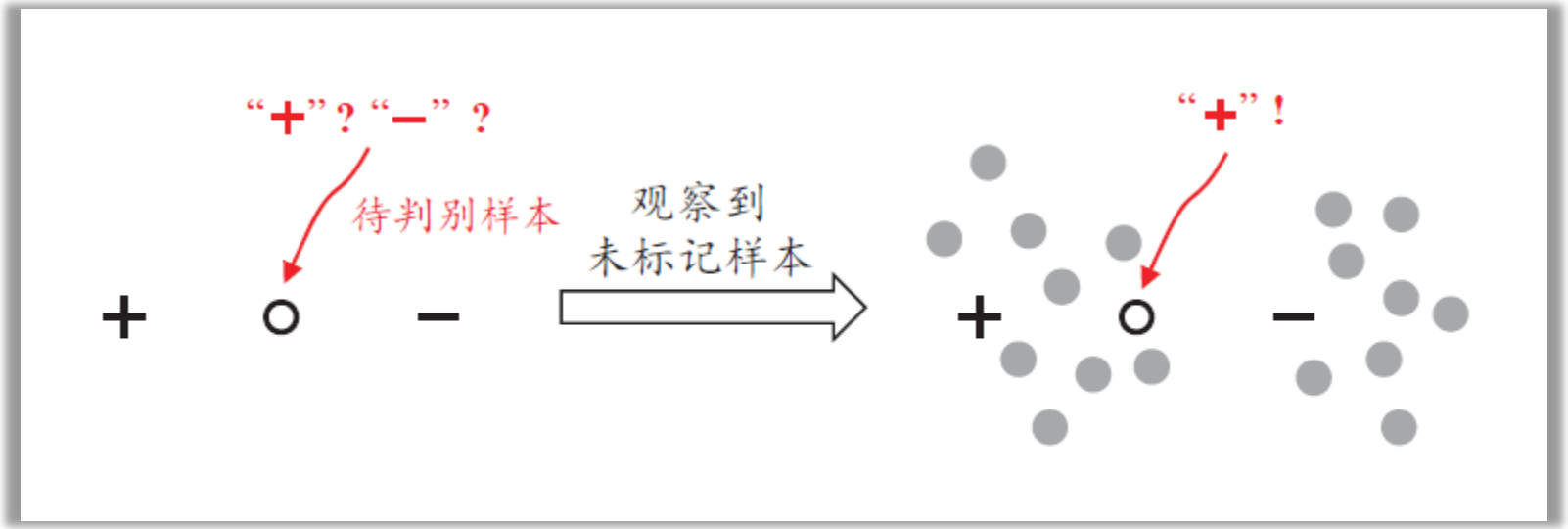
生物识别



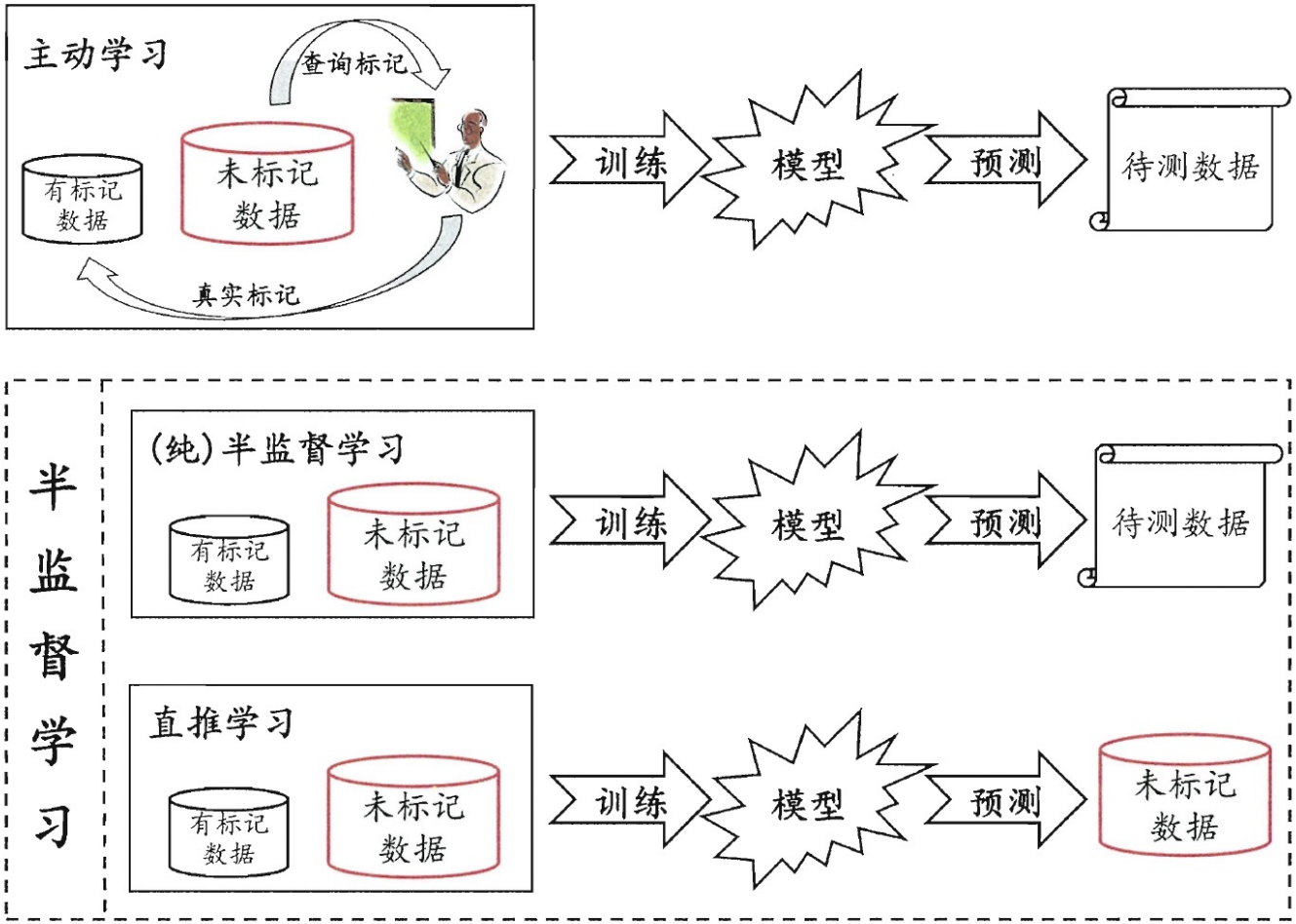
专业设备**贵**

无标注样本

能否同时利用有标注样本和无标注样本
构建泛化性能良好的模型？



无标注样本利用


主动学习
查询标记
有标记数据
未标记数据
真实标记
训练
模型
预测
待测数据
半监督学习
(纯)半监督学习
有标记数据
未标记数据
训练
模型
预测
待测数据
直推学习
有标记数据
未标记数据
训练
模型
预测
未标记数据

无标注样本的潜在假设

- 要使得半监督学习奏效，我们需要对未标记数据做假设，刻画它们内蕴**数据分布与类别标记的联系**
- 两种常见假设
  - 聚类假设（clustering assumption）：
    - 假设数据存在簇结构，同一簇的样本属于同一类别
  - 流形假设（manifold assumption）：
    - 假设数据分布在一个流形结构上，邻近的样本具有相似的输出值

主流的半监督学习方法就是在合理假设上成功构建的

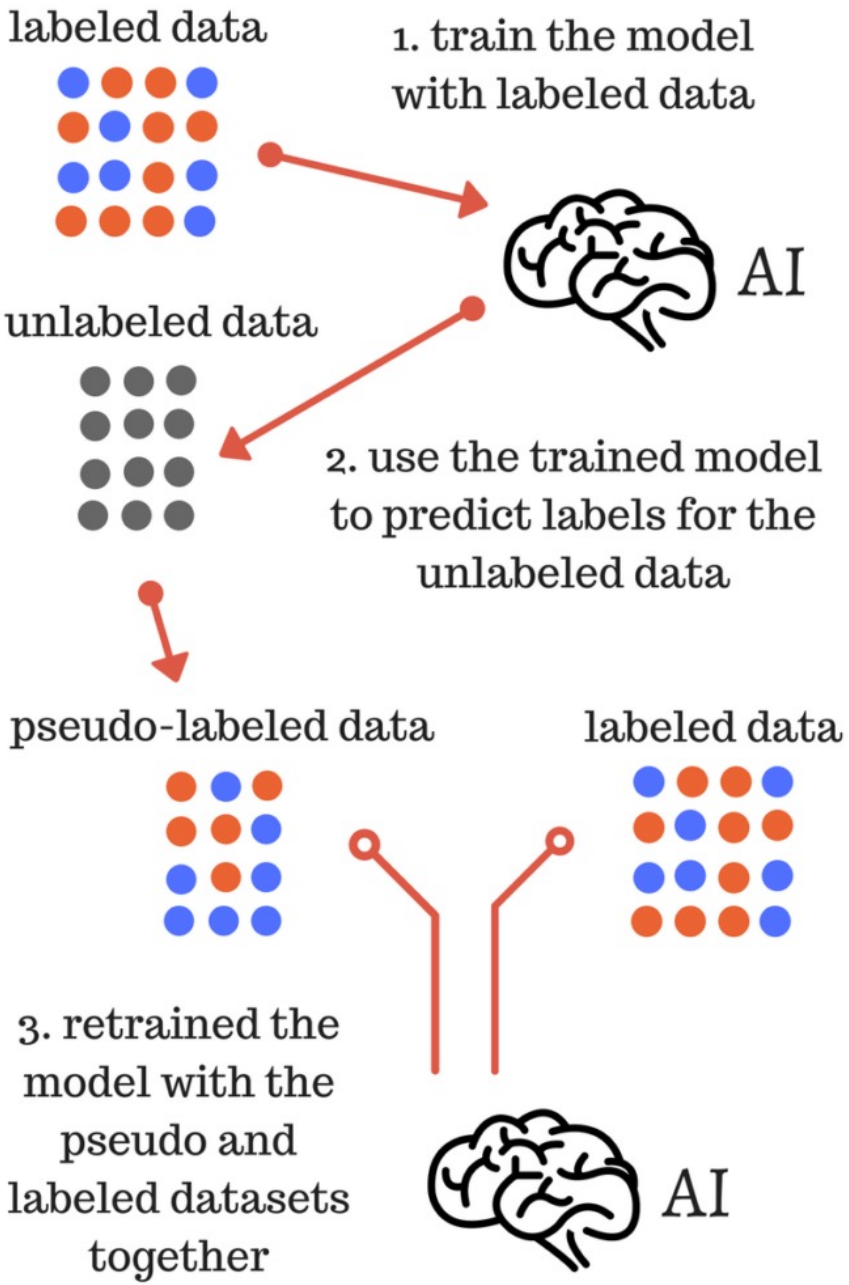
大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

自训练(Self-Training)

- 基本假设：
  - 置信度较高的预测结果是正确的
- 算法流程：
  - 基于有标注数据训练一个机器学习模型
  - 用这个模型对无标注数据进行分类，产生伪标注(pseudo label)
  - 挑选认为分类正确的无标签样本（此处应该有一个挑选准则，例如预测置信度）
  - 把选出来的伪标注样本加入有标注数据集重新训练模型
  - 迭代上述过程，直至满足停止条件

自训练(Self-Training)



自训练(Self-Training)的变种

- 挑选认为分类正确的无标签样本加入有标注数据集
  - 选择少量预测置信度较高的无标注样本
  - 选择所有的无标注样本
  - 选择所有的无标注样本，基于预测置信度对样本进行加权
  - …

自训练(Self-Training)

- 优势：
  - 最简单的半监督学习算法
  - 通用性强，可与各种机器学习算法结合
- 劣势：
  - 错误累计
  - 收敛性难判断

大纲

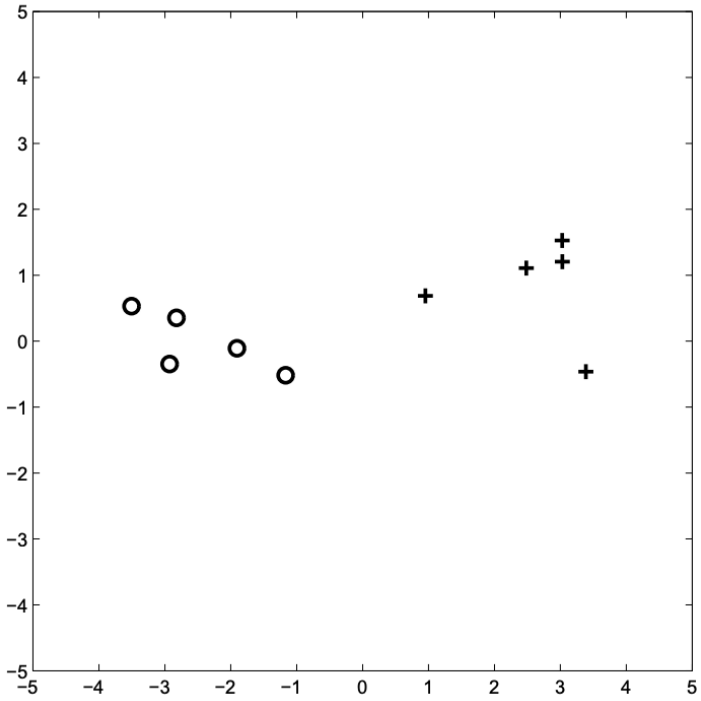
- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

生成式方法(generative methods)

- 基本假设：
  - 所有数据（无论是否有标注）都是由一个潜在模型生成的
- 算法流程：
  - 给定潜在生成模型，基于有标注数据和无标注数据去估计潜在模型的参数（通常可基于EM算法进行极大似然估计）
- 此类方法的区别主要在于生成式模型的假设，不同的模型假设将产生不同的方法

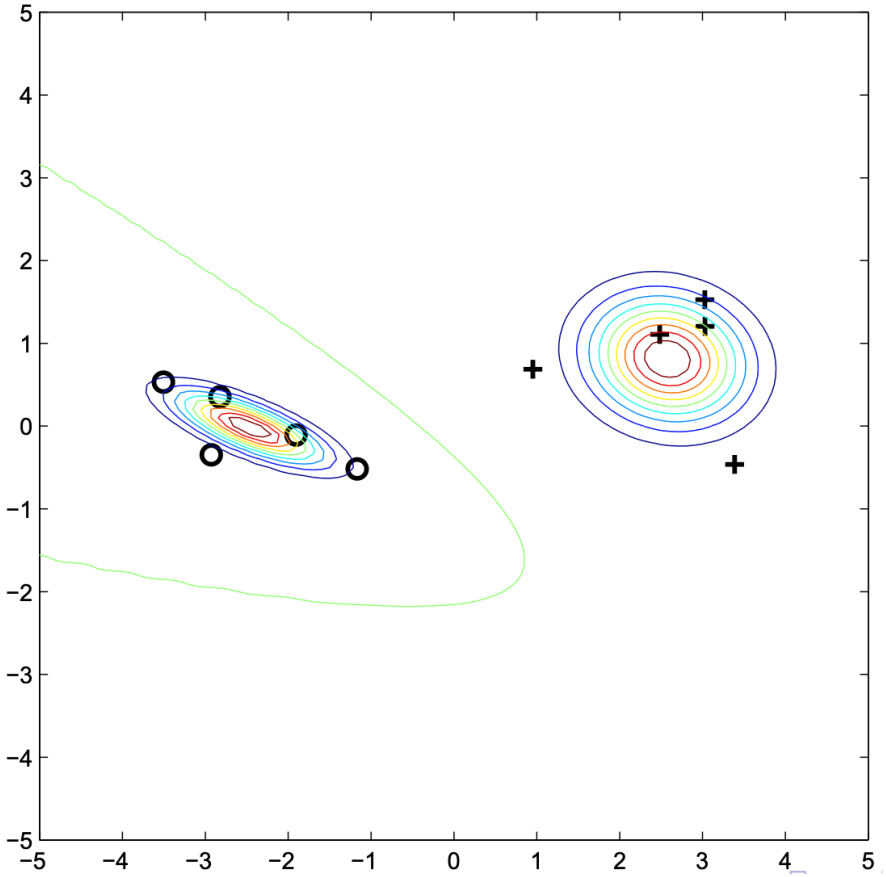
生成式方法(generative methods)

有标注数据



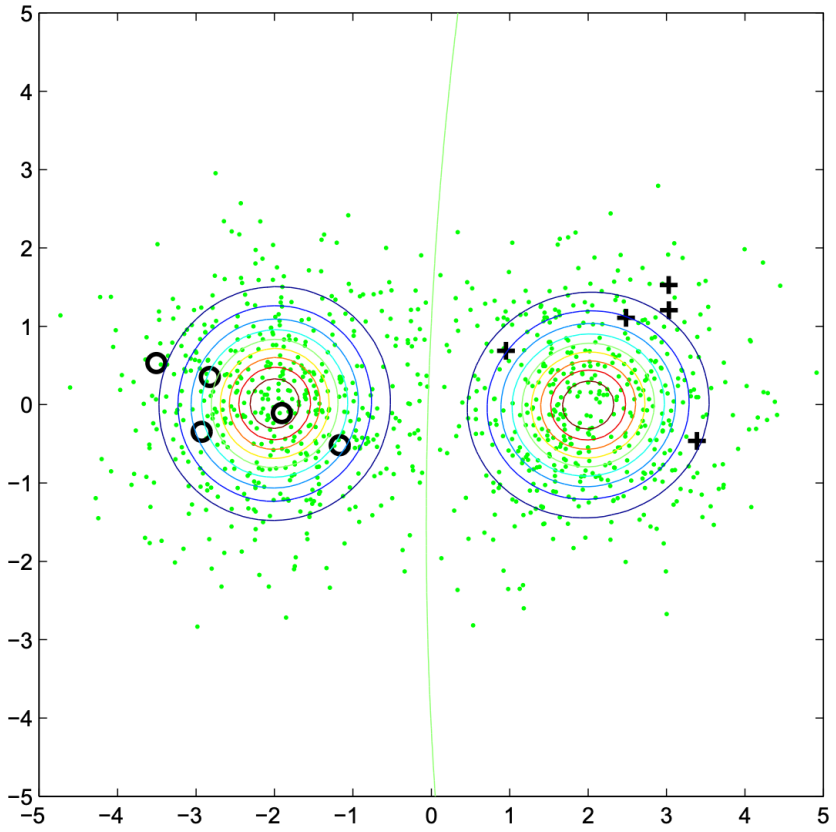
假设每个类对应一个高斯分布，决策边界？

生成式方法(generative methods)



生成式方法(generative methods)

有标注数据+无标注数据



生成式半监督

- 假设样本由且每个类别对应一个高斯混合成分

$$p(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

其中 $\alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1$

$$p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}$$

生成式半监督

- 由最大化后验概率可知：

$$f(\boldsymbol{x}) = \underset{j \in \mathcal{Y}}{\operatorname{argmax}}\, p(y = j|\boldsymbol{x})$$

$$= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y = j, \Theta = i|\boldsymbol{x}) \qquad p(y = j|\Theta = i)$$

$$= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y = j|\Theta = i, \boldsymbol{x}) \cdot p(\Theta = i|\boldsymbol{x})$$

其中 $p(\Theta = i|x) = \frac{\alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$

生成式半监督

- 假设样本独立同分布，且由同一个高斯混合模型生成，则对数似然函数是：

$$\ln p(D_l \cup D_u) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot p(y_j \mid \Theta = i, \boldsymbol{x}_j) \right)$$

$$+ \sum_{\boldsymbol{x}_j \in D_u} \ln \left(\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) .$$

生成式半监督

- 高斯混合的参数估计可以采用EM算法求解，迭代更新式如下：

- M步：根据当前模型参数计算未标记样本x_j属于各高斯混合成分的概率：

$$\gamma_{ji} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

生成式半监督

- M步：基于γ_{ji} 更新模型参数

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_i, y_i) \in D_l \wedge y_j = i} \boldsymbol{x}_j \right)$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \right.$$

$$\left. + \sum_{(\boldsymbol{x}_i, y_i) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \right)$$

$$\alpha_i = \frac{1}{m} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right)$$

生成式方法

- 将上述过程中的高斯混合模型换成混合专家模型，朴素贝叶斯模型等即可推导出其他的生成式半监督学习算法

- 此类方法简单、易于实现，在有标记数据极少的情形下往往比其他方法性能更好

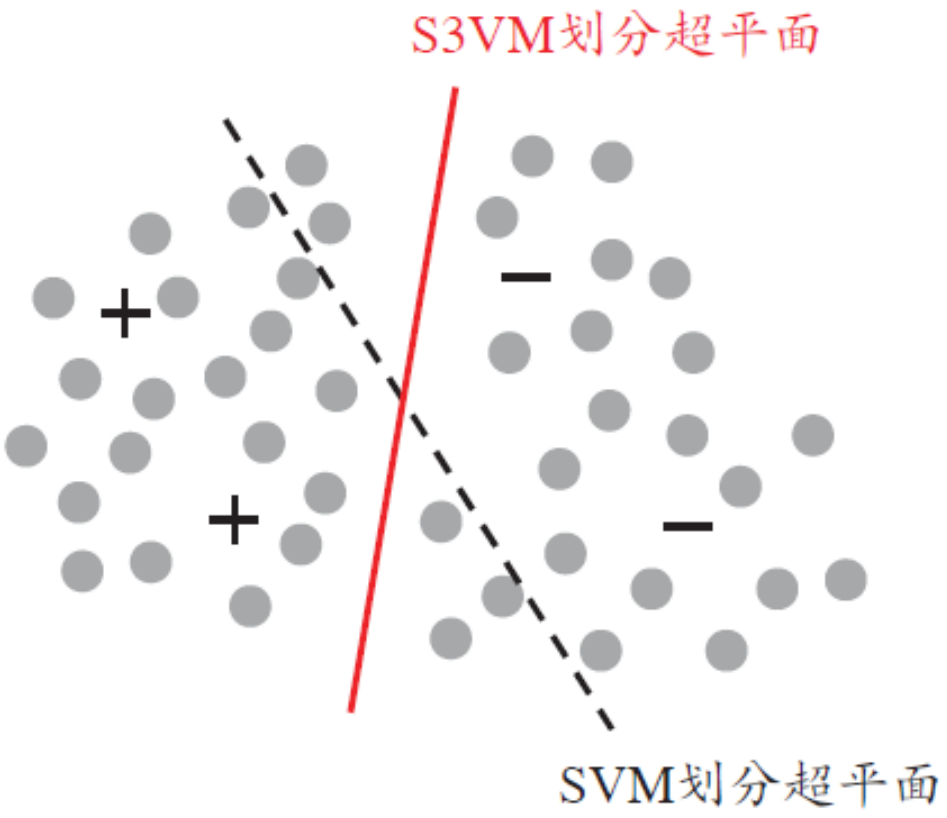
- 然而，此类方法有一个关键：模型假设必须准确，即假设的生成式模型必须与真实数据分布吻合；否则利用未标记数据反而会显著降低泛化性能

大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

半监督支持向量机

基本假设：低密度分隔(low-density separation)



半监督SVM

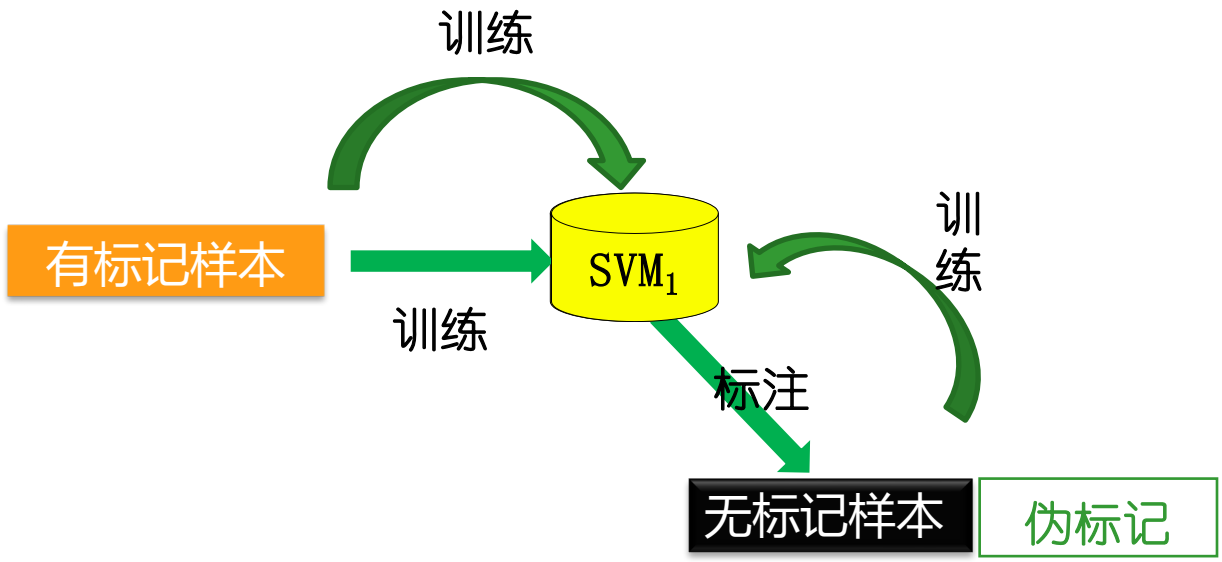
半监督支持向量机中最著名的是TSVM(Transductive Support Vector Machine)

$$\min_{\boldsymbol{w},b,\hat{\boldsymbol{y}},\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i$$

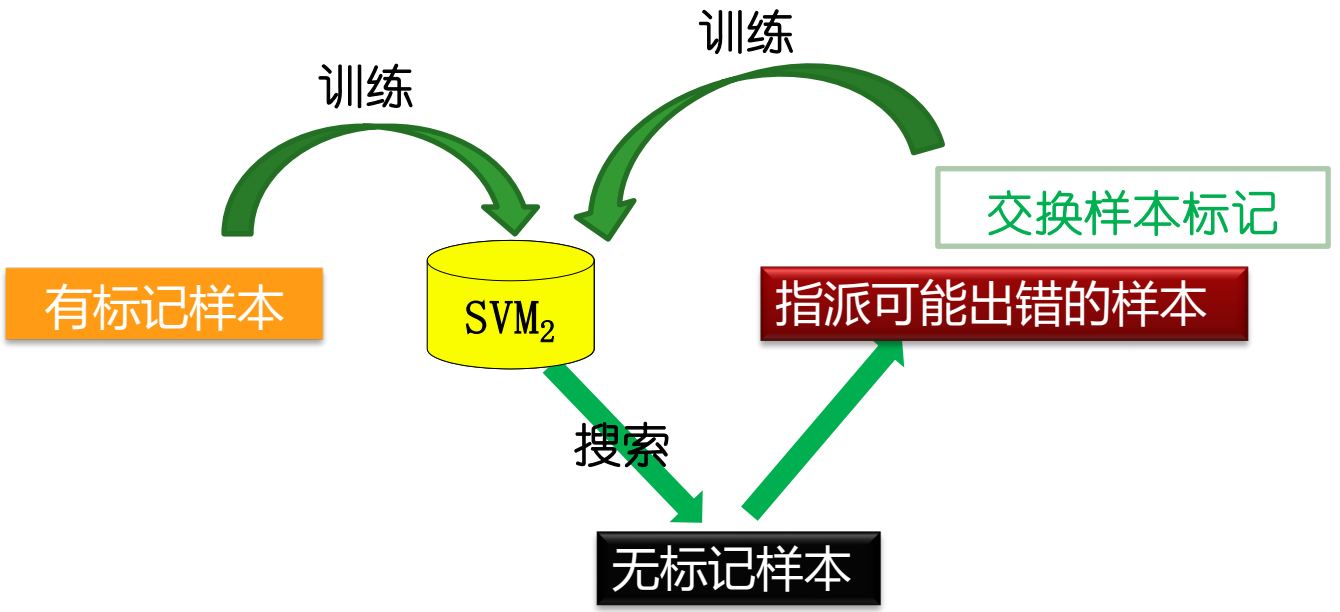
$$\begin{aligned} \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = l+1, \ldots, m, \\ & \xi_i \geq 0, \quad i = 1, \ldots, m, \end{aligned}$$

半监督SVM

- TSVM采用局部搜索来迭代地寻找近似解



半监督SVM



半监督SVM

未标记样本的伪标记不准确

输入: 有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$;
未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
折中参数 C_l, C_u.

过程:

用 D_l 训练一个 SVM_l;
用 SVM_l 对 D_u 中样本进行预测, 得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$;
初始化 $C_u \ll C_l$;
while $C_u < C_l$ **do**
  基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$;
  while $\exists\{i, j \mid (\hat{y}_i \hat{y}_j < 0) \wedge (\xi_i > 0) \wedge (\xi_j > 0) \wedge (\xi_i + \xi_j > 2)\}$ **do**
    $\hat{y}_i = -\hat{y}_i$;
    $\hat{y}_j = -\hat{y}_j$;
    基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 重新求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$
  end while
  $C_u = \min\{2C_u, C_l\}$
end while

输出: 未标记样本的预测结果: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

图 13.4　TSVM 算法

半监督SVM

- 显然, 搜寻标记指派可能出错的每一对未标记样本, 是一个涉及巨大计算开销的大规模优化问题

- 因此, 半监督SVM研究的一个重点是如何设计出高效的优化求解策略

- 例如基于图核(graph kernel)函数梯度下降的Laplacian SVM [Chapelle and Zien, 2005]、基于标记均值估计的meanS3VM [Li et al., 2009]等

大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

图半监督学习

- 给定一个数据集，我们可将其映射为一个图，数据集中每个样本对应于图中一个结点，若两个样本之间的相似度很高(或相关性很强)，则对应的结点之间存在一条边，边的强度(strength)正比于样本之间的相似度(或相关性)

- 基本假设：
  - 关联性越强的样本，具有相似标注的可能性越大

图半监督学习

- 图中的节点
  - 有标注样本+无标注样本
- 图中的边：基于样本特征计算相似度
  - K近邻图：与K近邻样本连边
  - ϵ近邻图：距离≤ ϵ的样本连边
  - 全连接图：$w = \exp\left(-\|x_i - x_j\|^2/\sigma^2\right)$
  - …

Mincut 算法

- 有标注损失+无标注损失

$$\min_{Y\in\{0,1\}^n} \sum_{i=1}^{l}(y_i - Y_{li})^2 + \sum_{ij} w_{ij}(y_i - y_j)^2$$

整数规划，但通常可以找到多项式规模的算法

标签传播算法

- 假定从图中学得一个实值函数 $f : V \to \mathbb{R}$

- 直观上看，相似的样本应具有相似的标记，于是可定义关于f的能量函数(energy function) :

$$E(f) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{W}_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2$$
$$= \boldsymbol{f}^T(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{f}$$

$$\mathrm{D}_{ii} = \sum_{j=1}^{m} W_{ij}$$

D − W 为拉普拉斯矩阵
(Laplacian matrix)

标签传播算法

- 采用分块矩阵表示方式：

$$E(f) = (\boldsymbol{f}_l^\top \ \boldsymbol{f}_u^\top) \left(\begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix}$$

$$= \boldsymbol{f}_l^\top (\mathbf{D}_{ll} - \mathbf{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^\top \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^\top (\mathbf{D}_{uu} - \mathbf{W}_{uu}) \boldsymbol{f}_u \ .$$

- 由 $\frac{\partial E(f)}{\partial \boldsymbol{f}_u} = \mathbf{0}$ 可得：

$$\boldsymbol{f}_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1} \mathbf{W}_{ul} \boldsymbol{f}_l$$

图半监督学习

- 图半监督学习方法在概念上相当清晰，且易于通过对所涉矩阵运算的分析来探索算法性质
- 但此类算法的缺陷也相当明显．首先是在存储开销高
- 另一方面，由于构图过程仅能考虑训练样本集，难以判知新样本在图中的位置，因此，在接收到新样本时，或是将其加入原数据集对图进行重构并重新进行标记传播，或是需引入额外的预测机制

大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

基于分歧的方法

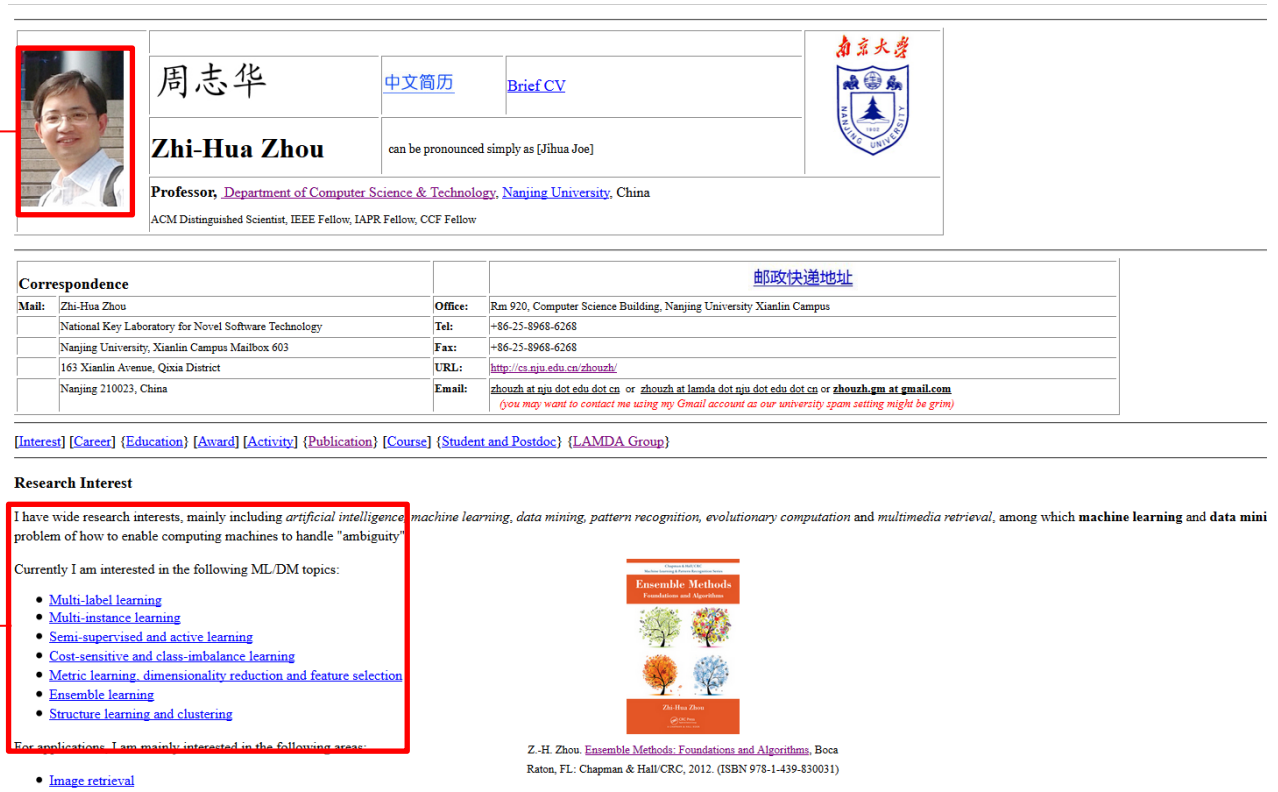
- 基于分歧的方法(disagreement-based methods)使用多学习器，而学习器之间的“分歧”(disagreement)对未标记数据的利用至关重要

- 协同训练(co-training)[Blum and Mitchell, 1998]是基于分歧的方法的重要代表，它最初是针对“多视图”(multi-view)数据设计的，因此也被看作“多视图学习”(multi-view learning)的代表

基于分歧的方法

图片视图

文字视图



周志华 中文简历 Brief CV

Zhi-Hua Zhou can be pronounced simply as [Jihua Joe]

Professor, Department of Computer Science & Technology, Nanjing University, China

ACM Distinguished Scientist, IEEE Fellow, IAPR Fellow, CCF Fellow

Correspondence			邮政快递地址
Mail:	Zhi-Hua Zhou	**Office:**	Rm 920, Computer Science Building, Nanjing University Xianlin Campus
	National Key Laboratory for Novel Software Technology	**Tel:**	+86-25-8968-6268
	Nanjing University, Xianlin Campus Mailbox 603	**Fax:**	+86-25-8968-6268
	163 Xianlin Avenue, Qixia District	**URL:**	http://cs.nju.edu.cn/zhouzh/
	Nanjing 210023, China	**Email:**	zhouzh at nju dot edu dot cn or zhouzh at lamda dot nju dot edu dot cn or **zhouzh.gm at gmail.com** *(you may want to contact me using my Gmail account as our university spam setting might be grim)*

[Interest] [Career] {Education} [Award] [Activity] {Publication} [Course] {Student and Postdoc} {LAMDA Group}

Research Interest

I have wide research interests, mainly including *artificial intelligence*, *machine learning*, *data mining*, *pattern recognition*, *evolutionary computation* and *multimedia retrieval*, among which **machine learning** and **data mini** problem of how to enable computing machines to handle "ambiguity".

Currently I am interested in the following ML/DM topics:

- Multi-label learning
- Multi-instance learning
- Semi-supervised and active learning
- Cost-sensitive and class-imbalance learning
- Metric learning, dimensionality reduction and feature selection
- Ensemble learning
- Structure learning and clustering

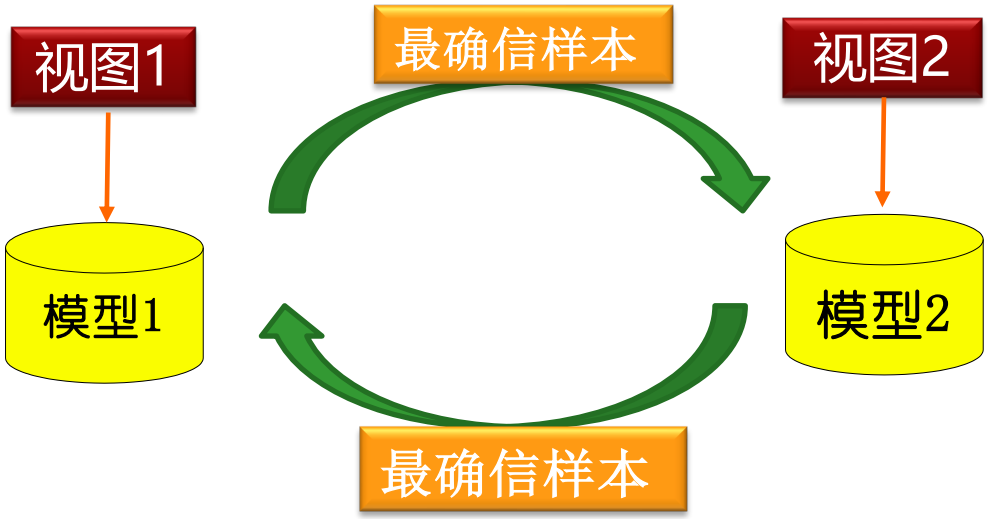
For applications, I am mainly interested in the following areas:

- Image retrieval

Z.-H. Zhou. Ensemble Methods: Foundations and Algorithms, Boca Raton, FL: Chapman & Hall/CRC, 2012. (ISBN 978-1-439-830031)

网页分类任务中的双视图

协同训练(co-training)



若两个视图“充分”(sufficient)且“条件独立”，可利用无标注样本通过协同训练将弱学习器性能提升到任意高

协同训练(co-training)

- 协同训练算法本身是为多视图数据而设计的，但此后出现了一些能在单视图数据上使用的变体算法

- 它们或是使用不同的学习算法[Goldman and Zhou,2000]、或使用不同的数据采样[Zhou and Li, 2005b]、甚至使用不同的参数设置[Zhou and Li, 2005a]来产生不同的学习器，也能有效地利用未标记数据来提升性能

- 后续理论研究发现，此类算法事实上无需数据拥有多视图，仅需弱学习器之间具有显著的分歧(或差异)，即可通过相互提供伪标记样本的方式来提高泛化性能[周志华, 2013]

协同训练(co-training)

- 基于分歧的方法只需采用合适的基学习器，就能较少受到模型假设、损失函数非凸性和数据规模问题的影响，学习方法简单有效、理论基础相对坚实、适用范围较为广泛

- 为了使用此类方法，需能生成具有显著分歧、性能尚可的多个学习器，但当有标记样本很少、尤其是数据不具有多视图时，要做到这一点并不容易

大纲

- 无标注样本
- 自训练(self-training)
- 生成式方法(generative methods)
- 半监督支持向量机(semi-supervised SVM,S3VM)
- 图半监督学习(graph-based SSL)
- 协同训练(co-training)
- 深度半监督学习(deep SSL)

深度半监督学习

- 深度学习，需要用到大量的有标记数据，即使在大数据时代，干净能用的有标记数据也是不多的

- 近年来，深度学习与半监督学习思想相结合，产生的半监督深度学习已成为深度学习领域热门的新方向

- 包括MIT、Stanford、Google Brain、Facebook等学术界和工业界，在半监督深度学习领域做了大量的工作

两个常用方法

- 一致性正则 (Consistency Regularization)

- 熵最小化(Entropy Minimization)

两个常用方法

- 一致性正则：
  - 如果对一个无标注样本添加扰动，预测应保持一致

- 具体来说，给定一个未标记的数据样本及其扰动的形式 ，目标是最小化两个输出之间的距离：

$$d(f_\theta(x) \ , \quad f(\hat{x}))$$

一致性正则

流行的距离度量 $d(f_\theta(x)\ ,\ \ f(\hat{x}))$ 均方误差(Mean-Squared Error, MSE)，Kullback-Leiber散度 (KL Divergence) 和 Jensen-Shannon 散度(JS Divergence)

$$d_{\mathrm{MSE}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{C}\sum_{k=1}^{C}(f_\theta(x)_k - f_\theta(\hat{x})_k)^2$$

$$d_{\mathrm{KL}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{C}\sum_{k=1}^{C} f_\theta(x)_k \log \frac{f_\theta(x)_k}{f_\theta(\hat{x})_k}$$

$$d_{\mathrm{JS}}(f_\theta(x), f_\theta(\hat{x})) = \frac{1}{2}d_{\mathrm{KL}}(f_\theta(x), m) + \frac{1}{2}d_{\mathrm{KL}}(f_\theta(\hat{x}), m)$$

一致性正则

基于一致性正则(Consistency Regularization)的思想，衍生出一批深度半监督学习算法，如：

Π-Model、Temporal Ensembling、Mean Teacher、VAT、UDA等

具体到每一种算法，核心思想是没有变化的，即最小化未标记数据与其扰动两者之间预测值的距离，主要区别在于：

1）进行数据扰动的方式不同

2）距离的计算方式不同

Π-Model

TEMPORAL ENSEMBLING FOR SEMI-SUPERVISED LEARNING

Samuli Laine
NVIDIA
slaine@nvidia.com

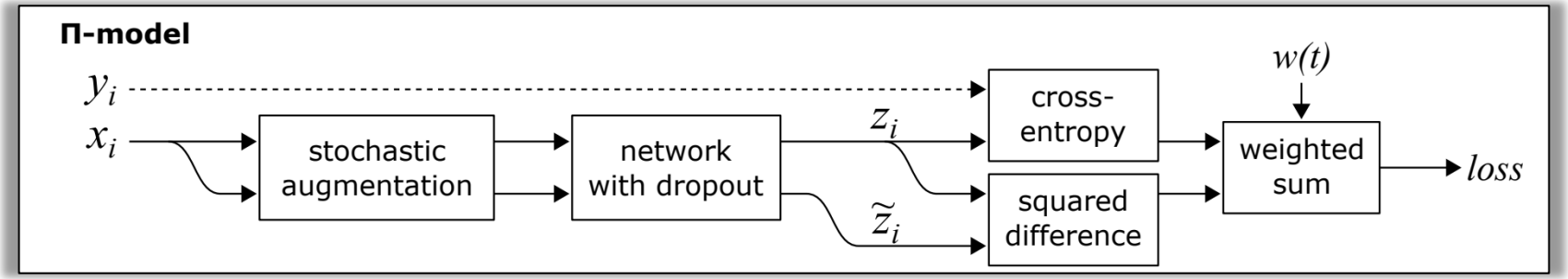
Timo Aila
NVIDIA
taila@nvidia.com

论文链接：
https://openreview.net/forum?id=BJ6oOfqge&noteId=BJ6oOfqge

代码链接：
https://github.com/smlaine2/tempens

Π-Model

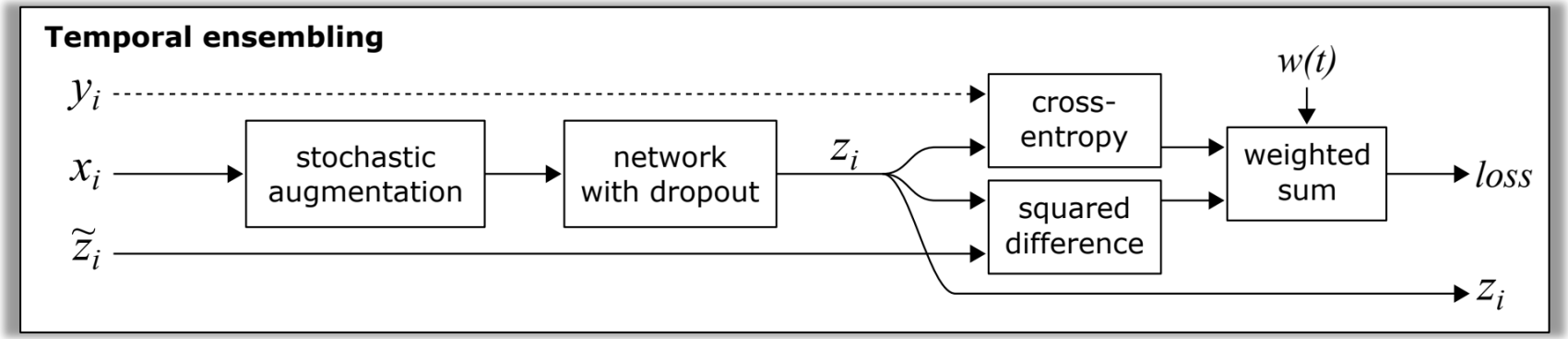
- 对给定的样本x，采用不同的数据增广，得到两次预测结果，目标是最小化两次预测结果之间的不一致性



- 数据扰动方式：随机反转、平移、剪切等策略
- 无监督损失：两次前向运算结果的均方误差(MSE)

Temporal Ensembling

- Π-Model：最小化两次随机增广预测值的均方误差
- Temporal Ensembling：最小化当前模型预测结果与历史模型预测结果的平均值之间的均方误差



Temporal Ensembling

- 数据扰动方式：当前轮样本的预测值和该样本在历史轮数上的预测值
- 无监督损失：两次预测值的均方误差(MSE)

✓ 用空间来换取时间，总的前向计算次数减少了一半

✓ 通过历史预测做平均，有利于减小单次预测中的噪声

Mean Teacher

Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results

Antti Tarvainen
The Curious AI Company
and Aalto University
antti.tarvainen@aalto.fi

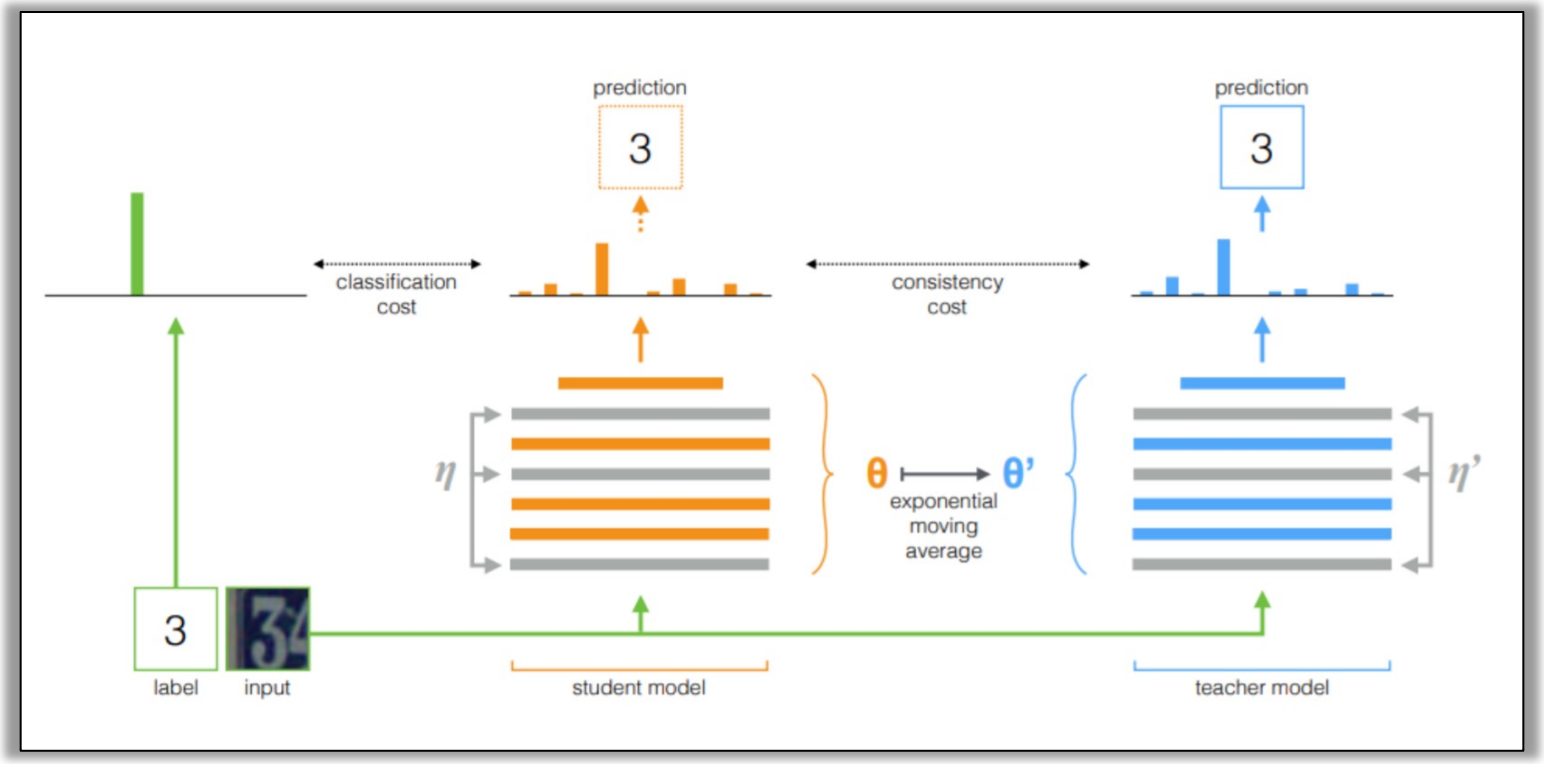
Harri Valpola
The Curious AI Company
harri@cai.fi

论文链接：
https://arxiv.org/abs/1703.01780

代码链接：
https://github.com/CuriousAI/mean-teacher

Mean Teacher

- Temporal Ensembling： 保存模型对样本的预测值
- Mean Teacher： 直接保存模型的权重



Mean Teacher

- 数据扰动方式：当前模型对该样本的预测值和历史轮数上模型的集成对该样本的预测值
- 无监督损失：两次预测值的均方误差(MSE)
- 保存模型的参数比保存历史预测值更一般化，性能也会更好，但是会带来更大的存储消耗

Virtual Adversarial Training (VAT)

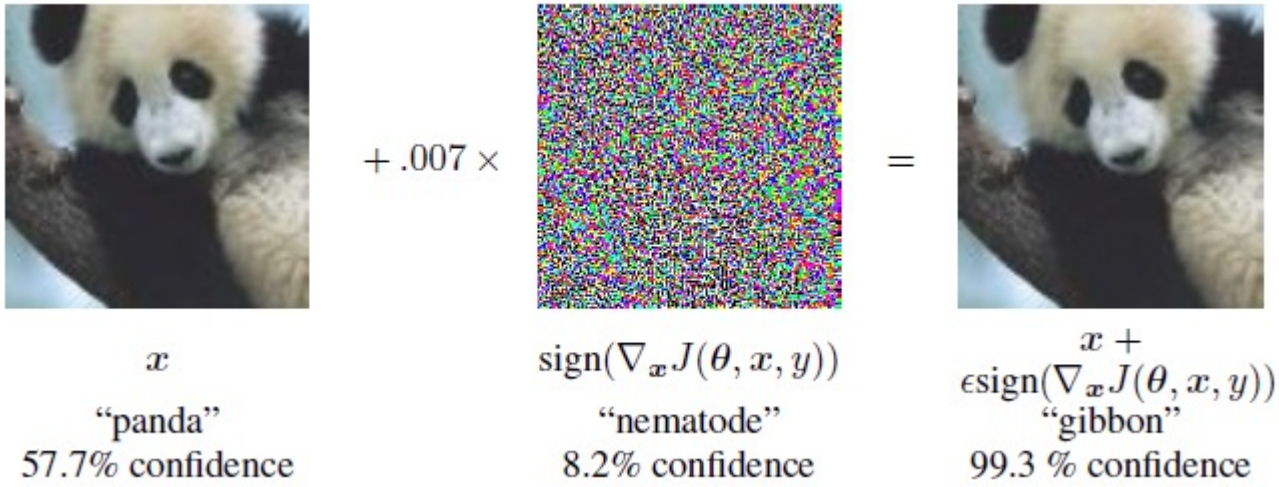
Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning

Takeru Miyato[*,†,‡], Shin-ichi Maeda[*,†], Masanori Koyama[§,†] and Shin Ishii[†,‡]

论文链接：

https://arxiv.org/abs/1704.03976

对抗样本



原有的模型以57.7%的置信度判定图片为熊猫，但添加微小的扰动后，模型以99.3%的置信度认为扰动后的图片是长臂猿

对抗样本可以让训练优秀的分类网络进行错误的分类，然而人类去看对抗样本的话和真实的样本几乎无异

Virtual Adversarial Training (VAT)

- 最小化原始样本预测值和对抗样本预测值之间的KL Divergence

$$\sum_x D[p_\theta(\mathbf{y}|\mathbf{x}), p_\theta(\mathbf{y}|\mathbf{x} + \mathbf{r}_{adv})]$$

- 相比以往的方法，提升了半监督深度学习模型在对抗扰动下的鲁棒性

Unsupervised Data Augmentation

Unsupervised Data Augmentation for Consistency Training

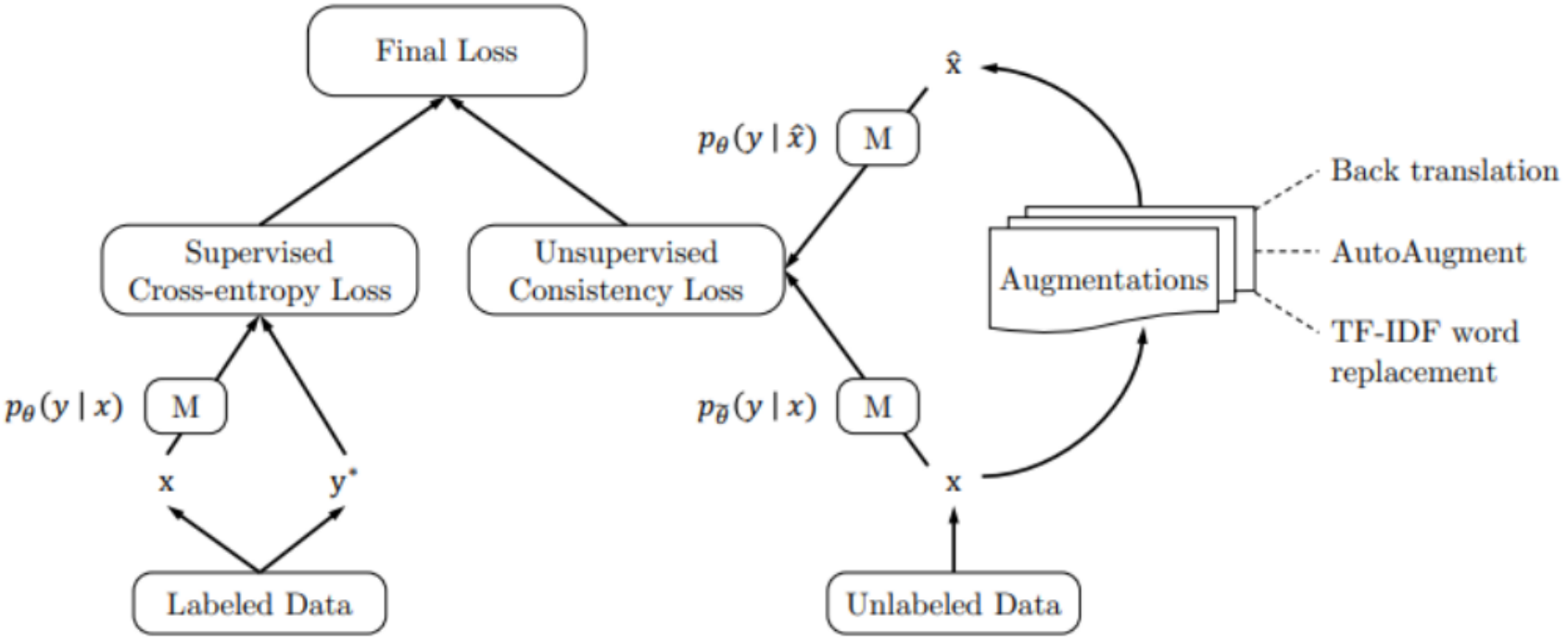
Qizhe Xie[1,2], **Zihang Dai**[1,2], **Eduard Hovy**[2], **Minh-Thang Luong**[1], **Quoc V. Le**[1]
[1] Google Brain, [2] Carnegie Mellon University
{qizhex, dzihang, hovy}@cs.cmu.edu, {thangluong, qvl}@google.com

论文链接：
https://arxiv.org/pdf/1904.12848v2.pdf

代码链接：
https://github.com/google-research/uda

Unsupervised Data Augmentation

- 该工作提出，对数据进行增广的方式不应该是一成不变的，需要对数据采取更多样化，更领域相关的数据增强方式



Unsupervised Data Augmentation

- 最小化未标记数据和增强未标记数据上预测分布之间的 KL Divergence

$$\min_{\theta} \mathcal{J}_{\text{UDA}}(\theta) = \mathop{\mathbb{E}}_{x \in U} \mathop{\mathbb{E}}_{\hat{x} \sim q(\hat{x}|x)} \left[\mathcal{D}_{\text{KL}} \left(p_{\tilde{\theta}}(y \mid x) \,\|\, p_{\theta}(y \mid \hat{x}) \right) \right]$$

- UDA证明了针对性的数据增强效果明显优于无针对性的数据增强

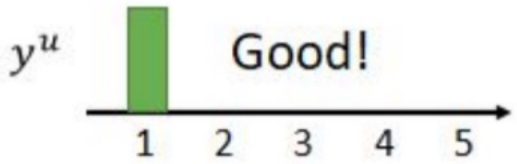
基于一致性正则的方法小结

- 一致性正则(Consistency Regularization)这类方法的主要思想：
  - 对于无标记样本，添加扰动后模型的预测值也应该尽可能保持不变
- 各方法的主要区别在于如何找到更适合的数据增广
- 没有哪种数据增广是万能的，数据增广方法不应该是一成不变的，要结合领域知识，不同的任务采用不同的增广方法！

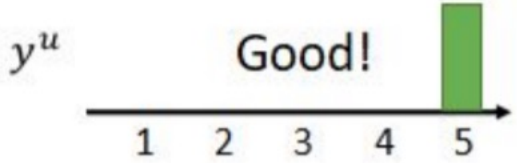
熵最小化(entropy minimization)

- 熵最小化：鼓励模型输出的预测值置信度尽可能高

Distribution



$E(y^u) = 0$



$E(y^u) = 0$

$$E(y^u) = -ln\left(\frac{1}{5}\right) = ln5$$

Entropy of y^u :
Evaluate how concentrate the distribution y^u is

$$E(y^u) = -\sum_{m=1}^{5} y_m^u ln(y_m^u)$$

As small as possible

$$L = \sum_{x^r} C(y^r, \hat{y}^r)$$

labelled data

$$+\lambda \sum_{x^u} E(y^u)$$

unlabeled data

熵最小化(entropy minimization)

- 如何分配伪标签：最受欢迎的两种是锐化 (sharpening) 方法和 Argmax 方法

- 前者在保持预测值分布的同时使分布有些极端
- 后者仅使用对预测具有最高置信度的预测标签进行标记

- 我们还可以对无标签数据进行过滤，如果预测结果大于预定阈值τ，再将其添加训练中

Holistic Methods

- 一致性正则和熵最小化各有优劣，
- 能不能同时考虑两种正则化方法呢？

- Google Brain提出MixMatch、FixMatch算法，称为Holistic Method
- 即试图在一个框架中整合当前的 SSL 的主要方法，从而获得更好的性能

MixMatch

MixMatch: A Holistic Approach to Semi-Supervised Learning

David Berthelot
Google Research
dberth@google.com

Nicholas Carlini
Google Research
ncarlini@google.com

Ian Goodfellow
Work done at Google
ian-academic@mailfence.com

Avital Oliver
Google Research
avitalo@google.com

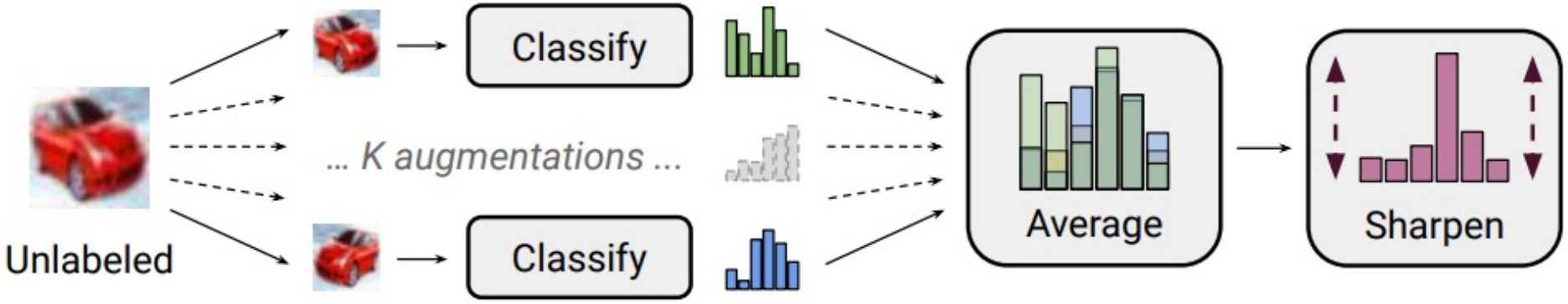
Nicolas Papernot
Google Research
papernot@google.com

Colin Raffel
Google Research
craffel@google.com

论文链接：
https://arxiv.org/pdf/1905.02249.pdf

代码链接：
https://github.com/google-research/mixmatch

MixMatch



- 对样本做k次增广，然后将预测值平均，再进行锐化操作

- 无监督损失：上述操作得到的预测值和模型直接预测值之间的均方误差

FixMatch

FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence

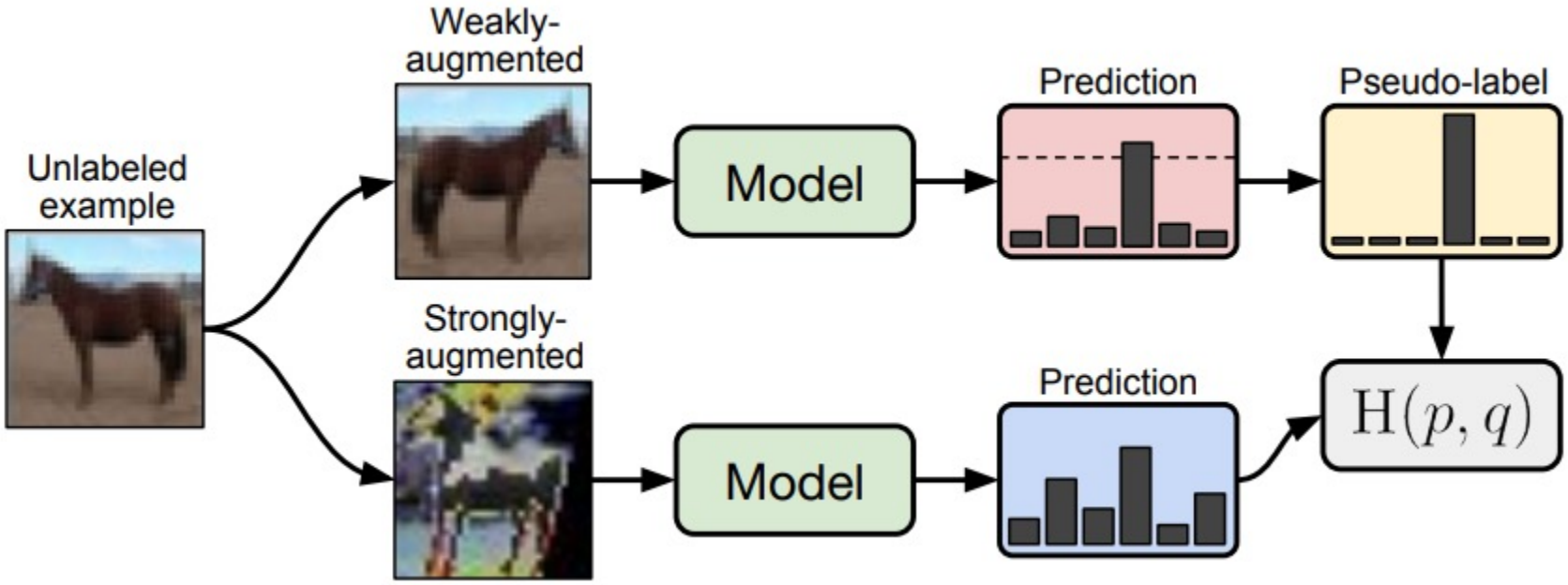
Kihyuk Sohn* David Berthelot* Chun-Liang Li Zizhao Zhang Nicholas Carlini
Ekin D. Cubuk Alex Kurakin Han Zhang Colin Raffel
Google Research
{kihyuks,dberth,chunliang,zizhaoz,ncarlini,
cubuk,kurakin,zhanghan,craffel}@google.com

论文链接：
https://arxiv.org/ftp/arxiv/papers/2001/2001.07685.pdf

代码链接：
https://github.com/google-research/fixmatch

FixMatch

- MixMatch 对预测值进行锐化操作
- FixMatch 利用Argmax的方式得到伪标记，并且只有预测结果大于预定阈值τ，才将该样本加入训练



FixMatch

- 两种增强
  - 弱增强：用标准的翻转和平移策略
  - 强增强：输出严重失真的输入图像

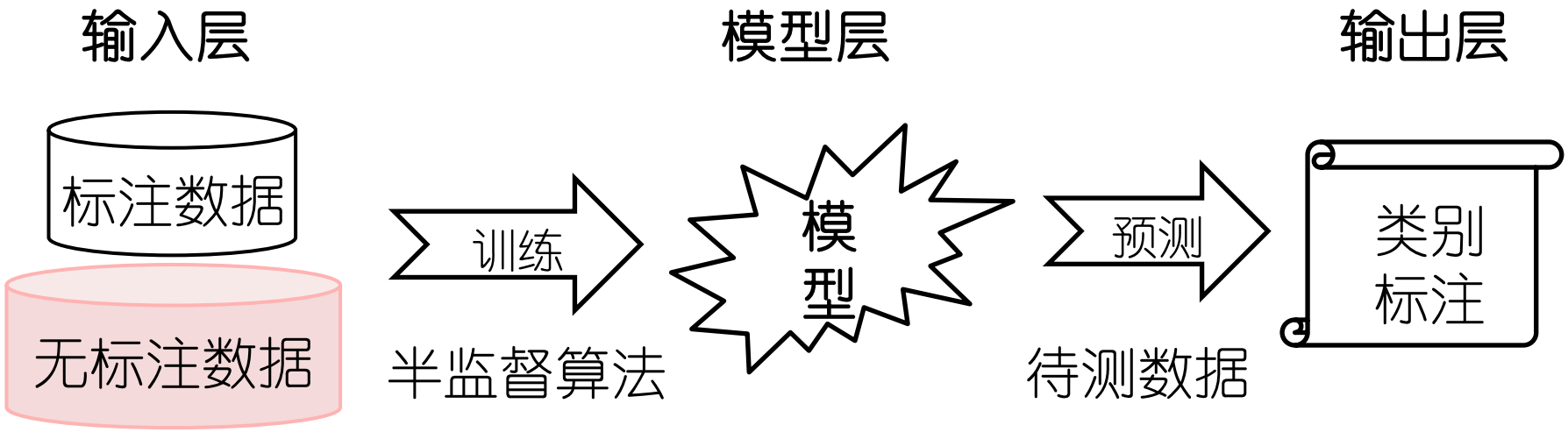
- 无监督损失：
  - 利用弱增强的样本得到伪标记后，计算模型预测值和伪标记之间的交叉熵(和监督损失一致)

半监督深度学习小结

- 一致性正则化方法通过鼓励无标签数据扰动前后的预测相同，增加了模型对数据变化的鲁棒性，减缓了标记数据不足时容易过拟合的问题
- 熵最小化的方法主要是通过对未标记数据制作满足熵最小化的伪标签然后加入训练，以得到更好的决策边界
- 而众多方法中，混合方法表现出了良好的性能，是近来的研究热点

小结

当标注数据不足时，利用大量无标注数据辅助提升性能



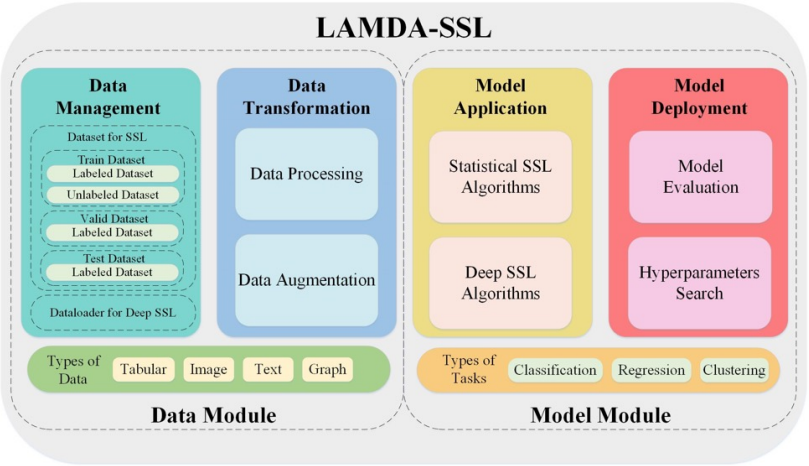
半监督学习

- 统计学习时期：生成式半监督、半监督SVM
  图半监督、基于分歧的半监督
- 深度学习时期：熵最小化方法、一致性正则方法
  混合式方法

SSL工具包

半监督学习工具包：LAMDA-SSL

- ✓ 支持30余种半监督学习算法，4种数据类型，16种评价指标
- ✓ Github: https://github.com/YGZWQZD/LAMDA-SSL




```python
from LAMDA_SSL.Dataset.Vision.CIFAR10 import CIFAR10
from LAMDA_SSL.Algorithm.Classification.FixMatch import FixMatch
from LAMDA_SSL.Evaluation.Classifier.Accuracy import Accuracy
# Initialize CIFAR10 dataset
dataset=CIFAR10(root='..\Download\cifar-10-python',labeled_size=4000)
labeled_X, labeled_y=dataset.labeled_X,dataset.labeled_Y
unlabeled_X=dataset.unlabeled_X
test_X, test_y=dataset.test_X, dataset.test_y
# Initialize FixMatch algorithm
model=FixMatch(threshold=0.95,lambda_u=1.0,T=0.5,mu=7,
               epoch=1,num_it_epoch=2**20,device='cuda:0')
# Call the fit() method to Train the model
model.fit(X=labeled_X,y=labeled_y,unlabeled_X=unlabeled_X)
# Call the predict() method to predict the labels of new samples
y_pred=model.predict(test_X)
# Evaluate the model' s performance.
performance=Accuracy().scoring(test_y,y_pred)
```