半监督学习

卿来云

复习：降维

- 特征选择/降维的目的
  - 选择出重要的特征可以缓解维数灾难问题
  - 去除不相关特征可以降低学习任务的难度
- 常用的特征选择方法
  - 手工定义规则
  - 过滤法
  - 包裹法
  - 嵌入式

复习：降维

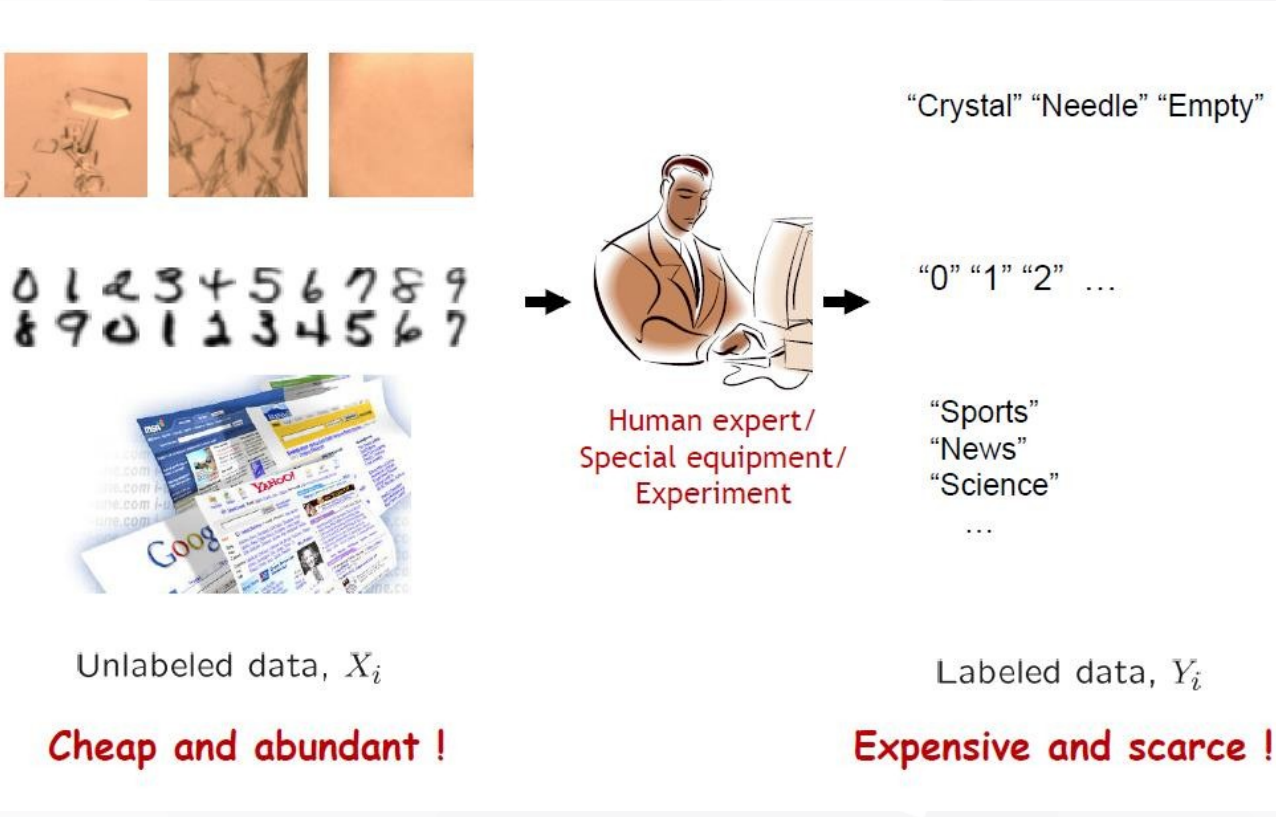
- 常用的线性降维技术
  - 保距：MDS（当距离度量为欧式距离时，MDS等价于PCA ）
  - 重构：PCA
  - 监督：LDA
- 常用的非线性降维技术
  - 重构：核化PCA、深度自编码器
  - 保距：ISOMAP、拉普拉斯映射、T-NSE

大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

监督学习

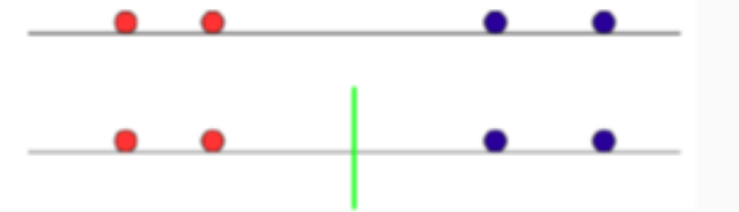
- 监督学习模型需要标注数据
- 学习一个可靠的模型需要大量标注数据



不是缺数据，而是缺有标签的数据

无标注数据能有帮助么?

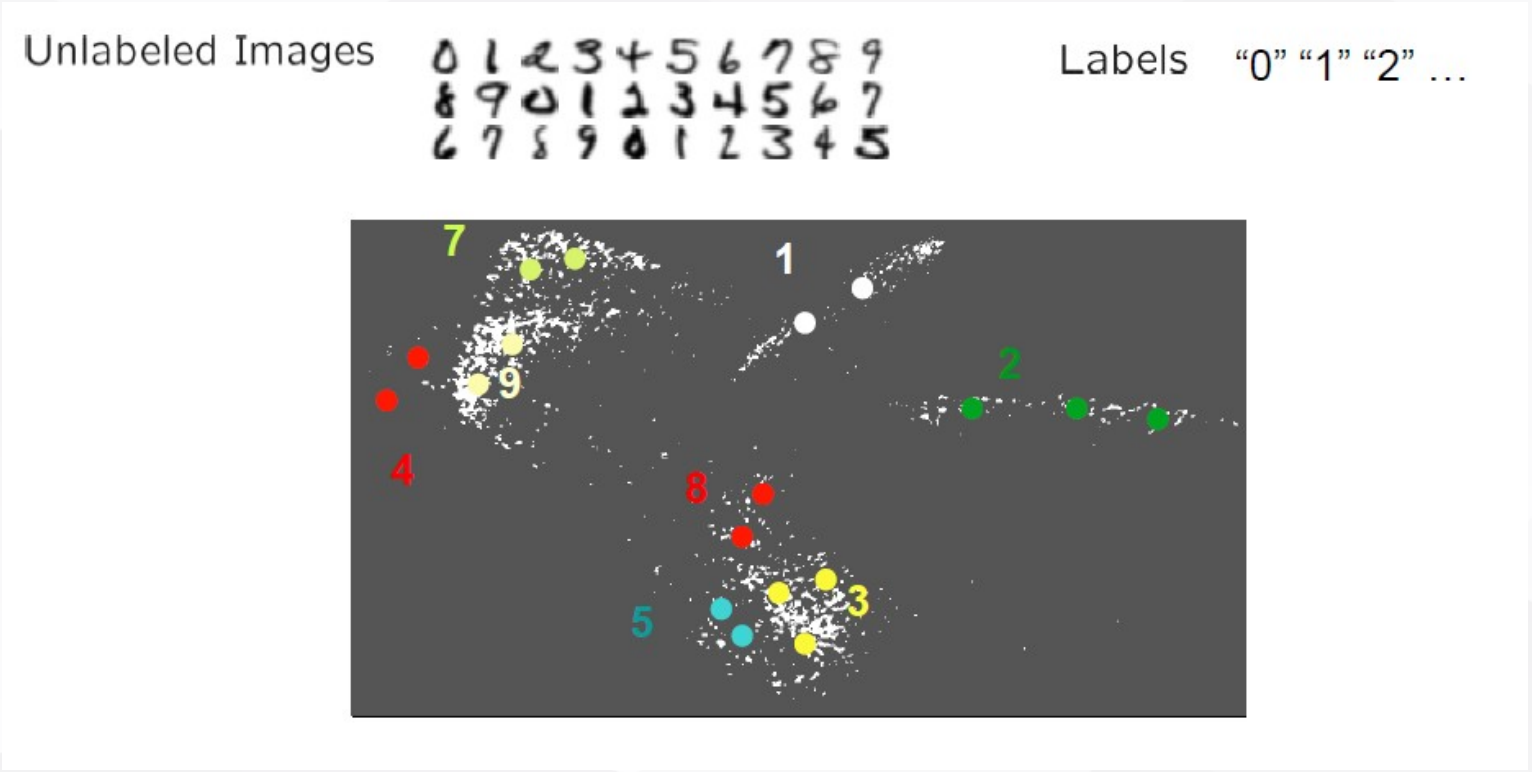
- 红色球: +1, 深蓝:-1



- 让我们包含额外的无标注数据 (浅蓝色的点)

- 假设: 同一个类别的样本内在服从一致的分布
- 无标注数据能够给出更有意义的分类边界

无标注数据能有帮助么?



- 假设: “相似” 的数据点有“相似”的标签

半监督学习

- 通用想法：同时利用有标注数据和无标注数据学习
- 半监督分类/回归
  - **给定:** 标注数据 $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots (\mathbf{x}_L, y_L)\}$，无标注数据 $\mathcal{D}_U = \{\mathbf{x}_{L+1}, \mathbf{x}_{L+2}, \dots \mathbf{x}_{L+U}\}$（ $U \gg L$ ）
  - 目标: 学习一个分类器 f 比只用标注数据学的更好
- 半监督聚类/降维
  - **给定:** 标注数据 $\{\mathbf{x}_i\}_{i=1}^N$, 但另外给出对数据的一些限制
    - 聚类: 两个点必须在一个簇, 或两个点一定不能在一个簇;
    - 降维: 两个点降维后必须接近

归纳学习vs 直推学习

- 归纳学习（Inductive learning）
  - 给定训练数据$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots (\mathbf{x}_L, y_L)\}$，无标注数据 $\mathcal{D}_U = \{\mathbf{x}_{L+1}, \mathbf{x}_{L+2}, \dots \mathbf{x}_{L+U}\}$（$U \gg L$）
  - 学习一个函数f用于预测新来的测试数据的标签
  - 学习一个函数能够被应用到测试数据上
- 直推学习（Transductive learning）
  - 给定训练数据 $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots (\mathbf{x}_L, y_L)\}$，无标注数据 $\mathcal{D}_U = \{\mathbf{x}_{L+1}, \mathbf{x}_{L+2}, \dots \mathbf{x}_{L+U}\}$
  - 可以没有显示的学习函数, 我们所关心的是在 $\mathcal{D}_U$ 上的预测
  - $\mathcal{D}_U$ 是测试数据集合并且在训练时可以使用

为什么叫半监督学习？

监督学习（分类, 回归） $\{(\mathbf{x}_{1:N}, y_{1:N})\}$

↕

半监督学习 分类/回归 $\{(\mathbf{x}_{1:L}, y_{1:L}), \mathbf{x}_{L+1:N}, \mathbf{x}_{test}\}$

直推分类/回归 $\{\{(\mathbf{x}_{1:L}, y_{1:L}), \mathbf{x}_{L+1:N}\}$

↕

半监督聚类 $\{\mathbf{x}_{1:N}, must-, cannot-links\}$

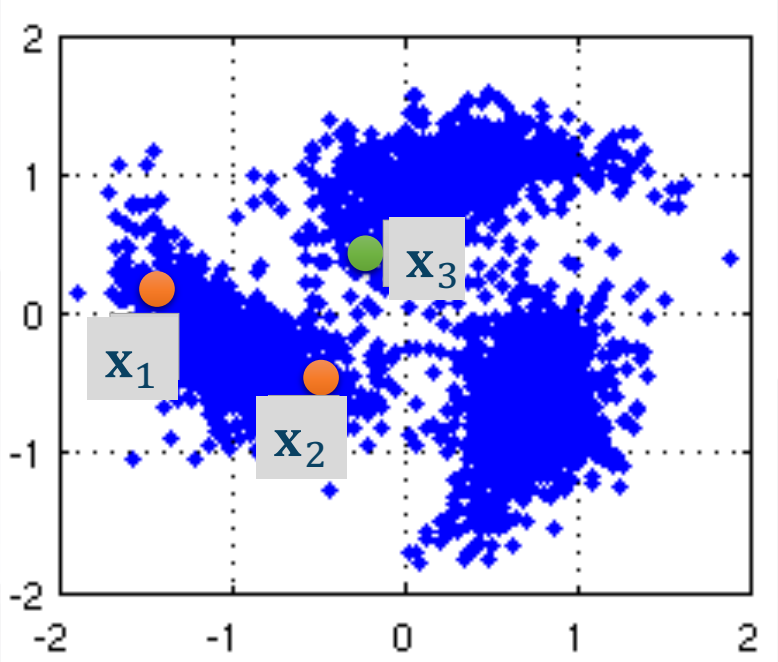
↕

无监督学习（聚类） $\{\mathbf{x}_{1:N}\}$

平滑假设（smoothness assumption）

- 半监督学习能有效，必须满足一些假设
- 半监督平滑假设：
  - 如果高密度区域中两个点$\mathbf{x}_1, \mathbf{x}_2$距离较近，那么对应的输出$y_1, y_2$也应该接近
    - $\mathbf{x}_1, \mathbf{x}_2$之间有一条高密度路径

$\mathbf{x}_1, \mathbf{x}_2$标签相同
$\mathbf{x}_1, \mathbf{x}_3$标签不同
虽然$\mathbf{x}_1, \mathbf{x}_2$之间的欧式距离与
$\mathbf{x}_1, \mathbf{x}_3$之间的欧式距离差不多



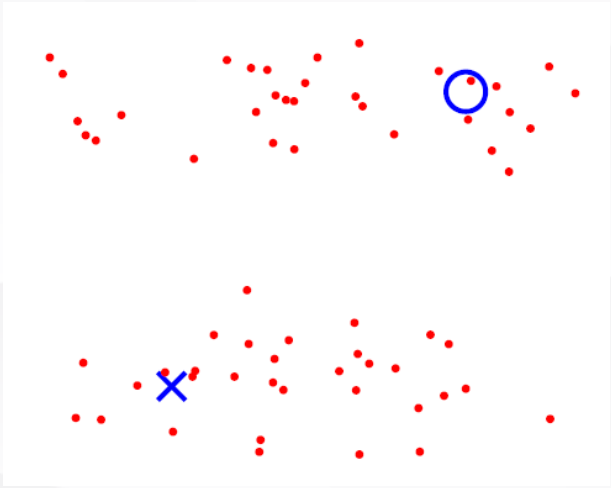
近朱者赤，近墨者黑

"You are known by the company you keep"

聚类假设（cluster assumption）

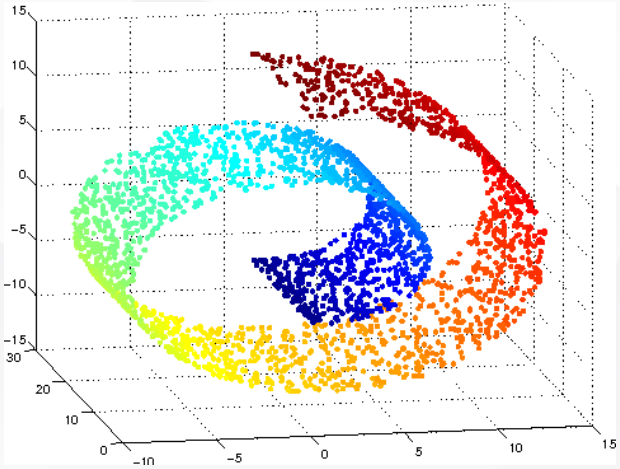
- 聚类假设
  - 如果点在同一个簇，那么它们很有可能属于同一个类

- 聚类假设的等价公式：
  - 低密度分隔：决策边界应该在低密度区域



流形假设（manifold assumption）

- 流形假设
  - 高维数据大致会分布在一个低维的流形上
  - 流形上邻近的样本拥有相似的输出
  - 邻近的程度常用“相似”程度来刻画

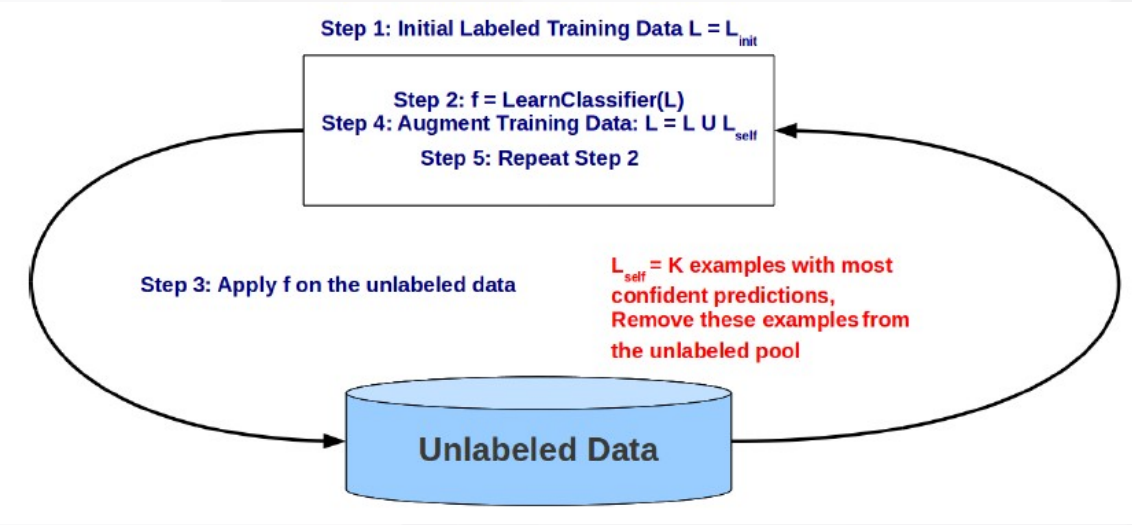


大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

自学习算法

- 假设
  - 输出的高度置信的预测是正确的
- 自学习算法
  - 从 $(\mathbf{X}_L, \mathbf{y}_L)$ 学习f
  - 对 $\mathbf{x} \in \mathcal{D}_U$，计算预测结果$f(\mathbf{x})$
  - 把 $(\mathbf{x}, f(\mathbf{x}))$ 加入到标注数据
  - 重复上述过程



如何从无标签数据中挑选数据加入到带标签数据集合仍是一个开放问题。

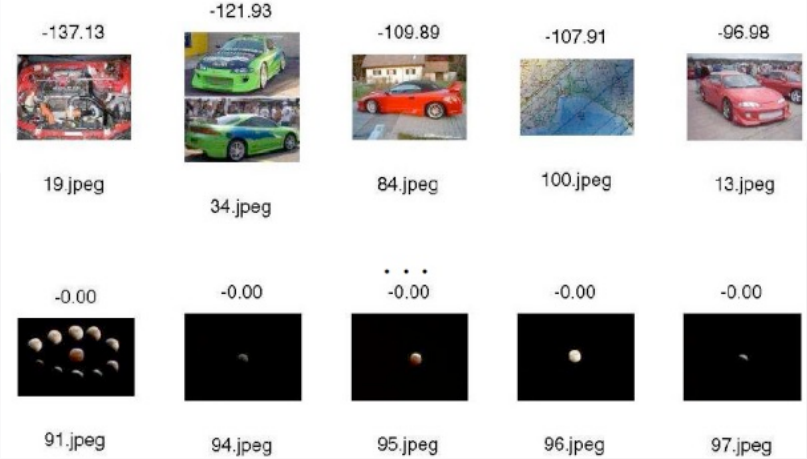
自学习的变体

- 加入一些置信度最高的$(\mathbf{x}, f(\mathbf{x}))$到标注数据集
- 把所有$(\mathbf{x}, f(\mathbf{x}))$加到标注数据
- 把所有$(\mathbf{x}, f(\mathbf{x}))$加到标注数据，为每条数据按置信度赋予权重

例：自动驾驶中的图像分类

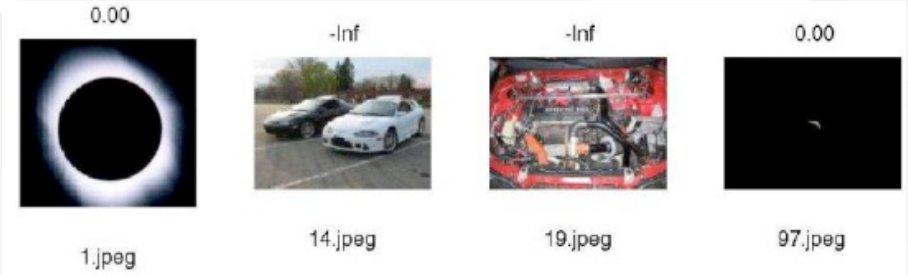
- 在两个初始图像上训练朴素贝叶斯分类器

- 对无标记的数据分类, 根据置信度 $\log p(y = astronomy|\mathbf{x})$ 排序

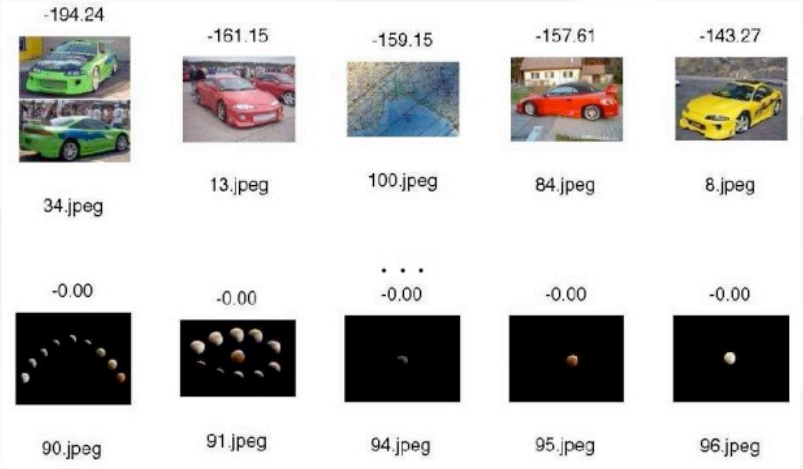


自动驾驶的例子: 图像分类

- 将最置信的图像及其预测标签加入到标注数据

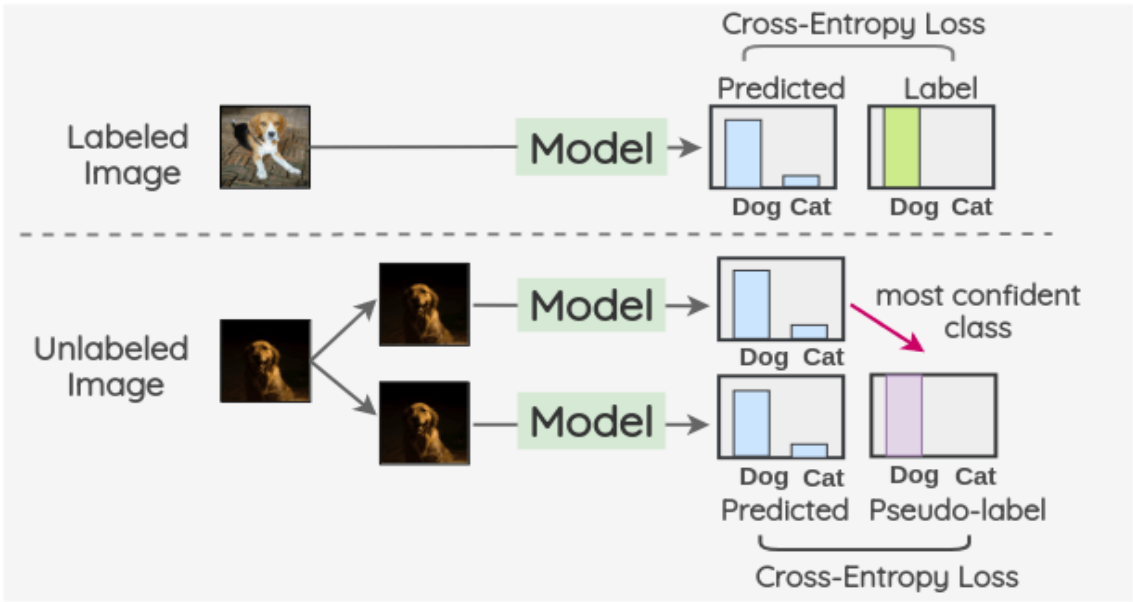


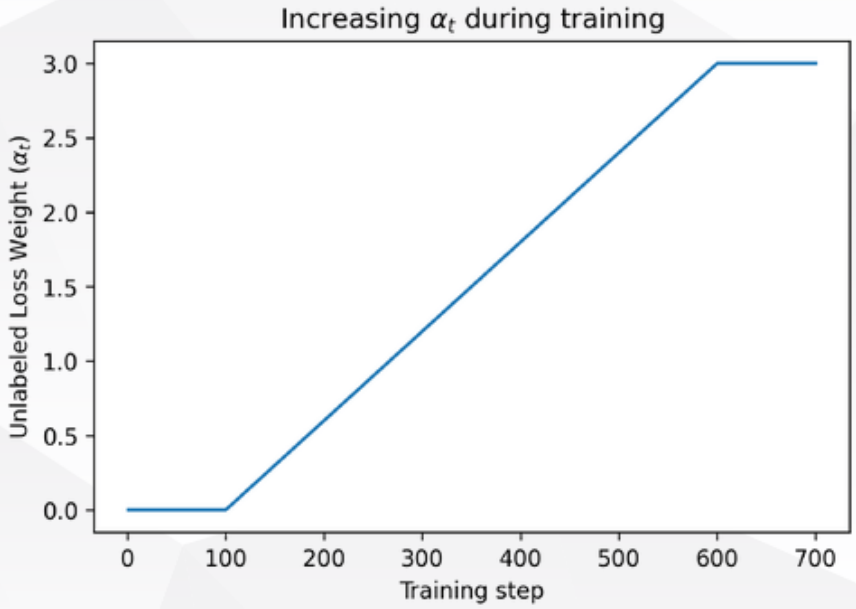
- 重新训练分类器，重复上述过程



例：伪标签（Pseudo-label）

■损失函数：$\mathcal{L} = \mathcal{L}_{labbel} + \alpha_t \times \mathcal{L}_{unlabbel}$





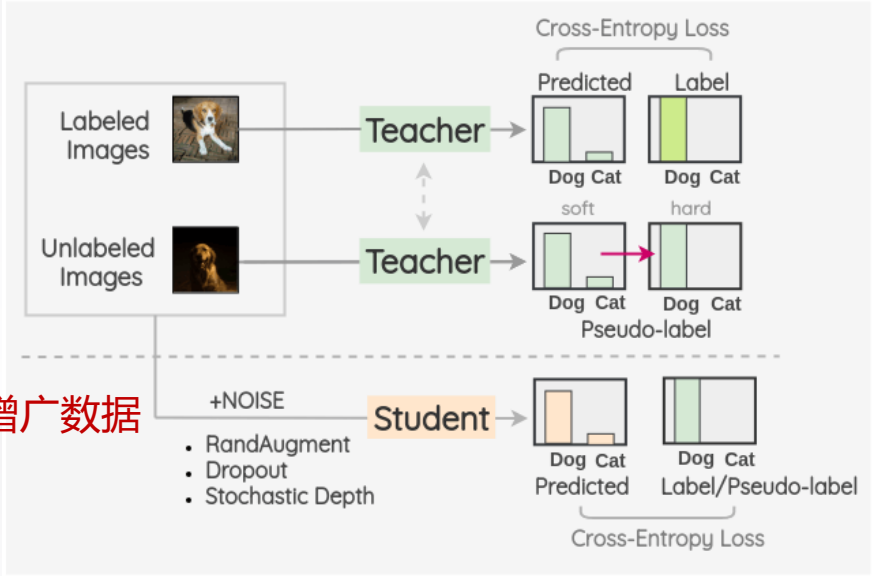
对无标签数据，计算模型预测结果与伪标签之间的交叉熵

•Dong-Hyun Lee, “Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”，ICML203

例：Noisy Student

■训练两个模型：

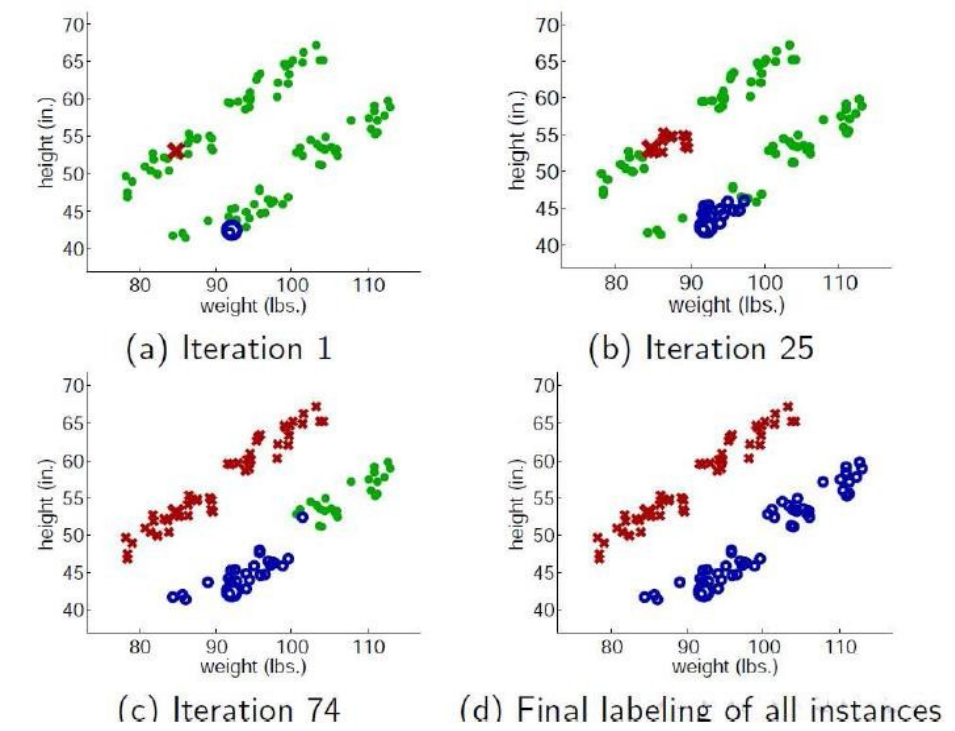
- 教师模型：有标签数据上训练，训练好的模型对无标签数据产生伪标签
- 学生模型：标签数据+伪标签数据+增广数据



•Qizhe Xie et al., "Self-training with Noisy Student improves ImageNet classification", CVPR2020

自我训练的优点

- 最简单的半监督学习方法，效果不错
- 这是一种wrapper方法，可以应用到已有的（复杂）分类器上

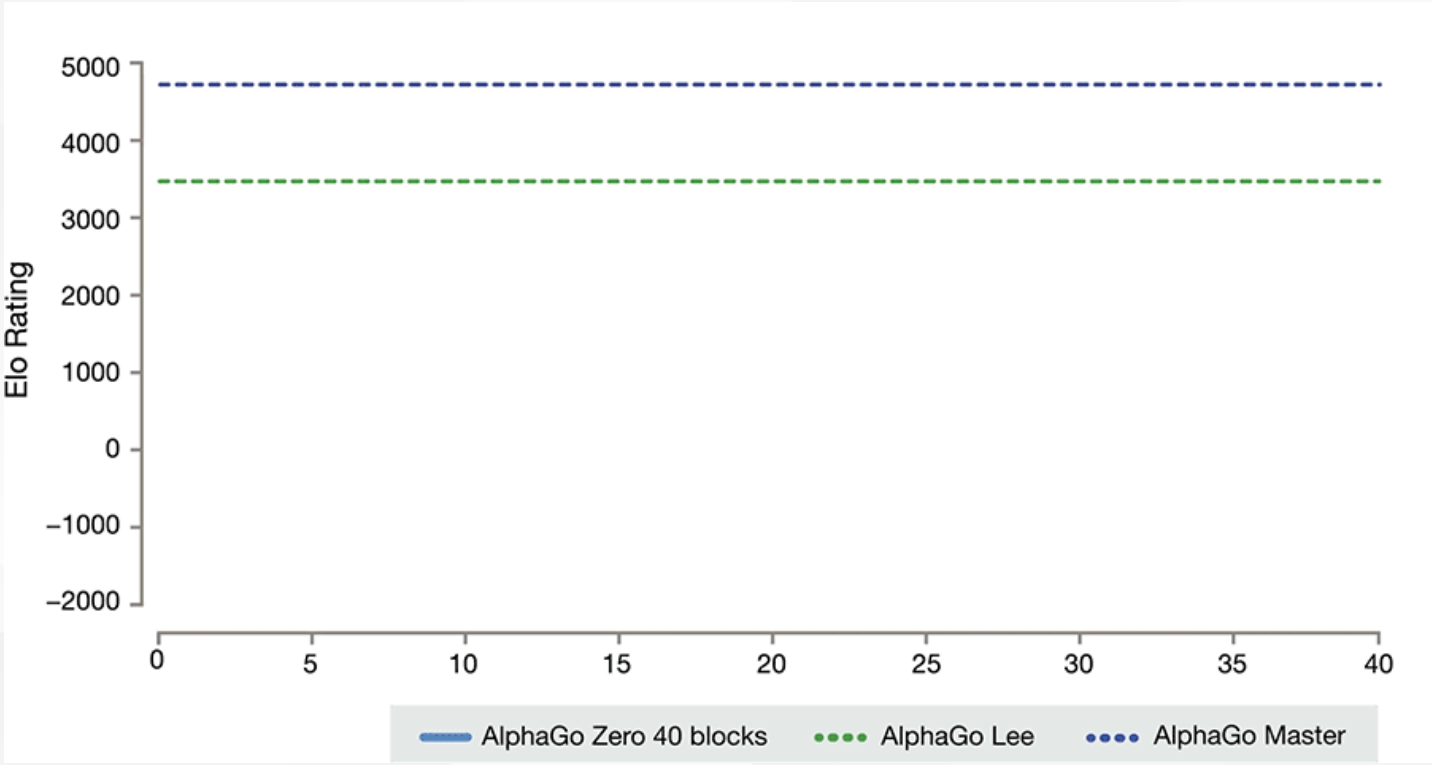


一个好的例子 基学习器: KNN

自我训练的缺点

- 早期的错误会强化
  - 启发式的缓解方案：如果数据的置信分数低于某个阈值再把它的标签去掉
- 在收敛性方面没有保障
  - 但是也有特例，自我训练等价于EM算法
  - 有部分存在封闭解的特殊情形(如线性函数)

知名的自学习例子（AlphaGo Zero）



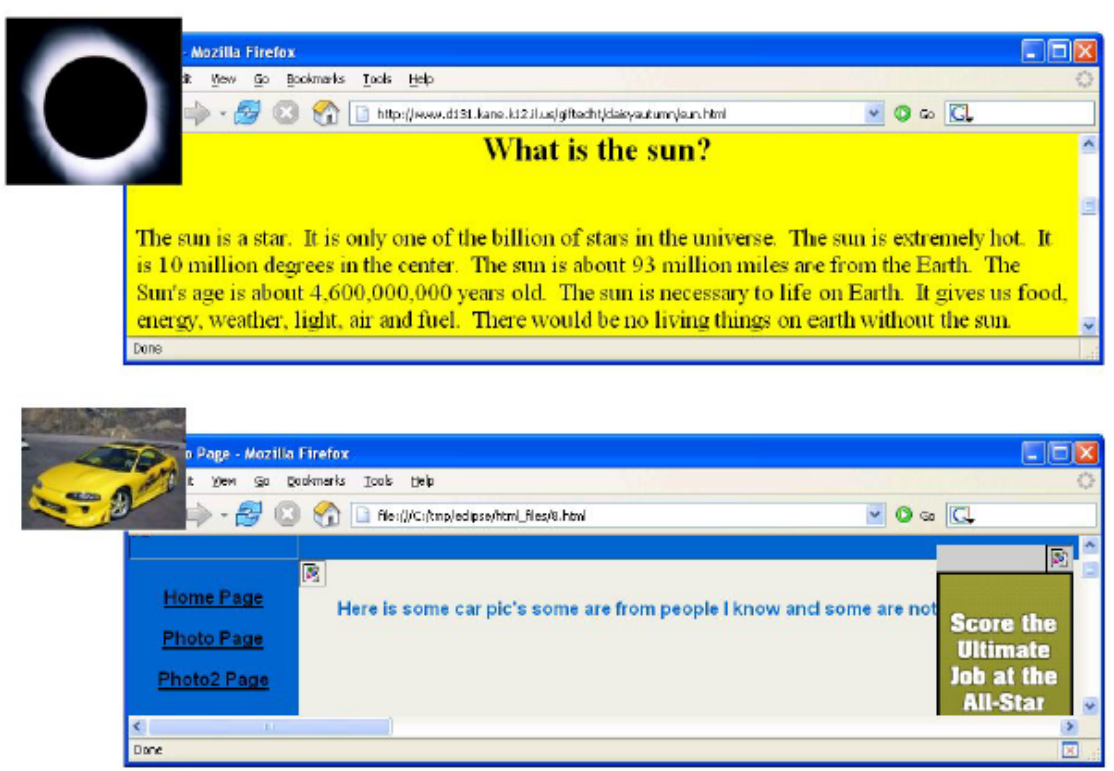
https://www.discovermagazine.com/technology/the-ai-that-dominated-humans-in-go-is-already-obsolete

大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

协同训练（co-training）

- 一个对象的两个视角: 图像和HTML文本

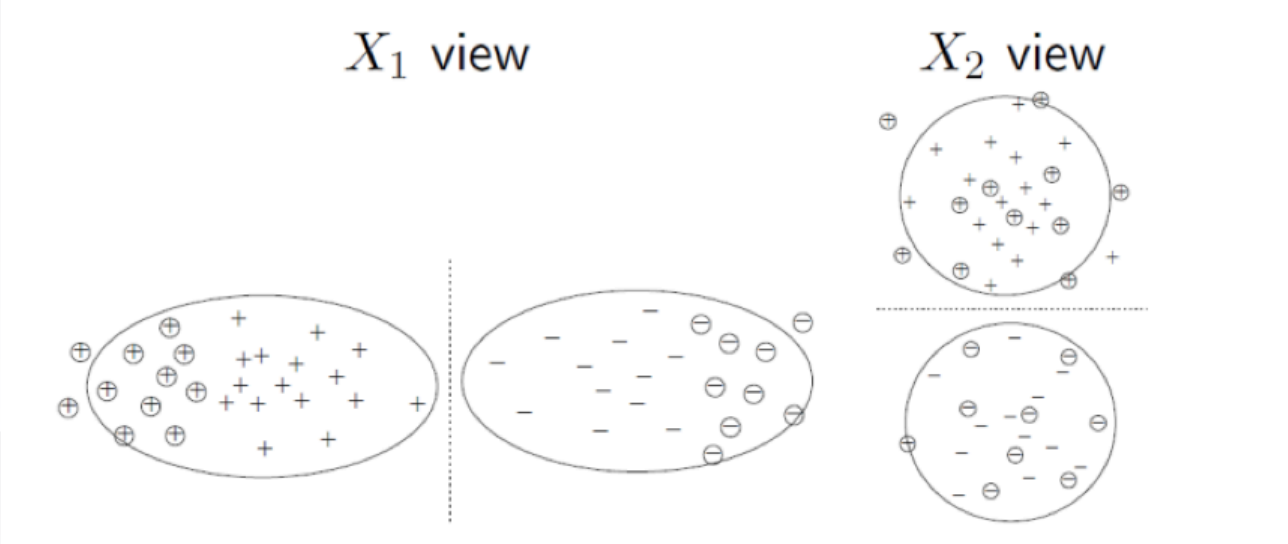


特征分裂

- 每个实例由两个特征集合$\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}]$表示
  - $\mathbf{x}^{(1)}$ = 图像特征
  - $\mathbf{x}^{(2)}$ = web 页面文本
  - 这是一个自然的特征分裂 (或者称为多视角)
- 协同训练的想法:
  - 训练一个图像分类器和一个文本分类器
  - 两个分类器互相教对方

协同训练的假设

- 假设
  - 特征可分裂$\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}]$
  - $\mathbf{x}^{(1)}$ 或 $\mathbf{x}^{(2)}$单独对于训练一个好的分类器是充分的
  - $\mathbf{x}^{(1)}$ 和 $\mathbf{x}^{(2)}$在给定类别后是条件独立的



协同训练

- 协同训练算法
  - 训练两个分类器: 从$\left(\mathbf{X}_L^{(1)}, \mathbf{y}_L\right)$学习$f^{(1)}$, 从$\left(\mathbf{X}_L^{(2)}, \mathbf{y}_l\right)$学习$f^{(2)}$
  - 用$f^{(1)}$和$f^{(2)}$分别对$\mathcal{D}_U$分类
  - 把$f^{(1)}$的k个置信度最高的预测结果$(\mathbf{x}, f^{(1)}(\mathbf{x}))$当做$f^{(2)}$的标注数据
  - 把$f^{(2)}$的k个置信度最高的预测结果$(\mathbf{x}, f^{(2)}(\mathbf{x}))$当做$f^{(1)}$的标注数据
  - 重复上述过程

协同训练的优点和缺点

- 优点
  - 简单的wrapper方法. 可以被用到已有的各种分类器
  - 相比较于自我训练，对于错误不那么敏感

- 缺点
  - 自然的特征分裂可能不存在
  - 使用全部特征的模型可能效果更好

多视角学习（Multi-view Learning）

- 半监督学习中一类通用的算法
- 基于数据的多个视角(特征表示)
  - 协同训练是多视角学习中一个特例

- 通用的想法：一致正则化
  - 多个分类器在无标签数据上应该达成一致

$$\min_f \sum_{v=1}^{M} \left(\sum_{i=1}^{L} \mathcal{L}\left(y_i, f_v(\mathbf{x}_i)\right) + \lambda_1 \|f\|_K^2 \right) + \lambda_2 \sum_{u,v=1}^{M} \sum_{i=L+1}^{N} \left(f_u(\mathbf{x}_i) - f_v(\mathbf{x}_i)\right)^2$$

M为view/学习器的数目，$\mathcal{L}(\quad)$是原来的损失函数，$\|f\|_K^2$为模型的正则项

多视角学习

- 为什么多视角学习能学得更好?
  - 学习过程实质上搜索最好的分类器
  - 通过强迫多个分类器的预测一致性， 我们减少了搜索空间
  - 希望在较少的训练数据能够找到最好的分类器
- 对于测试数据，融合多个分类器
  - 例如：投票、共识等
- 得到了一些理论结果的支持：基于多视角的半监督学习是半监督学习和集成学习的自然过渡
  - 一些集成学习者观点：只要能够使用多个学习器，即可将弱学习器性能提升到极高，无需使用未标记样本
  - 一些半监督学习者观点：只要能使用未标记样本，即可将弱学习器性能提升到极高，无需使用多学习器

构造view

■在标注数据上训练多个模型

- 相同数据，不同结构的神经网络或不同的学习算法
- 对有标签数据进行Bootstrap采样，对每个Bootstrap训练一个模型
- 对数据（标注数据和无标签数据）增加噪声

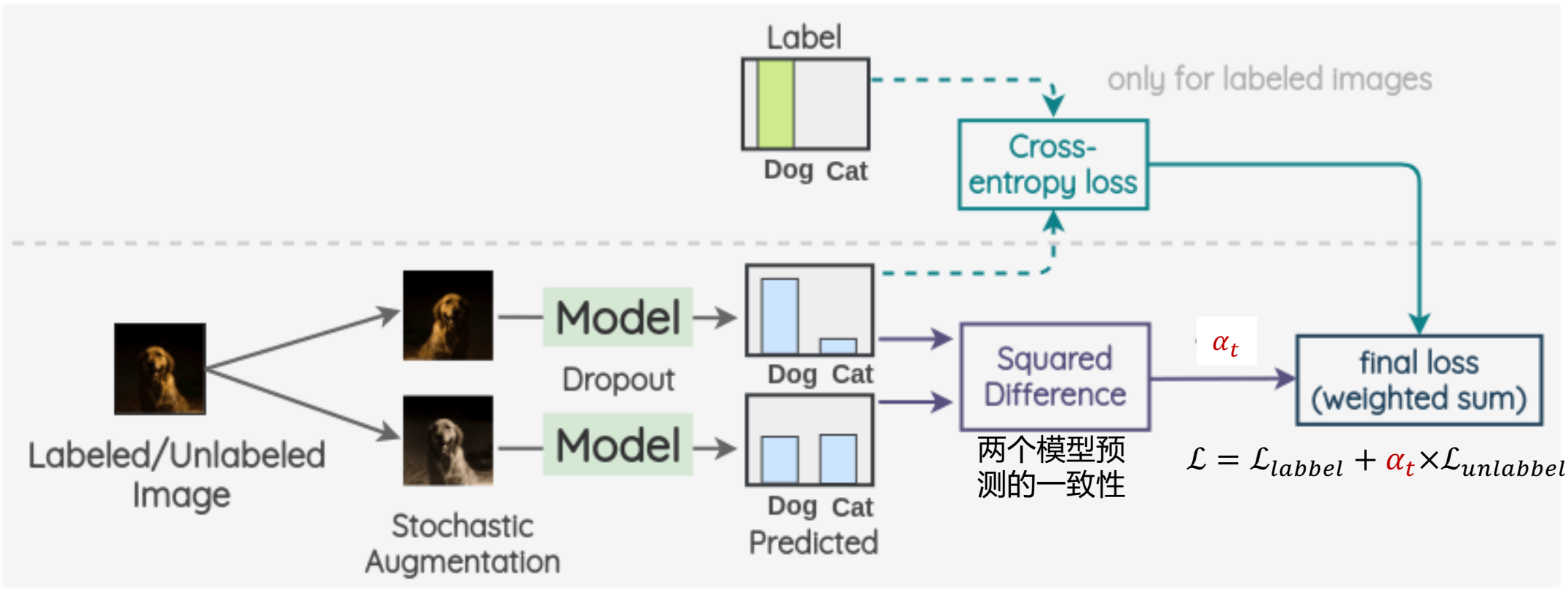
$$J_{\boldsymbol{\theta}} = CE\left(p(p(y|\mathbf{x}), \boldsymbol{\theta}), p(p(y|\mathbf{x} + \eta), \boldsymbol{\theta})\right)$$

其中η是一个较小的随机向量

■多个模型分别对无标签数据进行预测，若多个模型的结果一致（如超过一半模型）

- 将该无标签的数据的标签标为模型的预测结果，视为有标签数据
- 只将该数据加入到不一致的那些模型（view）中

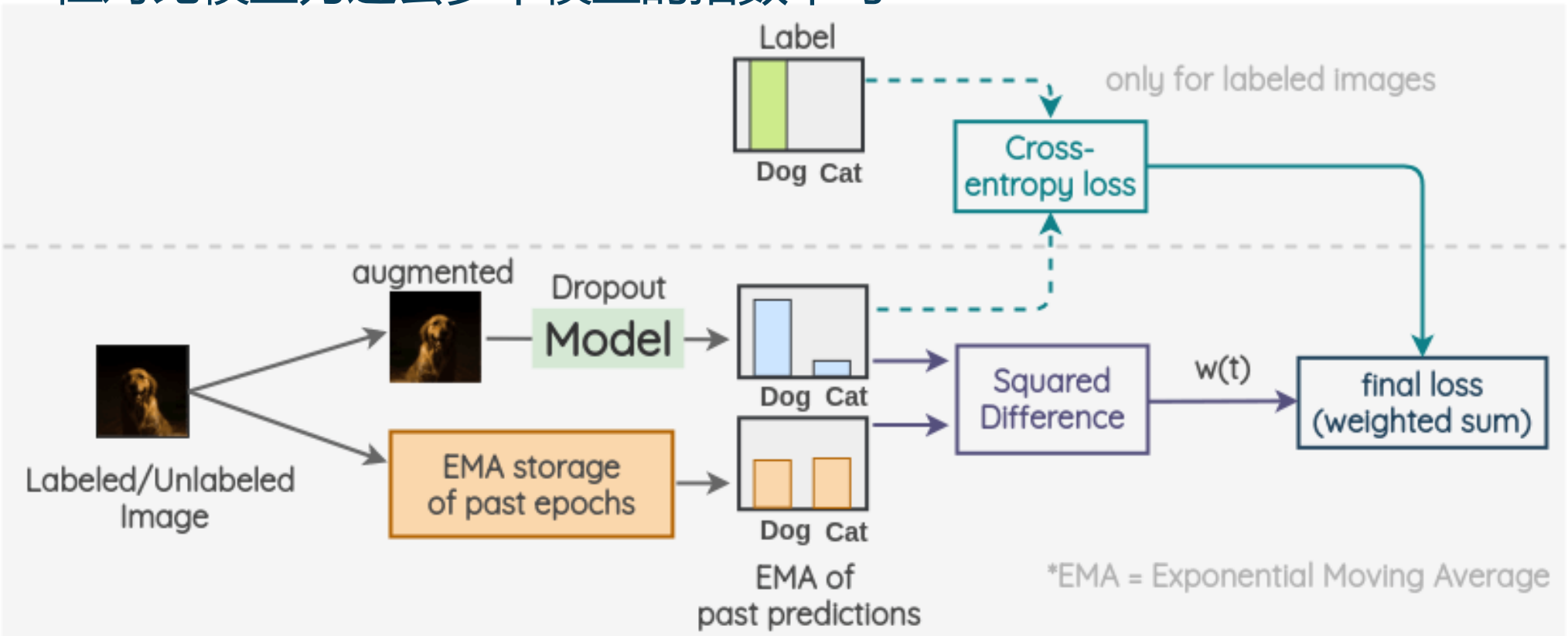
例：π-model



数据增广

例：时序集成

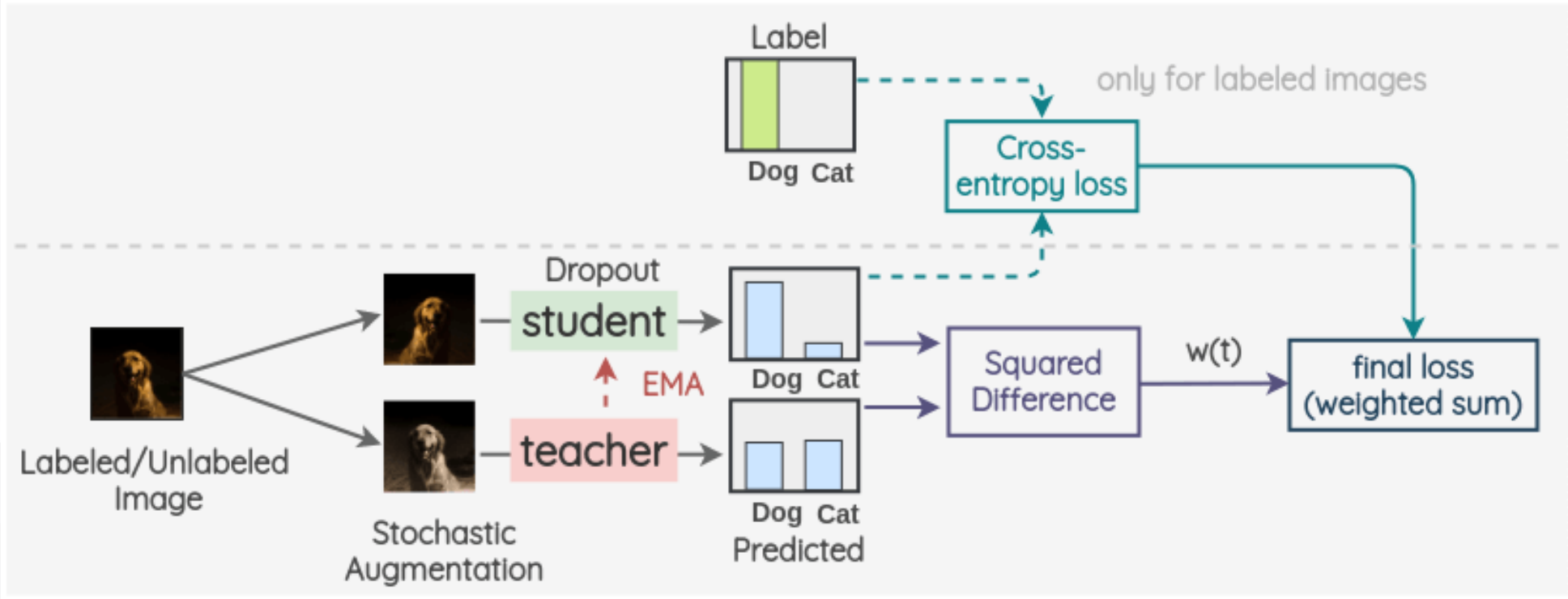
- 类似π模型
- 但对比模型为过去多个模型的指数平均



•Samuli Laine et al., "Temporal Ensembling for Semi-Supervised Learning"

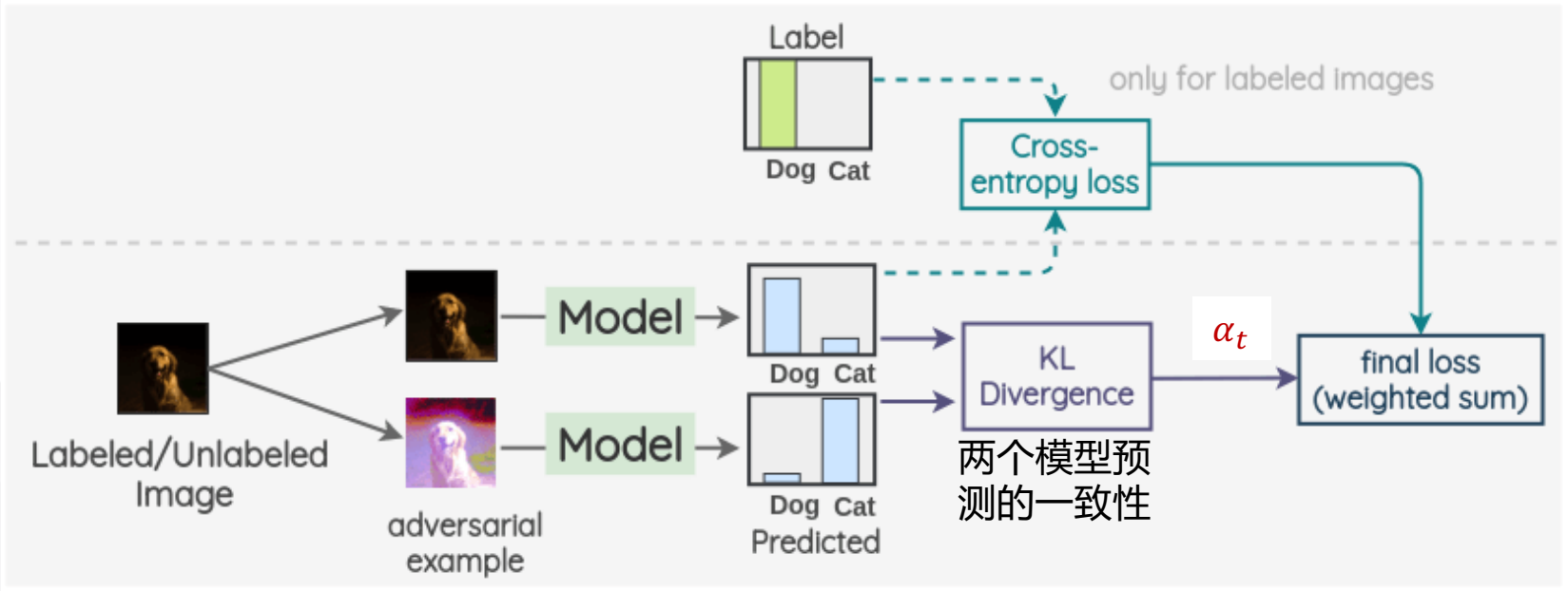
例：平均教师（Mean Teacher）

- 类似时序集成模型
- 但指数平均的不是预测值，而是模型参数



例：虚拟对抗训练

- 产生图像的对抗图像
  - 图像：first view
  - 对抗图像：second view
- 两个view的预测的KI散度：两个模型的一致性约束



一致性正则化

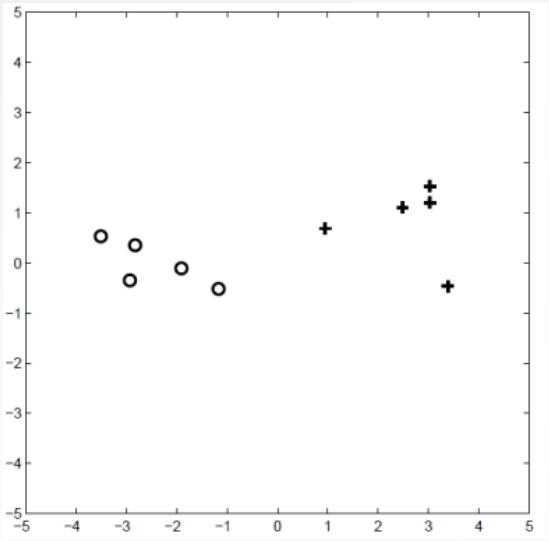
- 在图像上取得较好的效果
- 在NLP上如何应用，NLP的输入是离散的词
  - 为词向量加入噪声
  - 随机丢弃一些词

大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

生成模型的例子

- 带标签的数据 $(\mathbf{X}_L, \mathbf{y}_L)$:



假定每个类别采样自一个高斯分布, 决策的边界在哪里?

生成模型的例子

模型参数 $\boldsymbol{\theta} = \{\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sum_1, \sum_2\}$

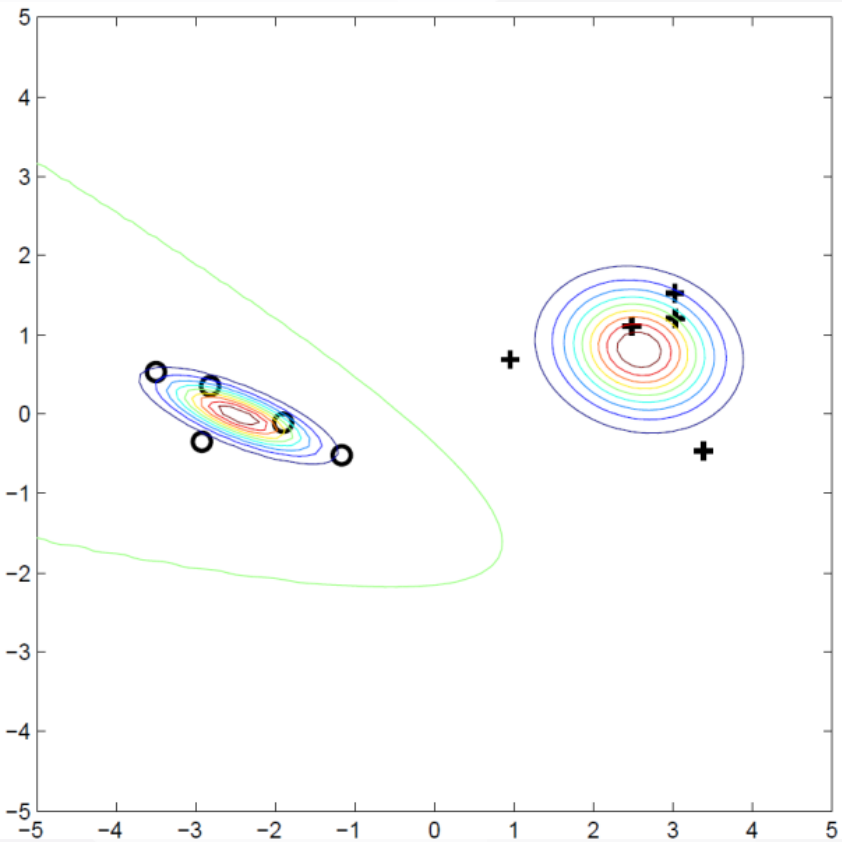
高斯混合模型:

$$p(\boldsymbol{x}, y|\boldsymbol{\theta}) = p(y|\boldsymbol{\theta})p(\mathbf{x}|y, \boldsymbol{\theta}) = \pi_y \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_y, \sum_y)$$

分类: $p(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}, y|\theta)}{\sum_{y'} p(\mathbf{x}, y'|\theta)}$

生成模型的例子

最可能的模型和它的决策边界

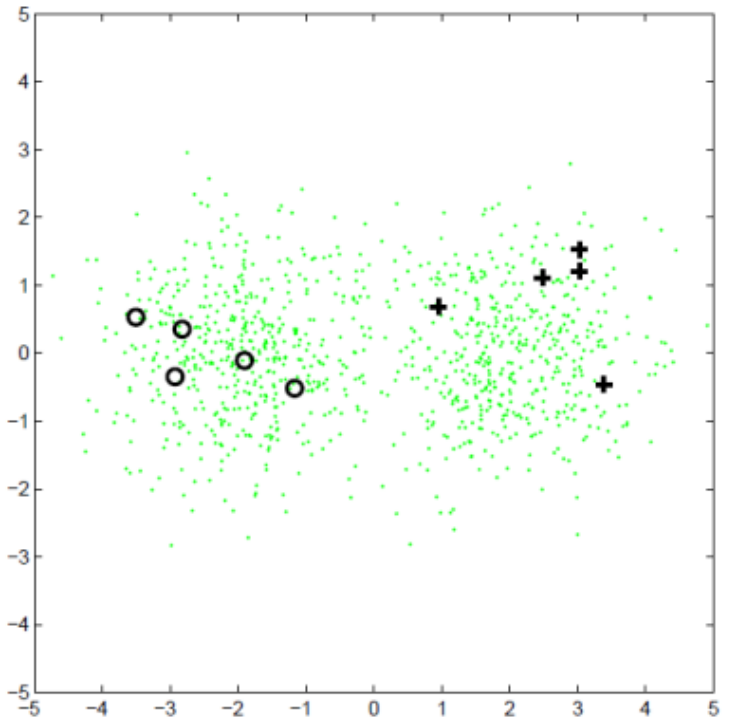


$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{y_i=c} \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{y_i=c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}}$$

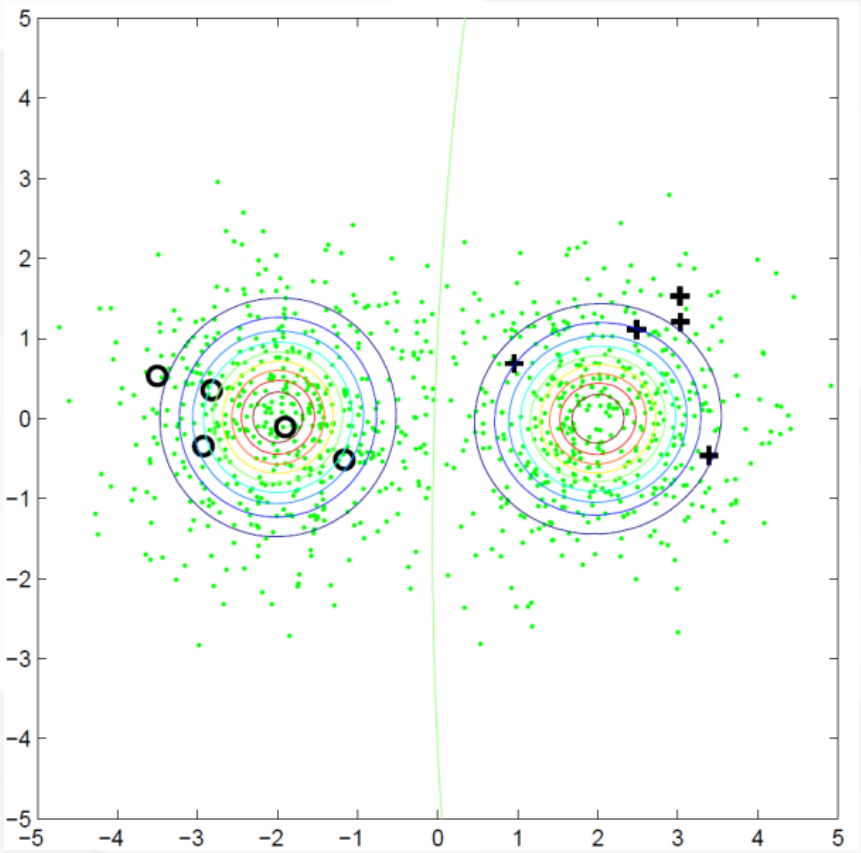
生成模型的例子

加入无标签数据



生成模型的例子

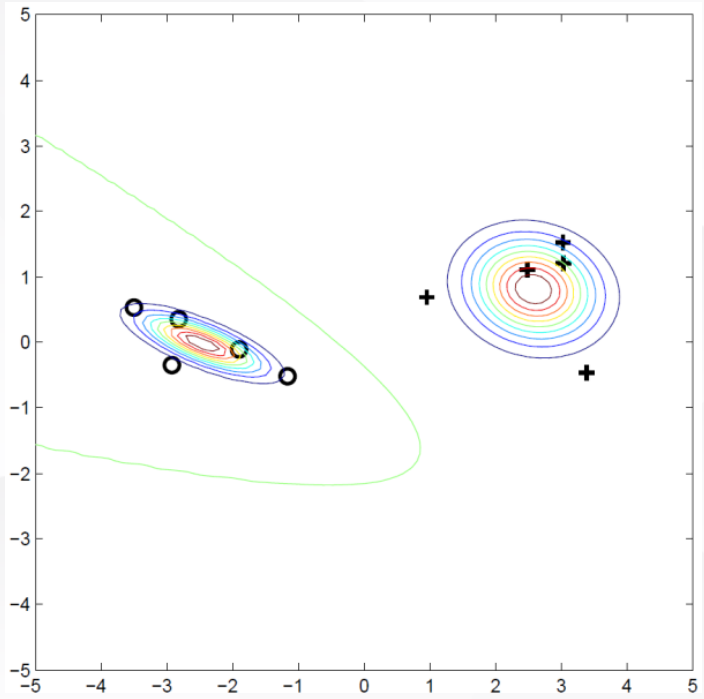
加入无标签数据, 最可能的模型和它的决策边界



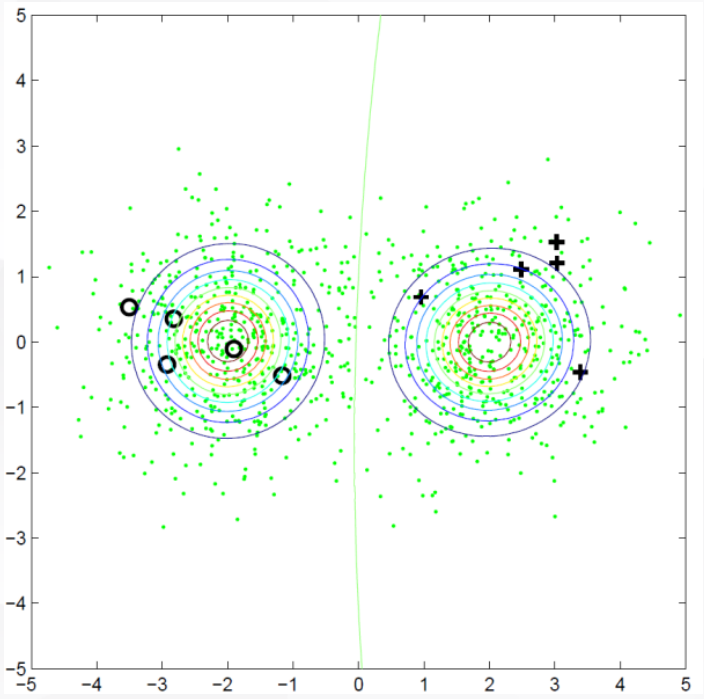
生成模型的例子

决策边界的不同，是由于模型最大化的目标不同

$p(\mathbf{X}_L, \mathbf{y}_L|\theta)$



$p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U|\theta)$



生成模型用于半监督学习

- 生成模型假设
  - 完全的生成模型 $p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})$
- 生成模型用于半监督学习:
  - 我们所感兴趣的量: $p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U|\boldsymbol{\theta}) = \sum_{\mathbf{y}_U} p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U, \mathbf{y}_U|\theta)$
- 寻找$\boldsymbol{\theta}$的极大似然估计，或最大后验估计（贝叶斯估计）

生成式模型的一些例子

在半监督学习中经常使用:

- 高斯混合模型(GMM)
- 混合多项分布 (朴素贝叶斯)
- 隐马尔科夫模型(HMM)

例分析: GMM

- 为简单起见，考虑GMM用在二分类任务，利用MLE计算参数
- 只使用标注数据
  - $\ln p(\mathbf{X}_L, \mathbf{y}_L|\boldsymbol{\theta}) = \sum_{i=1}^{L} \ln(p(y_i|\boldsymbol{\theta})p(\mathbf{x}_i|y_i, \boldsymbol{\theta}))$
  - 利用MLE 计算$\boldsymbol{\theta}$ (频率, 样本均值, 样本协方差)
- 同时考虑有标注和无标注数据

$$\ln p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U|\boldsymbol{\theta}) = \sum_{i=1}^{L} \ln(p(y_i|\boldsymbol{\theta})p(\mathbf{x}_i|y_i, \boldsymbol{\theta})) + \sum_{i=L+1}^{L+U} \ln\left(\sum_{y_i=1}^{2} p(y_i|\boldsymbol{\theta})p(\mathbf{x}_i|y_i, \boldsymbol{\theta})\right)$$

  - MLE 计算困难(包含隐变量) → EM 算法是寻找局部最优解的一个方法

EM算法用于高斯混合模型

1. 初始化：在$(\mathbf{X}_L, \mathbf{y}_L)$上用MLE估计 $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \sum\}_{1:2}$:
   - π_k=类别k的比例：$\pi_k = N_k / L$
   - $\boldsymbol{\mu}_k$=类别c的样本均值：$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{y_i=k} \boldsymbol{x}_i$
   - $\sum_k$=类别c的样本协方差：$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{y_i=k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}$

EM算法用于高斯混合模型

1. 初始化：在$(\mathbf{X}_L, \mathbf{y}_L)$上用MLE估计 $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \sum\}_{1:2}$:
2. E步:对所有$\mathbf{x}\epsilon\mathbf{X}_U$，计算类别的期望$p(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}|y, \boldsymbol{\theta})}{\sum_{y'} p(\mathbf{x}, y'|\boldsymbol{\theta})}$

$$\gamma(z_{i,k}) = p(y = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k\prime=1}^{K} \pi_{k\prime} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{k\prime}, \boldsymbol{\Sigma}_{k\prime})}, i = 1, \ldots, U, k = 1,2$$

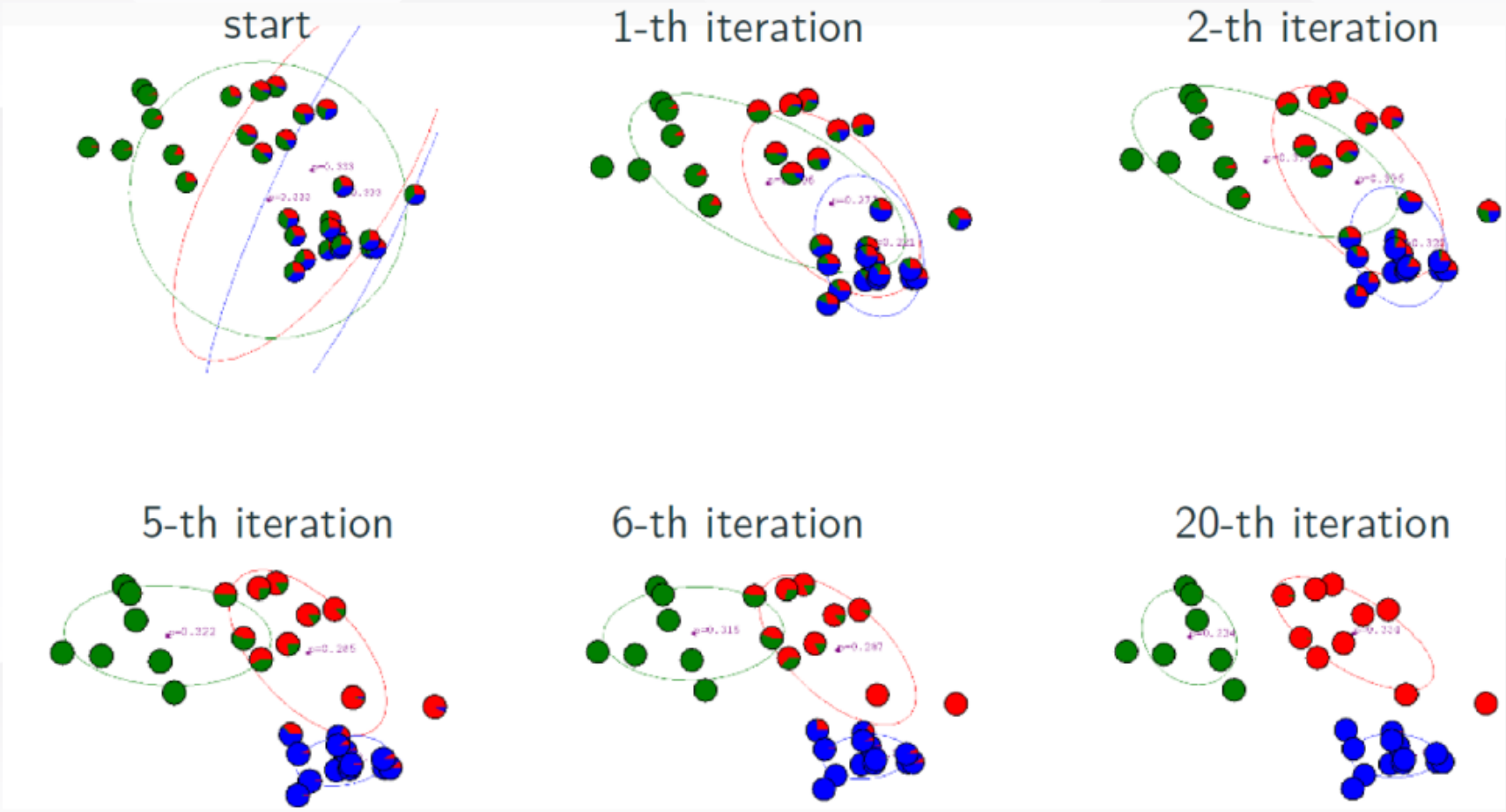
3. M步: 用有标签数据$\mathbf{X}_L$和标上标签的数据$\mathbf{X}_U$， 采用MLE估计参数 $\boldsymbol{\theta}$

$$\pi_k = \frac{\sum_{\mathbf{x}_i \epsilon \mathbf{X}_U} \gamma(z_{i,k}) + N_k}{N}, \boldsymbol{\mu}_k = \frac{\sum_{\mathbf{x}_i \epsilon \mathbf{X}_U} \gamma(z_{i,k}) \mathbf{x}_i + \sum_{\mathbf{x}_i \epsilon \mathbf{X}_L, y_i = k} \mathbf{x}_i}{\sum_{\mathbf{x}_i \epsilon \mathbf{X}_U} \gamma(z_{i,k}) + N_k},$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{\mathbf{x}_i \epsilon \mathbf{X}_U} \gamma(z_{i,k})(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} + \sum_{\mathbf{x}_i \epsilon \mathbf{X}_L, y_i = k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{\mathbf{x}_i \epsilon \mathbf{X}_U} \gamma(z_{i,k}) + N_k}$$

可以被看作自训练的一种特殊形式

EM


start
1-th iteration
2-th iteration
5-th iteration
6-th iteration
20-th iteration

生成模型用于半监督学习: 除了EM之外

- 核心是最大化$p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U|\boldsymbol{\theta})$
- EM只是最大化该概率的一种方式
- 其他能计算出使其最大化参数的方法也是可行的，如变分近似，或者直接优化

生成模型的优势

- 清晰，基于良好理论基础的概率框架
- 如果模型接近真实的分布，将会非常有效

生成模型的缺点

- 验证模型的正确性比较困难
- 模型可辨识问题(Model identifiability)
- EM局部最优
- 如果生成模型是错误，无监督数据会加重错误

减少风险的启发式方法

■ 需要更加仔细地构建生成模型，能正确建模目标任务
例如：每个类别用多个高斯分布，而不是单个高斯分布

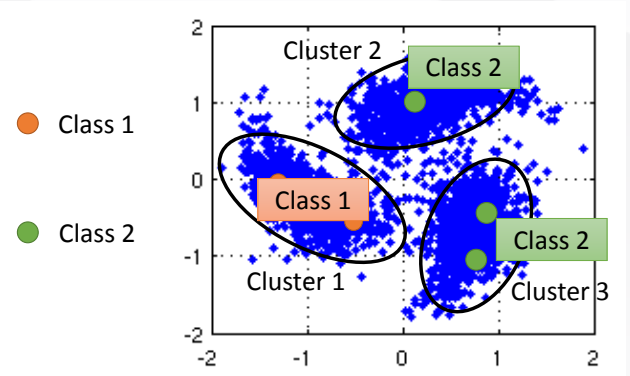
■降低无标注数据的权重

$$\ln p\left(\mathbf{X}_L, \mathbf{y}_L \mid \boldsymbol{\theta}\right) = \sum_{i=1}^{L} \ln\left(p(y_i|\boldsymbol{\theta}) p(\mathbf{x}_i|y_i, \boldsymbol{\theta})\right) + \lambda \sum_{i=L+1}^{L+U} \ln\left(\sum_{y_i=1}^{2} p(y_i|\boldsymbol{\theta}) p(\mathbf{x}_i|y_i, \boldsymbol{\theta})\right)$$

相关方法: 聚类标签法（Cluster-and-label）

除了使用概率生成模型，任何聚类算法都可以被用于半监督学习:

- 在$\mathbf{x}_1 \ldots \mathbf{x}_U$运行某种你挑选的聚类算法
- 通过计算簇内占多数的类别，将簇内所有的点标记为该类别
- 优点: 利用现有算法的一种简单方法
- 缺点: 很难去分析它的好坏。如果簇假设不正确，结果会很差

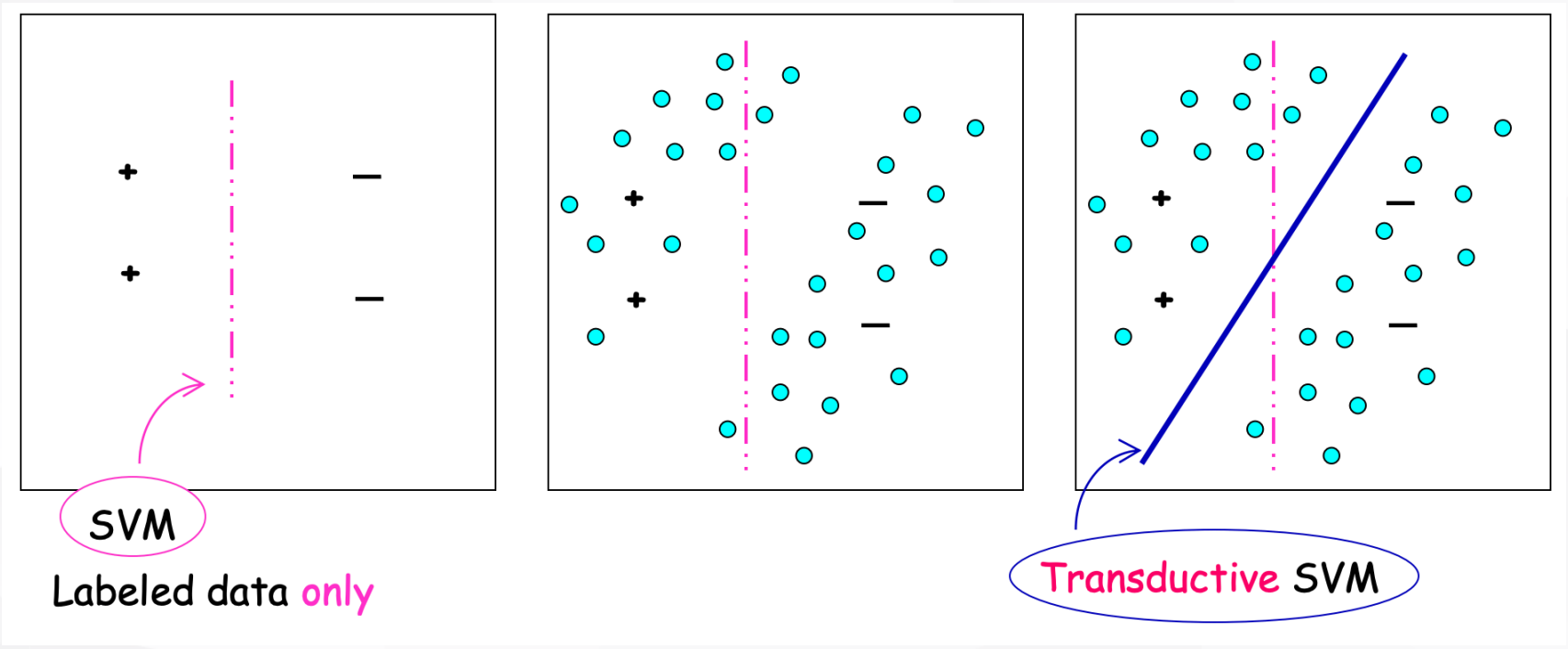


大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

半监督支持向量机

- 半监督支持向量机(Semi-supervised SVMs，简称S^3VMs) = 直推SVM (Transductive SVMs，简称TSVMs)
- 最大化“所有数据的间隔(margin)”



S³VMs

- 基本假设
  - 来自不同类别的无标记数据之间会被较大的间隔隔开
- S³VMs 的基本思想：
  - 遍历所有2^U 种可能的标注$\mathbf{X}_U$
  - 为每一种标注构建一个标准的SVM (包含 $\mathbf{X}_L$)
  - 选择间隔最大的SVM

标准SVM回顾

- 问题设置:
  - 两类$y \in \{+1, -1\}$
  - 标注数据$\{\mathbf{X}_L, \mathbf{y}_L\}$
  - 权重$\mathbf{w}$
- SVM 寻找一个函数$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$
- 通过$sign(f(\mathbf{x}))$分类$\mathbf{x}$

标准软间隔SVM

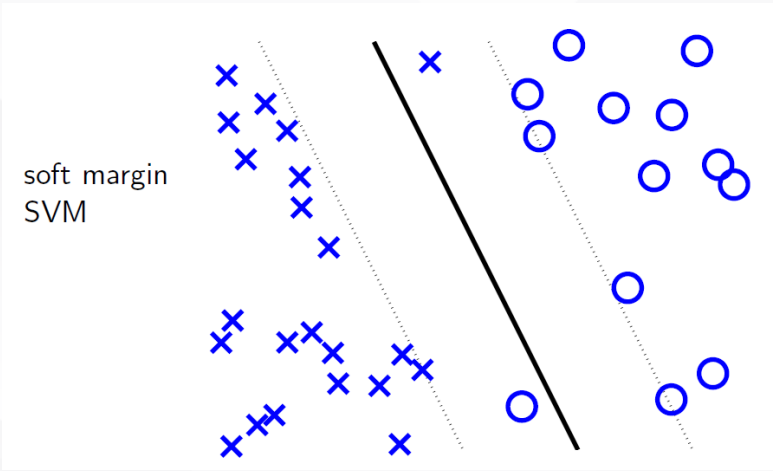
- 尝试去保持标注的点远离边界, 同时最大化间隔:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i$$

$$s.t.\, y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i = 1 \ldots L$$

$$\xi_i \geq 0$$

- ξ_i是松弛变量



SVM——合页损失

令 $z_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 - y_i f(\mathbf{x}_i)$, 目标函数

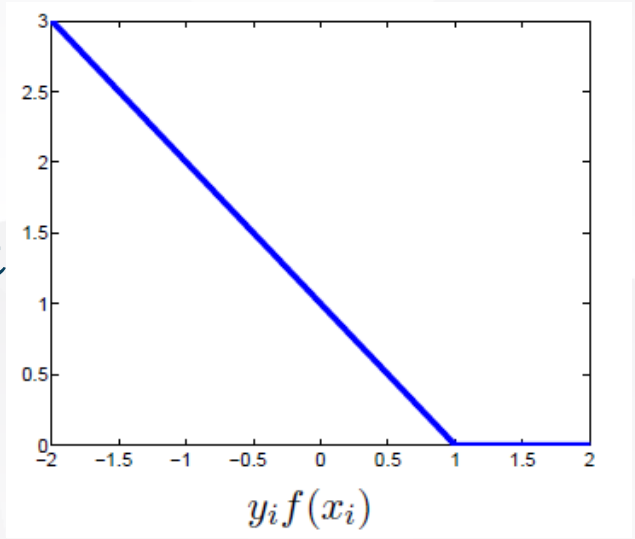
$$\min_{h,b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i$$

$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \text{ , } \forall i = 1 \cdots l$$

$$\xi_i \geq 0$$

等价于

$$\min_{f} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\underbrace{(1 - y_i f(\mathbf{x}_i))_+}_{\text{合页损失}}$$



倾向于让有标注的点在“正确”的一边

S³VM 目标函数

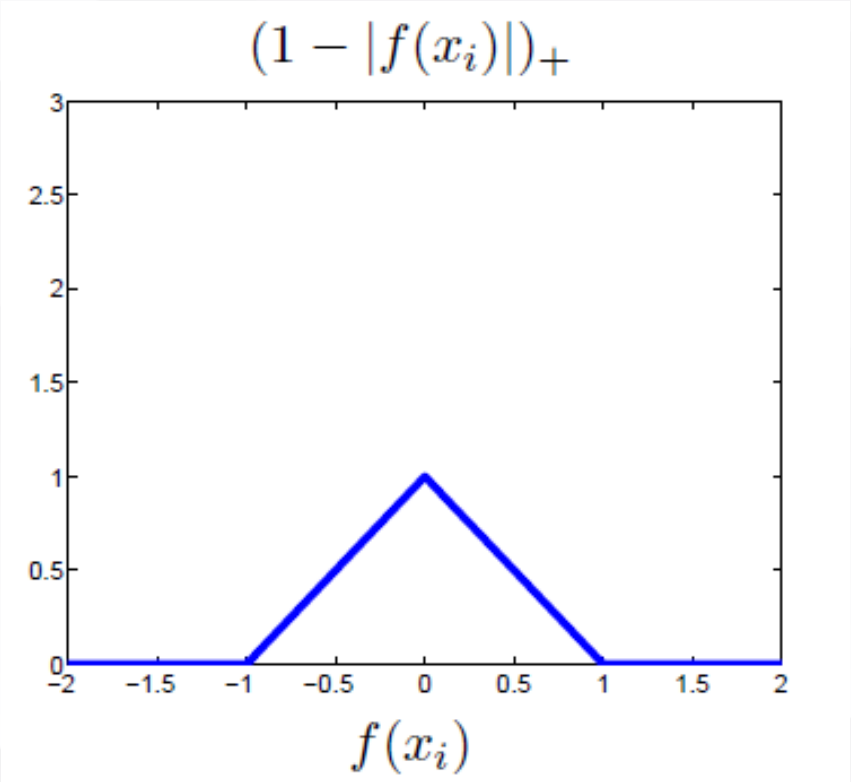
- 如何利用没有标注的点?
  - 分配标签$\text{sign}(f(\mathbf{x}))$ 给 $\mathbf{x} \in \mathbf{X}_U$
  - 无标注样本的合页损失为

$$(1 - y_i f(\mathbf{x}_i))_+ = (1 - |f(\mathbf{x}_i)|)_+$$

- S³VM 目标函数:

$$\min_f \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{L} (1 - y_i f(\mathbf{x}_i))_+ + C_2 \sum_{i=L+1}^{N} (1 - |f(\mathbf{x}_i)|)_+$$

无标注数据上的帽形损失 (hat loss)



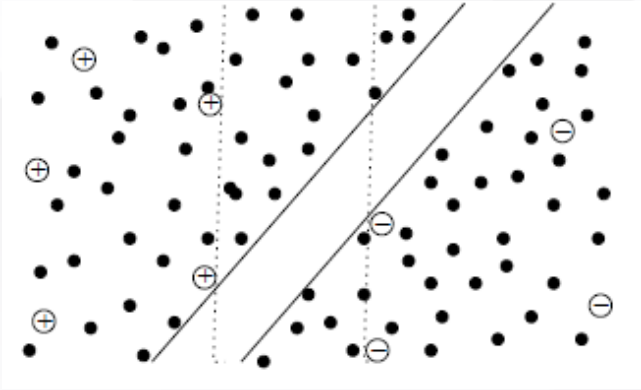
偏向$f(\mathbf{x}) \geq 1$ 或 $f(\mathbf{x}) \leq -1$, 即无标注数据远离决策边界$f(\mathbf{x}) = 0$.

避免无标签数据落在间隔内

S^{3}VM 目标函数

$$\min_f \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{L}(1 - y_i f(\mathbf{x}_i))_+ + C_2 \sum_{i=L+1}^{N}(1 - |f(\mathbf{x}_i)|)_+$$

第三项偏好无标注的点在间隔外。等价地，决策边界$f = 0$应该合理选择，使得尽可能少的无标注的点接近它。



类别平衡限制

- 直接优化S^3VM目标函数经常产生不均衡的分类 — 大多数点落在一个类内
- 启发式的类别平衡方法: $\frac{1}{N-L}\sum_{i=L+1}^{N} y_i = \frac{1}{L}\sum_{i=1}^{L} y_i$.
- 放松的类别均衡限制: $\frac{1}{N-L}\sum_{i=L+1}^{N} f(\mathbf{x}_i) = \frac{1}{L}\sum_{i=1}^{L} y_i$

S³VM 算法

- 输入: 权重$\mathbf{w}$, C_1, C_2, $(\mathbf{X}_L, \mathbf{y}_L), \mathbf{X}_U$
- 求解优化问题求 $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

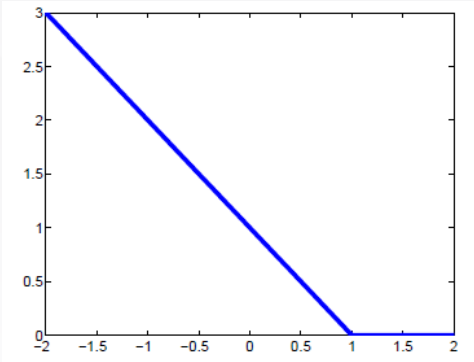
$$\min_f \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{L}(1 - y_i f(\mathbf{x}_i))_+ + C_2 \sum_{i=L+1}^{N}(1 - |f(\mathbf{x}_i)|)_+$$

$$\text{s. t.} \quad \frac{1}{N-L}\sum_{i=L+1}^{N} f(\mathbf{x}_i) = \frac{1}{L}\sum_{i=1}^{L} y_i$$

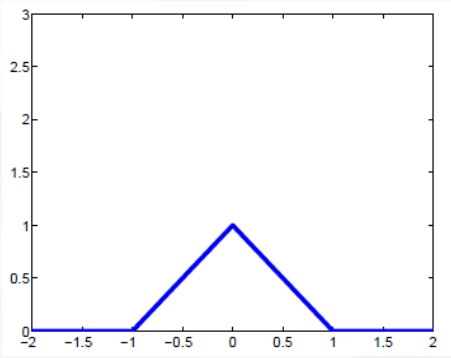
- 通过sign($f(\mathbf{x})$)分类新的测试点$\mathbf{x}$

S^3VM优化中的挑战

- SVM 目标函数是凸函数:



- 半监督SVM 目标函数是非凸的:



- 求解半监督SVM的解是困难的, 也是S^3VM研究主要关注的点

用于S^3VM训练的优化方法

- 精确方法:
  - 混合整数规划(Mixed Integer Programming)[Bennett, Demiriz; NIPS 1998]
  - 分支定界(Branch & Bound) [Chapelle, Sindhwani, Keerthi; NIPS 2006]
- 近似方法:
  - 自标注启发式S^3VM^{light} (self-labeling heuristic S^3VM^{light})[T. Joachims; ICML 1999]
  - 梯度下降(gradient descent)[Chapelle, Zien; AISTATS 2005]
  - CCCP-S^3VM [R. Collobert et al.; ICML 2006]
  - contS^3VM [Chapelle et al.; ICML 2006]

S³VM 实现1: SVM^{light}

- 局部组合搜索策略(Local combinatorial search)
- 分配一个“硬”标签到无标注数据
- 外层循环: C_2 从0开始向上退火
- 内层循环：成对标签切换

S³VM 实现1: SVM^{light}

- 用$(\mathbf{X}_L, \mathbf{y}_L)$训练一个SVM
- 根据$f(\mathbf{X}_u)$排序 $\mathbf{X}_u$ 以合适的比例标注$y = 1, -1$
- FOR $\tilde{C} \leftarrow 10^{-5}C_2 \ldots C_2$
  - REPEAT:
    - $\min_f \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{L}(1 - y_i f(\mathbf{x}_i))_+ + \tilde{C} \sum_{i=L+1}^{N}(1 - |f(\mathbf{x}_i)|)_+$
    - IF $\exists(i, j)$ 可交换THEN 交换y_i, y_j
  - UNTIL 没有标签可交换

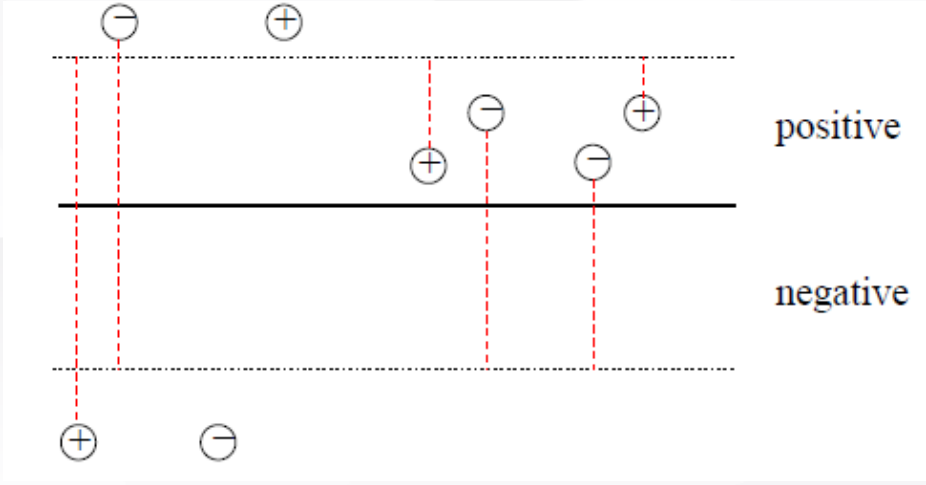
S³VM 实现1: SVM^{light}

$i, j \in \{L+1, \ldots, N\}$ 可交换 if $y_i = 1, y_j = -1$ and

$$loss\big(y_i = 1, f(\mathbf{x}_i)\big) + loss\left(y_j = -1, f\left(\mathbf{x}_j\right)\right)$$

$$> loss\big(y_i = -1, f(\mathbf{x}_i)\big) + loss(y_j = 1, f(\mathbf{x}_j))$$

hinge 损失 $loss(y, f) = (1 - yf)_+$



S^3VM 实现2: 分支定界(Branch and Bound)

- SVM^{light}实现存在局部最优的问题
- BB 能够找到精确的全局最优解
  - 使用AI中经典的分支定界搜索技术
  - 不足：只能处理数百个无标注的点

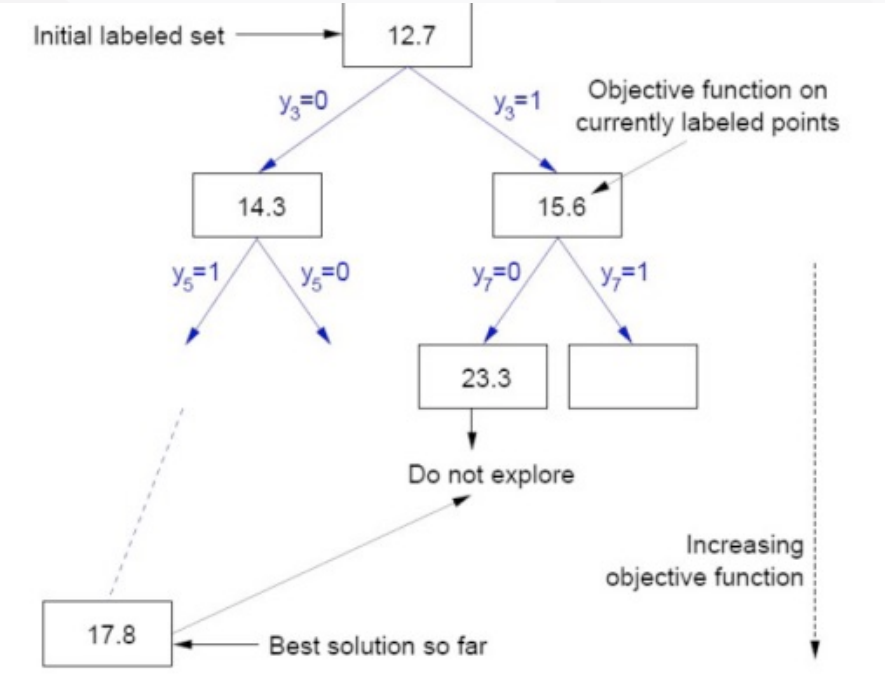
S³VM 实现2: 分支定界

- 组合优化问题
- 在$\mathbf{X}_U$上构建一棵部分标注的树
  - 根节点：$\mathbf{X}_U$没有标注
  - 子节点：比父节点多一个数据 $\mathbf{x} \in \mathbf{X}_U$被标注
  - 叶子节点: 所有 $\mathbf{x} \in \mathbf{X}_U$ 被标注
- 部分标注有一个非减(non-decreasing)的 S³VM 目标函数

$$\min_f \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{L} (1 - y_i f(\mathbf{x}_i))_+ + C_2 \sum_{i \in labeled\ so\ far} (1 - |f(\mathbf{x}_i)|)_+$$

S³VM 实现2: 分支定界

- 在树上进行深度优先搜索
- 记录一个到当前为止的完整目标函数值
- 如果它比最好的目标函数差，就进行剪枝(包括它的子树)



S^3VMs总结

- 优点:
  - 可以被用在任何SVMs 可以被应用的地方
  - 清晰的数学框架
- 缺点:
  - 优化困难
  - 可能陷入局部最优
  - 相比于生成模型和基于图的方法使用更弱的假设, 收益可能较小

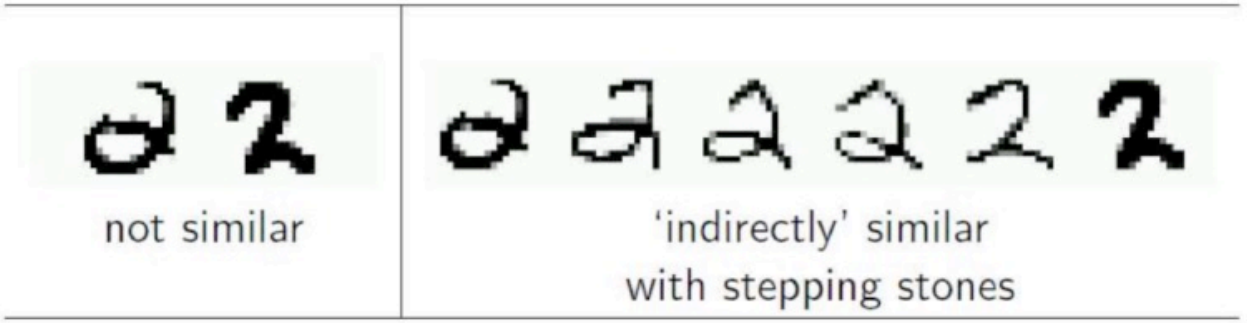
大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

基于图的半监督学习

- 假设
  - 假定的标在有标注和无标注数据上存在一个图，图中被“紧密”连接的点趋向于有相同签。
- 在图上标签的变化应该是平滑的。
  - 邻近结点应该有相似的标签

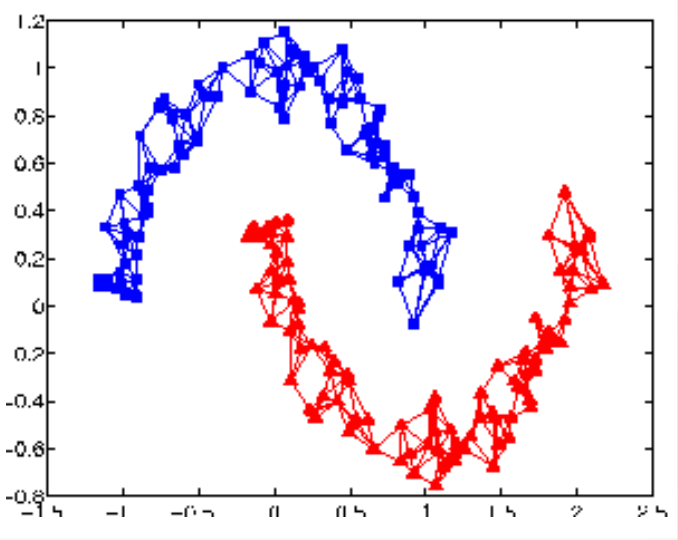
Handwritten digits recognition with pixel-wise Euclidean distance



我们称之为标签传播

图

- 节点: $\mathbf{X}_L \cup \mathbf{X}_U$
- 边：权重是基于特征来计算相邻节点之间的相似度，例如,
  - k最近邻图, 无权重(0, 1 权重)
  - 全连接图, 权重随距离衰减：$w_{i,j} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$
  - ε-半径(ε-radius)图
- 想要的结果：通过所有的路径来推导相似度



标签传播

- 标签沿图传播：标注数据影响其邻居
- 图上标签的变化是平滑的：邻近结点应该有相似的标签
- 图的平滑性定义：

对所有样本（有标签样本和无标签样本）

$$S = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(\hat{y}_i - \hat{y}_j)^2$$

越小越平滑

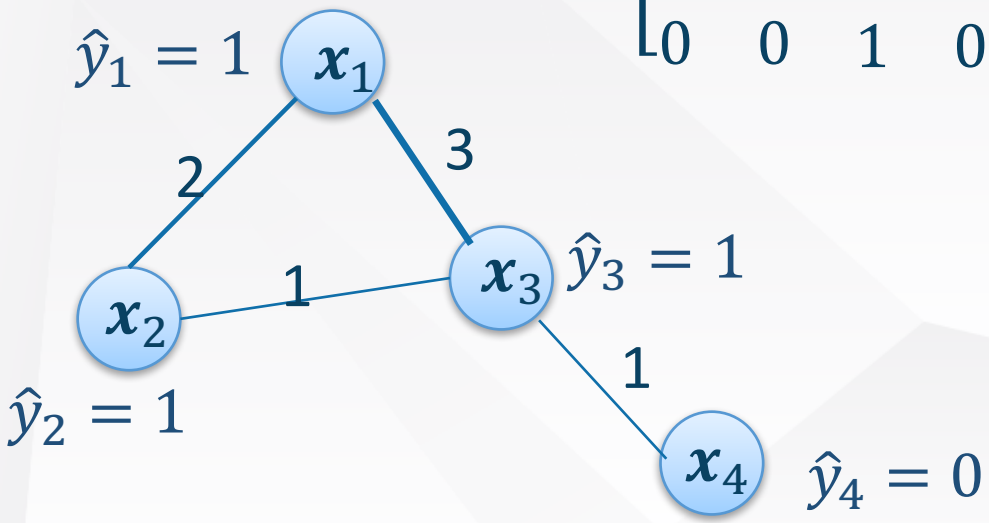
$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)^2$$

与拉普拉斯映射比较：

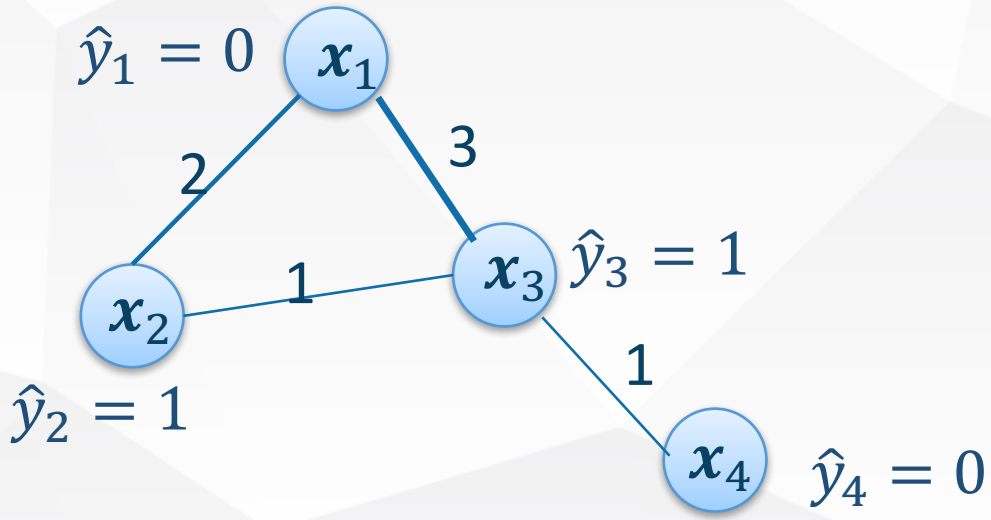
$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)^2$$

例：图的平滑性

- $\boldsymbol{W} = \begin{bmatrix} 0 & 2 & 3 & 0 \\ 2 & 0 & 1 & 0 \\ 3 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ $\mathbf{X}_L \cup \mathbf{X}_U$上$N \times N$ 权重矩阵$\mathbf{W}$



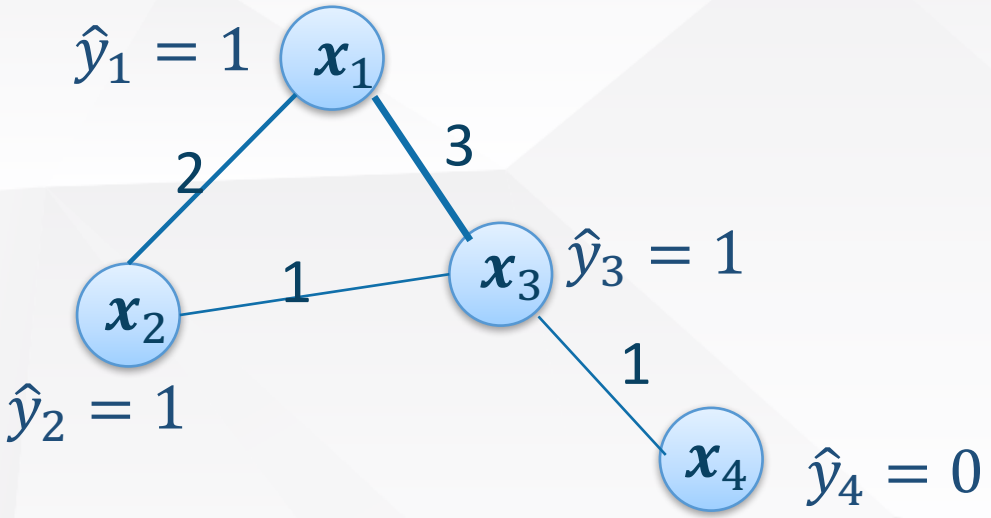
$S = 1$



$$S = \frac{1}{2}(5 + 2 + 3 + 1 + 1) = 6$$

越小越平滑

拉普拉斯矩阵



$$S = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(\hat{y}_i - \hat{y}_j)^2$$

$$= \frac{1}{2}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{L}\hat{\boldsymbol{y}}$$

其中 $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N)^{\mathrm{T}}$

$$\boldsymbol{W} = \begin{bmatrix} 0 & 2 & 3 & 0 \\ 2 & 0 & 1 & 0 \\ 3 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

对角的度矩阵（Diagonal degree matrix） $d_{i,i} = \sum_{j=1}^{N} w_{i,j}$

$$\boldsymbol{D} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

拉普拉斯矩阵：

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W} = \begin{bmatrix} 5 & -2 & -3 & 0 \\ -2 & 3 & -1 & 0 \\ -3 & -1 & 5 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

拉普拉斯矩阵

- 图的平滑性函数（亦可视为能量函数）：

$$S(f)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}w_{i,j}\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)^2=\frac{1}{2}\boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f}$$

- 对f求导：$\frac{\partial S(f)}{\partial f}=0\rightarrow \boldsymbol{L}\boldsymbol{f}=\mathbf{0}$
- 将矩阵分块表示：$\mathbf{W}=\begin{bmatrix}\mathbf{W}_{LL} & \mathbf{W}_{LU}\\ \mathbf{W}_{UL} & \mathbf{W}_{UU}\end{bmatrix}$，$\mathbf{D}=\begin{bmatrix}\mathbf{D}_{LL} & \mathbf{0}_{LU}\\ \mathbf{0}_{UL} & \mathbf{D}_{UU}\end{bmatrix}$，则
- 由于$\boldsymbol{L}\boldsymbol{f}=0$且$\boldsymbol{f}_L=\boldsymbol{Y}_L$，$\left(\begin{bmatrix}\mathbf{D}_{LL} & \mathbf{0}_{LU}\\ \mathbf{0}_{UL} & \mathbf{D}_{UU}\end{bmatrix}-\begin{bmatrix}\mathbf{W}_{LL} & \mathbf{W}_{LU}\\ \mathbf{W}_{UL} & \mathbf{W}_{UU}\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{f}_L\\ \boldsymbol{f}_U\end{bmatrix}=\mathbf{0}$
- $-\mathbf{W}_{UL}\boldsymbol{f}_L+(\mathbf{D}_{UU}-\mathbf{W}_{UU})\boldsymbol{f}_U=\mathbf{0}\quad\rightarrow \boldsymbol{f}_U=(\mathbf{D}_{UU}-\mathbf{W}_{UU})^{-1}\mathbf{W}_{UL}\boldsymbol{f}_L$

利用拉普拉斯计算最优解

■令$\mathbf{P}=\mathbf{D}^{-1}\mathbf{W}=\begin{bmatrix}\mathbf{D}_{LL}^{-1} & \mathbf{0}_{LU}\\ \mathbf{0}_{UL} & \mathbf{D}_{UU}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{W}_{LL} & \mathbf{W}_{LU}\\ \mathbf{W}_{UL} & \mathbf{W}_{UU}\end{bmatrix}=\begin{bmatrix}\mathbf{D}_{LL}^{-1}\mathbf{W}_{LL} & \mathbf{D}_{LL}^{-1}\mathbf{W}_{LU}\\ \mathbf{D}_{UU}^{-1}\mathbf{W}_{UL} & \mathbf{D}_{UU}^{-1}\mathbf{W}_{UU}\end{bmatrix}$，

■所以$\mathbf{P}_{UU}=\mathbf{D}_{UU}^{-1}\mathbf{W}_{UU}$，$\mathbf{P}_{UL}=\mathbf{D}_{UU}^{-1}\mathbf{W}_{UL}$

■$\boldsymbol{f}_U=(\mathbf{D}_{UU}-\mathbf{W}_{UU})^{-1}\mathbf{W}_{UL}\boldsymbol{f}_L$

$=\left(\mathbf{D}_{UU}\left(\mathbf{I}-\mathbf{D}_{UU}^{-1}\mathbf{W}_{UU}\right)\right)^{-1}\mathbf{W}_{UL}\boldsymbol{f}_L$

$=\left(\mathbf{I}-\mathbf{D}_{UU}^{-1}\mathbf{W}_{UU}\right)^{-1}\mathbf{D}_{UU}^{-1}\mathbf{W}_{UL}\boldsymbol{f}_L$

$=(\mathbf{I}-\mathbf{P}_{UU})^{-1}\mathbf{P}_{UL}\boldsymbol{f}_L$

最优解

最优解存在条件：$\mathbf{I}-\mathbf{P}_{UU}$可逆（每个连接组件都至少有一个已标记点）

扩展到多类

- 令标记矩阵$\mathbf{F}$为非负标记矩阵$[(L+U)\times C]$，其中C为标签y的类别数目
- $\mathbf{F}=(\mathbf{F}_1^T,\mathbf{F}_2^T,\ldots,\mathbf{F}_{L+U}^T)^T$
- $\mathbf{F}$的第i行元素$\mathbf{F}_i=\left((\mathbf{F})_{i,1},\ldots,(\mathbf{F})_{i,C}\right)$为示例$\mathbf{x}_i$的标记向量
  - 分类规则：$y_i=\arg\max_{1\le j\le C}(\mathbf{F})_{i,c}$
  - 对$i=1,2,\ldots,L+U, j=1,2,\ldots,C$，$\mathbf{F}$初始化为

$$(\mathbf{F})_{i,c}^{(0)}=(\mathbf{Y})_{i,c}=\begin{cases}1 & if\ (1\le i\le L)\wedge(y_i=c)\\ 0 & otherwise\end{cases}$$

标签的独热编码

扩展到多类

- 基于W构造标记传播矩阵：$\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$，其中$\mathbf{D}^{-1/2} =$

  $\mathrm{diag}\left(\frac{1}{\sqrt{d_1}}, \frac{1}{\sqrt{d_2}}, \ldots, \frac{1}{\sqrt{d_N}}\right)$

- 则迭代计算： $\mathbf{F}^{(t+1)} = \alpha\mathbf{S}\mathbf{F}^{(t)} + (1-\alpha)\,\mathbf{Y}$

- 会收敛到：$\mathbf{F}^* = \lim_{t\to\infty}\mathbf{F}^{(t)} = (1-\alpha)(\mathbf{I}-\alpha\mathbf{S})^{-1}\mathbf{Y}$，从$\mathbf{F}^*$得到无标签样的

  标签：$\hat{y}_i = \arg\max_{1\le j\le C}(\mathbf{F}^*)_{i,c}$

- 其中$\alpha \in (0,1)$为用户指定参数。

扩展到多类

- 该算法对应正则化框架

$$min_{\mathbf{F}} \frac{1}{2}\left(\sum_{i,j=1}(W)_{ij}\left\|\frac{1}{\sqrt{d_i}}\mathbf{F}_i-\frac{1}{\sqrt{d_j}}\mathbf{F}_j\right\|^2\right)+\mu\sum_{i=1}^{L}\|\mathbf{F}_i-\mathbf{Y}_i\|^2$$

- 其中$\mu=\frac{1-\alpha}{\alpha}$。
- 允许$f(\mathbf{X}_L)$ 不同于$\mathbf{y}_L$，但是加以惩罚
- 引入标注数据（全局）和图能量（局部）之间的平衡

更一般地，基于图的半监督学习

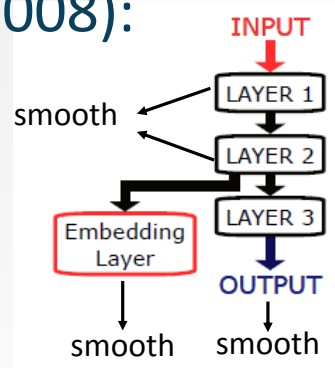
- 目标函数：

$$J = \sum_{i=1}^{L} \mathcal{L}(\hat{y}_i, y_i) + \lambda S$$

平滑性作为正则项

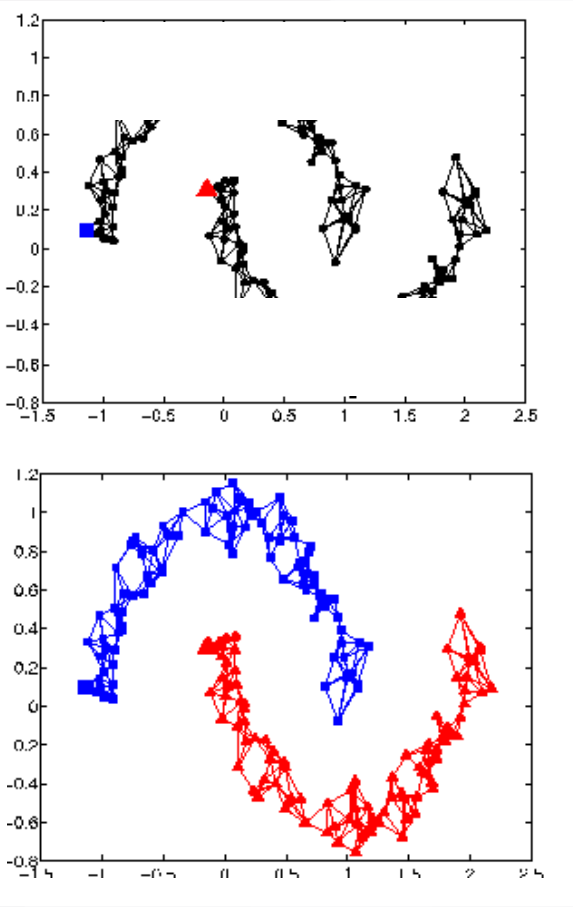
标注数据上的损失和

- LapSVM (Belkin et al., 2006)：$\mathcal{L}(\hat{y}_i, y_i)$ = 合页损失+w正则
- Deep learning via semi-supervised embedding(Weston et al., 2008):
  - 最终输出层和中间层输出均增加平滑性约束



基于图的算法的总结

- 优点:
  - 清晰的数学框架
- 缺点:
  - 图质量差的时候性能差
  - 对图的结构和权重敏感
  - 存储需求大

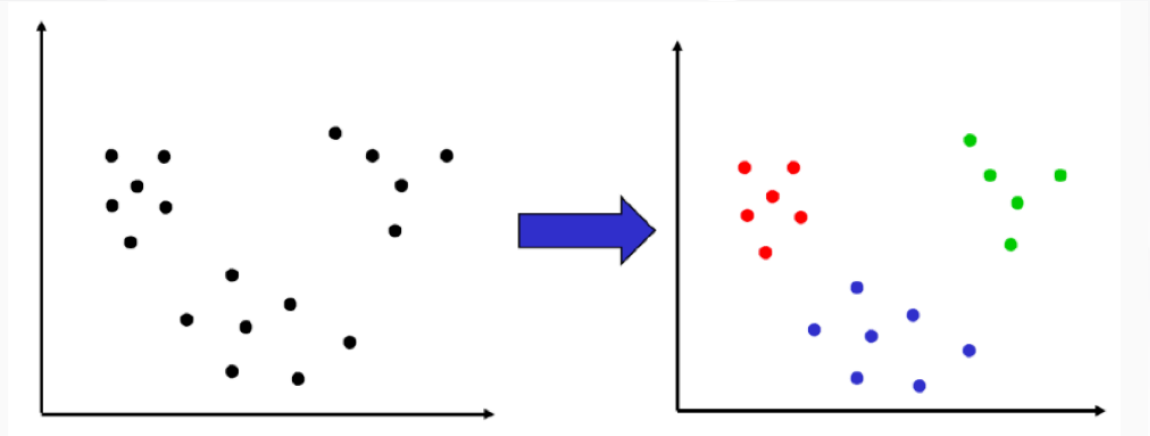


大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类

半监督聚类

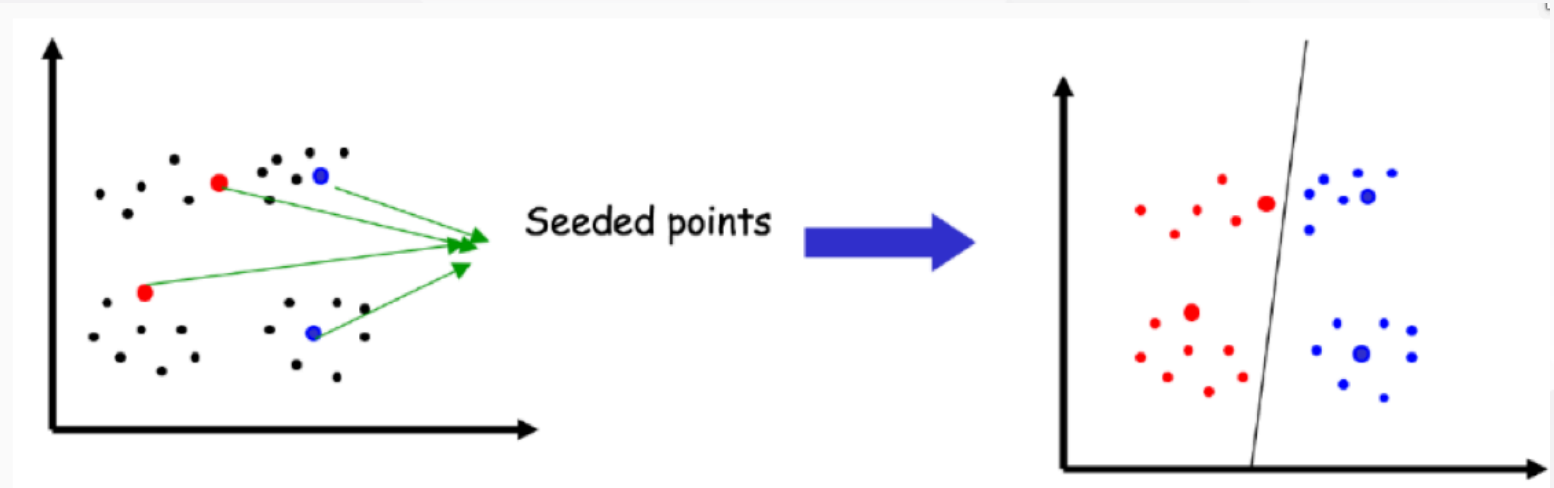
- 聚类是无监督学习的一种算法



- 半监督聚类: 聚类并加入一系列领域知识

半监督聚类

- 根据给定的不同的领域知识：
  - 用户预先提供一些种子样本的类别标签



半监督聚类

- 已知少量的标记信息

输入：样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$；
少量有标记样本 $S=\bigcup_{j=1}^{k}S_j$；
聚类簇数 k.

过程：

1: **for** $j=1,2,\ldots,k$ **do**
2: $\quad\boldsymbol{\mu}_j=\frac{1}{|S_j|}\sum_{\boldsymbol{x}\in S_j}\boldsymbol{x}$
3: **end for**
4: **repeat**
5: $\quad C_j=\varnothing\ (1\leqslant j\leqslant k)$;
6: $\quad$**for** $j=1,2,\ldots,k$ **do**
7: $\quad\quad$**for all** $\boldsymbol{x}\in S_j$ **do**
8: $\quad\quad\quad C_j=C_j\bigcup\{\boldsymbol{x}\}$
9: $\quad\quad$**end for**
10: $\quad$**end for**
11: $\quad$**for all** $\boldsymbol{x}_i\in D\setminus S$ **do**
12: $\quad\quad$计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j\ (1\leqslant j\leqslant k)$ 的距离：$d_{ij}=\|\boldsymbol{x}_i-\boldsymbol{\mu}_j\|_2$；
13: $\quad\quad$找出与样本 $\boldsymbol{x}_i$ 距离最近的簇：$r=\arg\min_{j\in\{1,2,\ldots,k\}}d_{ij}$；
14: $\quad\quad$将样本 $\boldsymbol{x}_i$ 划入相应的簇：$C_r=C_r\bigcup\{\boldsymbol{x}_i\}$
15: $\quad$**end for**
16: $\quad$**for** $j=1,2,\ldots,k$ **do**
17: $\quad\quad\boldsymbol{\mu}_j=\frac{1}{|C_j|}\sum_{\boldsymbol{x}\in C_j}\boldsymbol{x}$;
18: $\quad$**end for**
19: **until** 均值向量均未更新

输出：簇划分 $\{C_1,C_2,\ldots,C_k\}$

基于标签数据计算均值

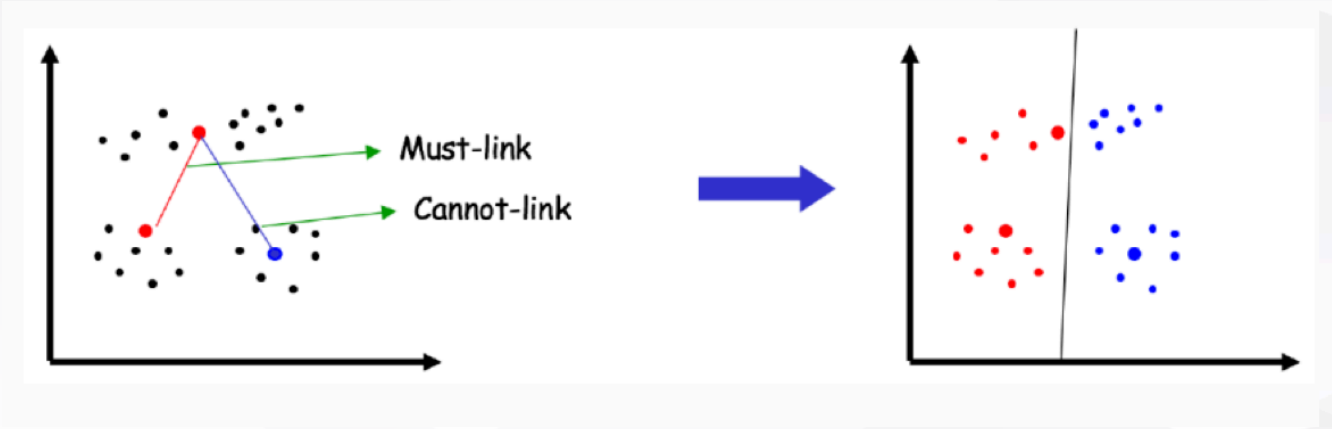
把种子节点放到对应的簇中

对无标签样本进行簇划分

更新簇中心

半监督聚类

- 根据给定的不同的领域知识：
  - 用户知道其中一些样本是相关 (must-link)的还是不相关(cannot-link)



半监督聚类

- 已知相关 (must-link)不相关 (cannot-link)

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
必连约束集合 $\mathcal{M}$;
勿连约束集合 $\mathcal{C}$;
聚类簇数 k.

过程:

```
从 D 中随机选取 k 个样本作为初始均值向量{μ_1, μ_2, ..., μ_k};
repeat
    C_j = ∅ (1 ⩽ j ⩽ k);
    for i = 1, 2, ..., m do
        计算样本 x_i 与各均值向量 μ_j (1 ⩽ j ⩽ k) 的距离: d_ij = ||x_i − μ_j||_2 ;
        K = {1, 2, ..., k};
        is_merged=false;
        while ¬ is_merged do
            基于 K 找出与样本 x_i 距离最近的簇: r = arg min_{j∈K} d_ij ;
            检测将 x_i 划入聚类簇 C_r 是否会违背 M 与 C 中的约束;
            if ¬ is_voilated then
                C_r = C_r ⋃ {x_i};
                is_merged=true
            else
                K = K \ {r};
                if K = ∅ then
                    break并返回错误提示
                end if
            end if
        end while
    end for
    for j = 1, 2, ..., k do
        μ_j = (1/|C_j|) Σ_{x∈C_j} x;
    end for
until 均值向量均未更新
```

输出: 簇划分 $\{C_1, C_2, \ldots, C_k\}$

总结：半监督学习中的各种约束

- 标注数据上的损失（监督）
- 伪标注数据上的损失（无监督）
- 模型输出的最小熵（无监督）
- 一致性约束（无监督）： L2损失、交叉熵损失
  - 多个view之间的一致性
  - 流形上标签的平滑性（马尔可夫性）
  - 合成相似样本（数据增广），相似样本标签的一致性

参考文献

- 周志华，《机器学习》
- Xiaojin Zhu and Andrew B. Goldberg. Introduction to Semi-Supervised Learning. Morgan & Claypool, 2009.
- Olivier Chapelle, Alexander Zien, Bernhard Scholkopf (Eds.). (2006). Semi- supervised learning. MIT Press.
- Jesper E Van Engelen, Holger H Hoos, A survey on semi-supervised learning, *Machine Learning* 109**,** 373–440 (2020)
- Yassine Ouali, Céline Hudelot, Myriam Tami, An Overview of Deep Semi-Supervised Learning, 2020, https://arxiv.org/pdf/2006.05278.pdf
- Awesome Semi-Supervised Learning
  - https://github.com/yassouali/awesome-semi-supervised-learning