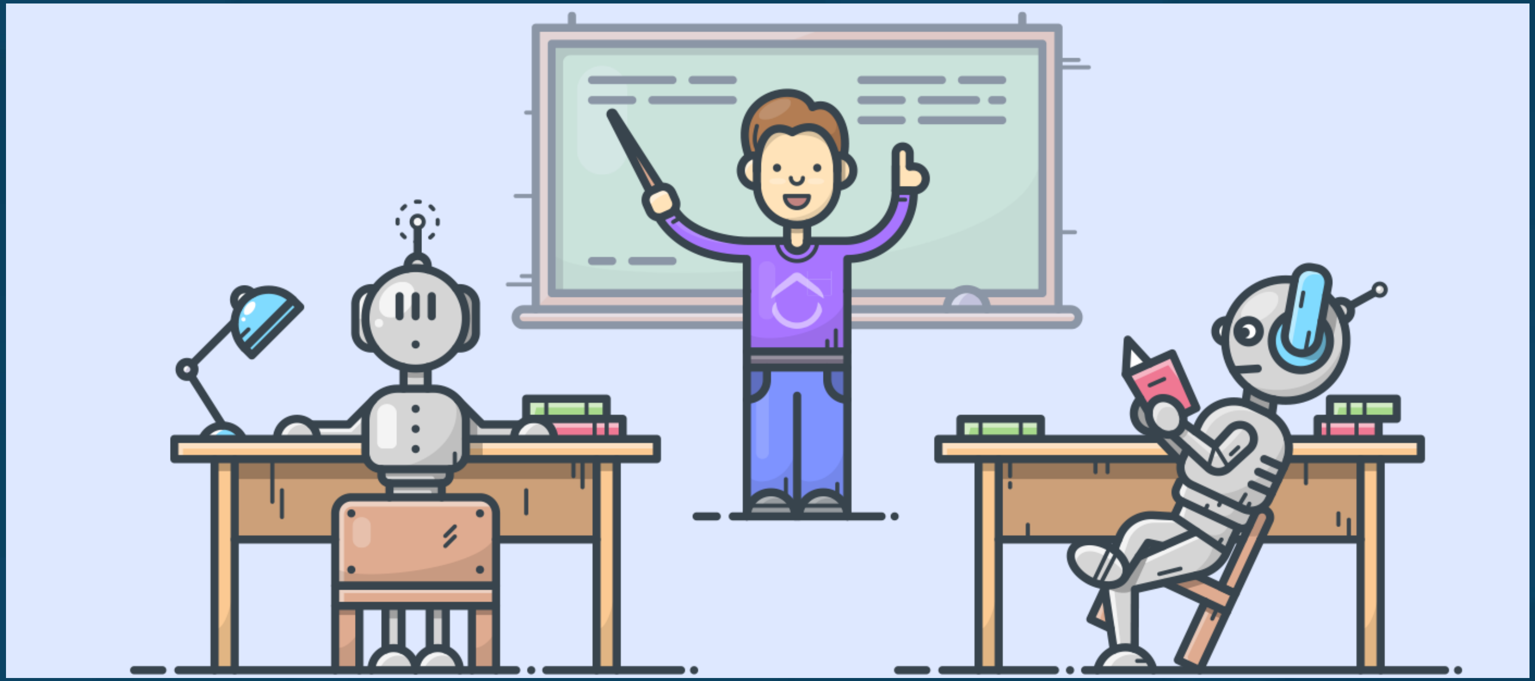
半监督学习



郭嘉丰

中国科学院大学，中国科学院计算技术研究所

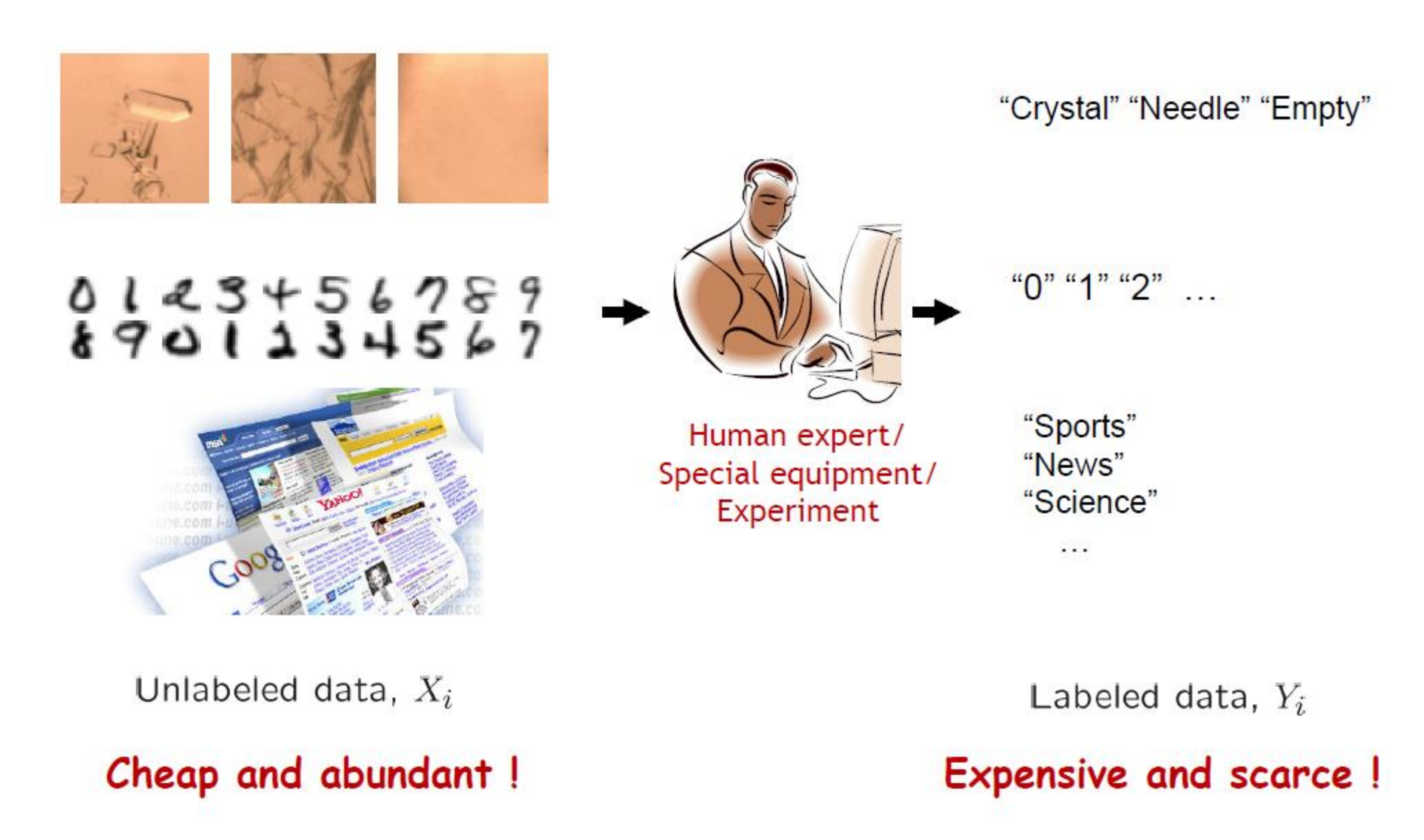
课程代码：https://github.com/lixinsu/tutorials2018/blob/master/semi_supervise.ipynb

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
- 前沿进展

监督学习

- 监督学习需要标注数据，学习一个可靠的模型需要大量标注数据
- 但是标注数据费时费力！

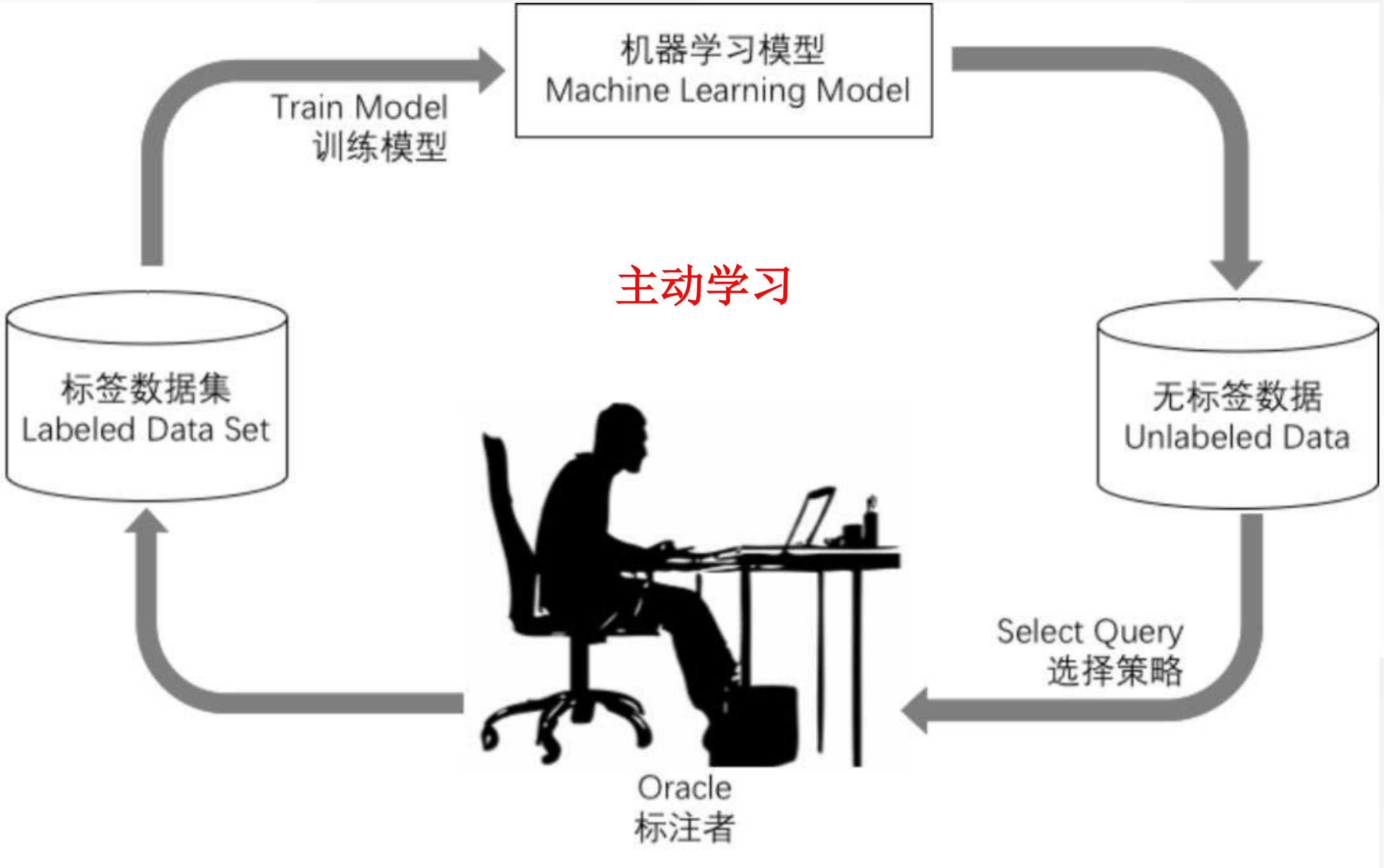


我们如何利用无标注数据?

Luis von Ahn: Games with a purpose (ReCaptcha)

ESP 游戏

我们如何利用无标注数据?



我们能否直接利用无标注数据学习出更好的模型?

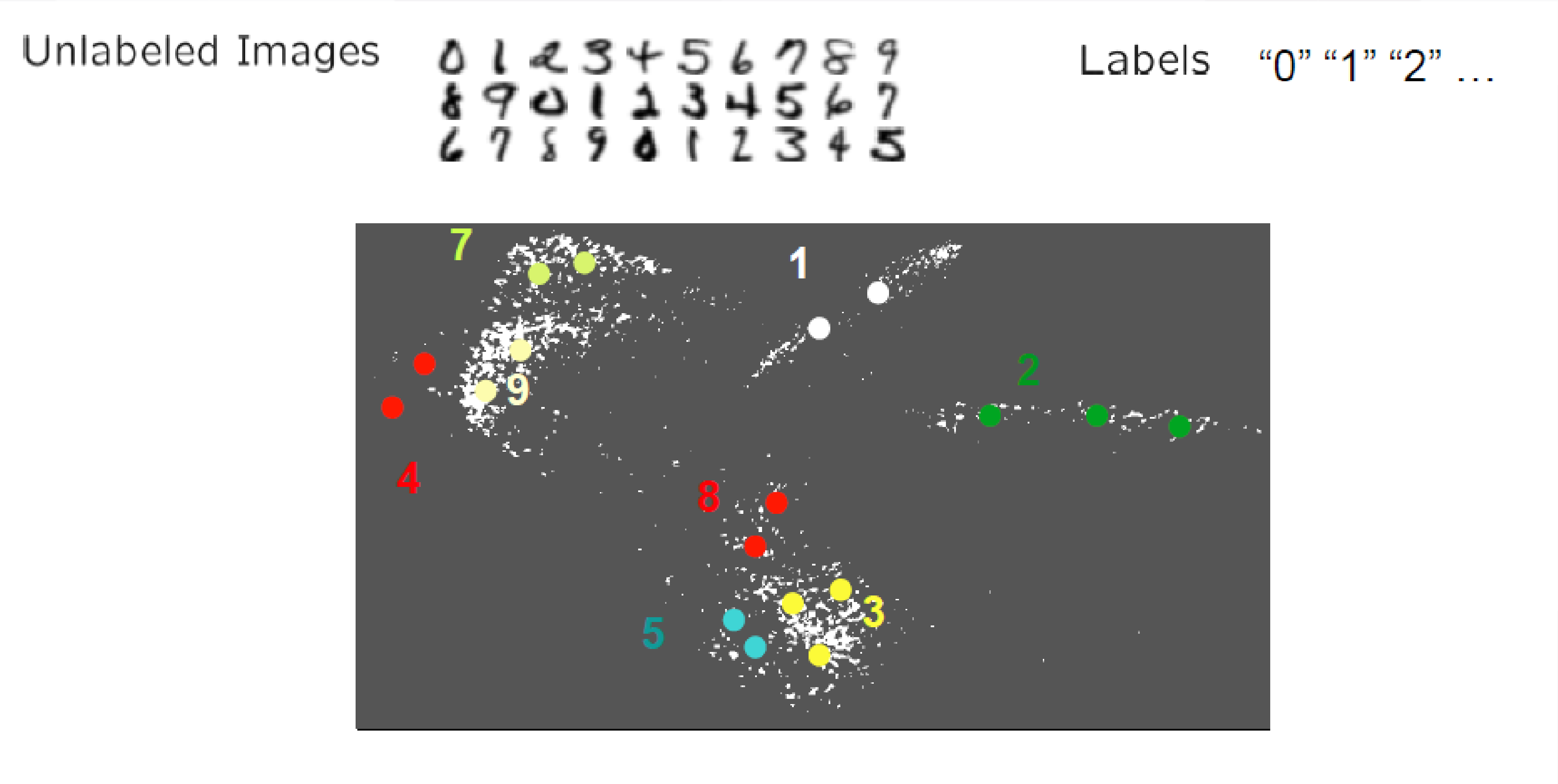
无标注数据能有帮助么?

- 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)

- 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

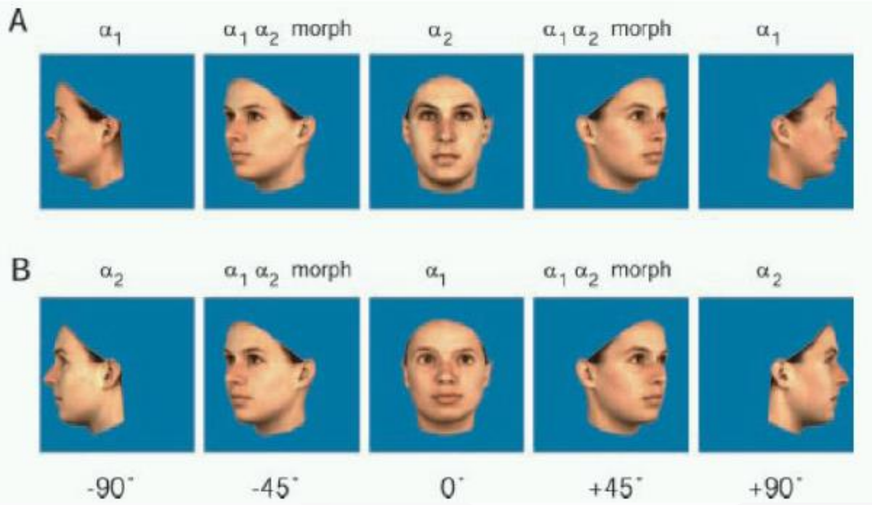
无标注数据能有帮助么?



■ “相似” 的数据点有”相似”的标签

自然界中的半监督学习

■ 人类就常常使用半监督学习



- 视觉识别中的时序关联
  - 人脸的两个角度是差别很大的，但是通过一个图像序列（无标注数据）就能把他们关联起来
  - 同样的机理，也可以让人造成误判

- 婴儿单词物体映射
  - 17个月的婴儿听单词，看物体
  - 如果这个单词听了很多遍，再看到物体，关联能力很强
  - 如果从没听过，关联能力很弱

半监督学习

- 半监督学习：让学习器不依赖外界交互、自动地利用无标注数据提升学习性能
- 半监督分类/回归
  - **给定:** 标注数据 $D = \{(x_1, y_1), (x_2, y_2) \dots (x_l, y_l)\}$，无标注数据 $D_u = \{x_{l+1}, x_{l+2}, \dots x_{l+u}\}$(通常 $u \gg l$)
  - 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 $\{x_i\}^m$, 目的是聚类或者降维。另外给出: 对数据的一些限制
    - 例如, 对于聚类: 两个点必须在一个簇, 或两个点一定不能在一个簇; 对于降维: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)：开放世界
  - 给定训练数据$D=\{(x_1,y_1),(x_2,y_2)\dots(x_l,y_l)\}$, 无标注数据 $D_u=\{x_{l+1},x_{l+2},\dots x_{l+u}\}$(通常$u\gg l$)
  - 学习函数f用于预测测试数据的标签
  - 学习到的函数f能够被应用到未见过的测试数据上
- 直推学习(Transductive learning)：封闭世界
  - 给定训练数据 $D=\{(x_1,y_1),(x_2,y_2)\dots(x_l,y_l)\}$, 无标注数据$D_u=\{x_{l+1},x_{l+2},\dots x_{l+u}\}$
  - 可以没有显式地学习函数，我们所关心的就是在D_u上的预测
  - 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) $\{(x_{1:n}, y_{1:n}), x_{test}\}$

$\updownarrow$

归纳学习 分类/回归$\{(x_{1:l}, y_{1:l}), x_{l+1:n}, x_{test}\}$

直推学习 分类/回归 $\{\{(x_{1:l}, y_{1:l}), x_{l+1:n}\}$

$\updownarrow$

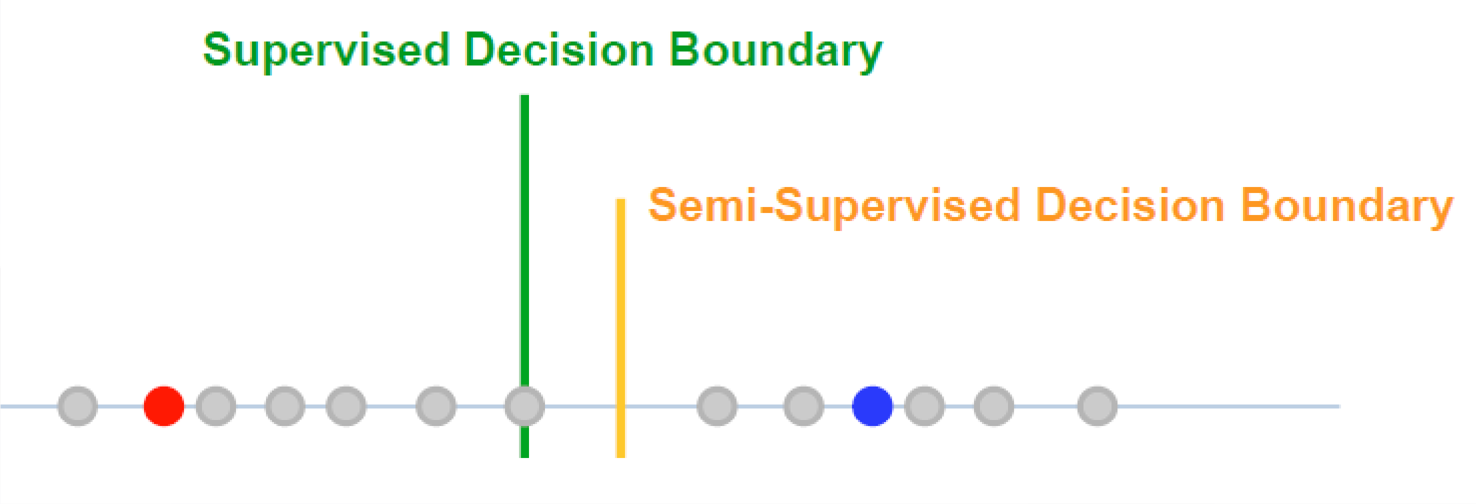
半监督聚类 $\{x_{1:n}, must-, cannot-links\}$

$\updownarrow$

无监督学习 (clustering) $\{x_{1:n}\}$

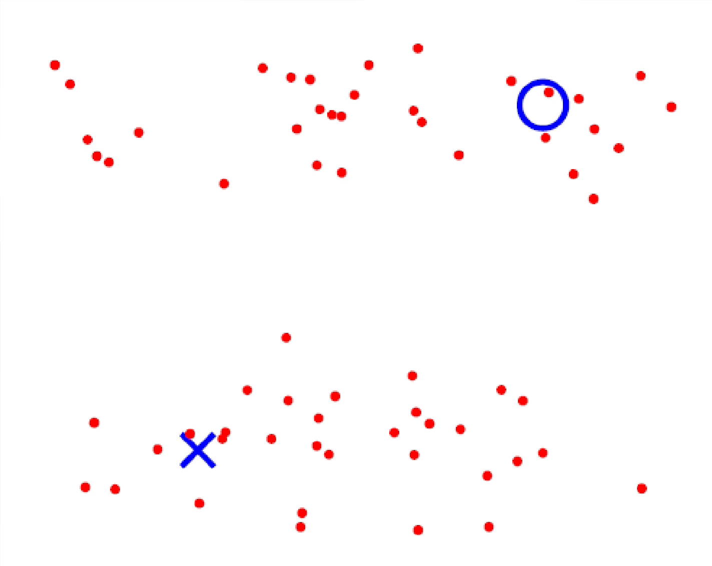
平滑假设（smoothness assumption）

- 半监督学习能有效，必须满足一些假设
- 半监督平滑假设:
  - 如果高密度空间中两个点 $x^{(1)}, x^{(2)}$ 距离较近, 那么对应的输出$y^{(1)}, y^{(2)}$也应该接近
- 监督学习的平滑假设 (用于对比):
  - 如果空间中两个点$x^{(1)}, x^{(2)}$距离较近, 那么对应的输出$y^{(1)}, y^{(2)}$也应该接近



聚类假设（cluster assumption）

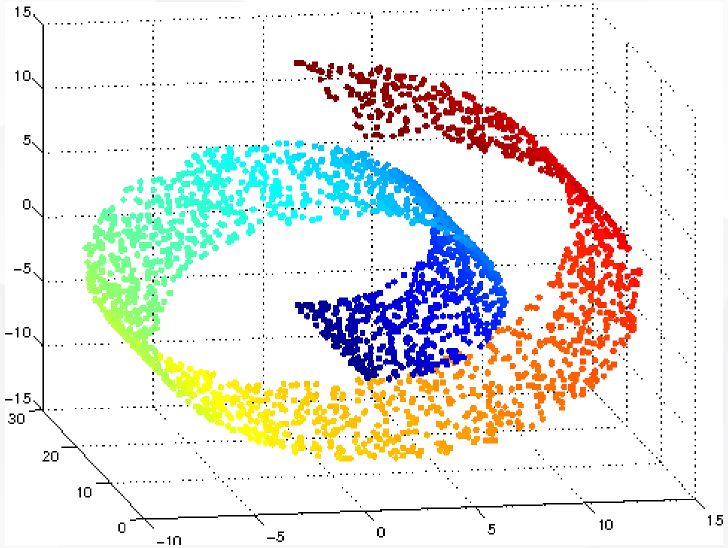
- 聚类假设
  - 如果点在同一个簇，那么它们很有可能属于同一个类
- 聚类假设的等价公式:
  - 低密度分隔:决策边界应该在低密度区域.
- 聚类假设可以被看作半监督平滑假设的一种特殊情形



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
- 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



主要的半监督学习模型

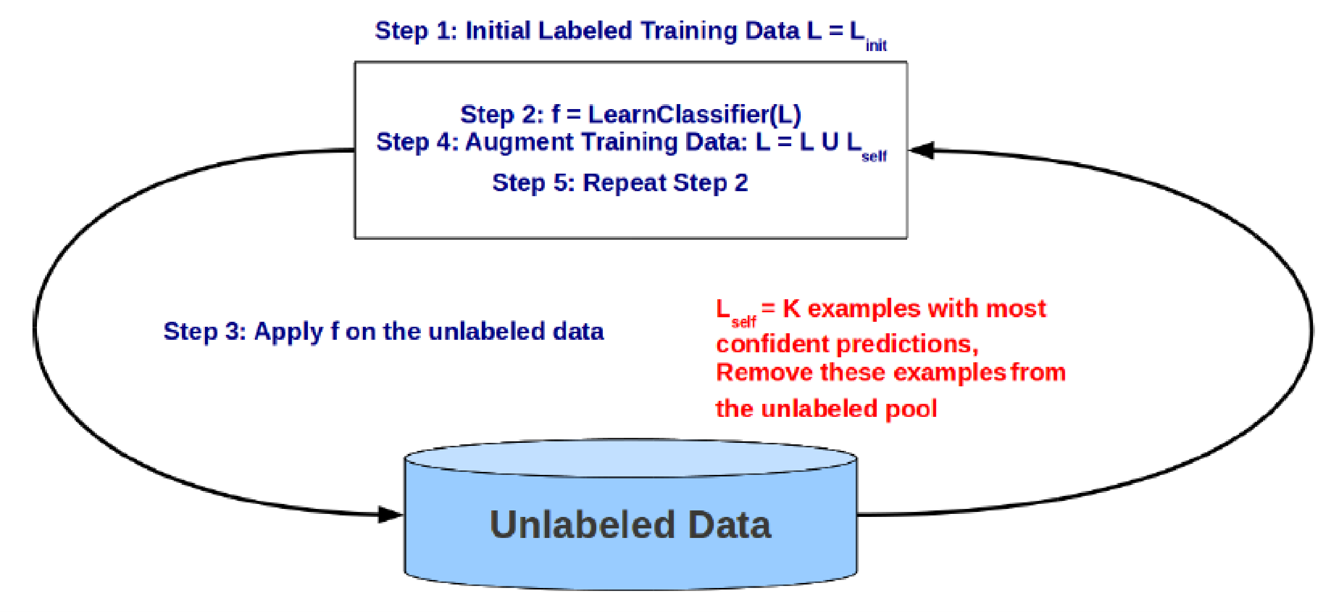
- 自学习
- 多视角学习
  - eg. 联合学习 [BM98]
- 生成模型
  - 数据采样自相同的生成模型.
  - eg. 混合高斯
- 低密度分割模型
  - 例如. Transductive SVM [Joa99]
- 基于图的算法
  - 数据被表示成图中的节点，边代表节点对的距离
  - 这些方法基于流形假设.
  - 例如. 标签传播
- 半监督聚类

大纲

- 简介
- 半监督学习算法
  - 自学习
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类
- 前沿进展

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(X_l, Y_l)学习f
  - 对$x \epsilon X_u$预测结果
  - 把$(x, f(x))$ 加入到标注数据
  - 重复上述过程

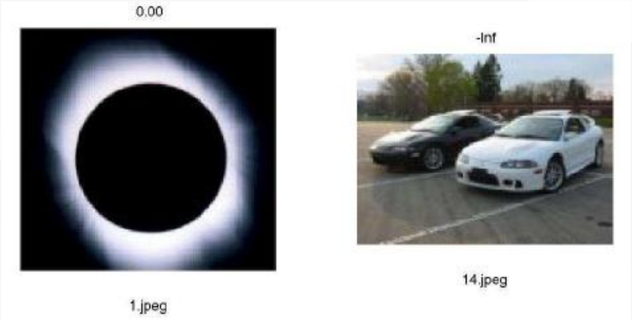


自学习的变体

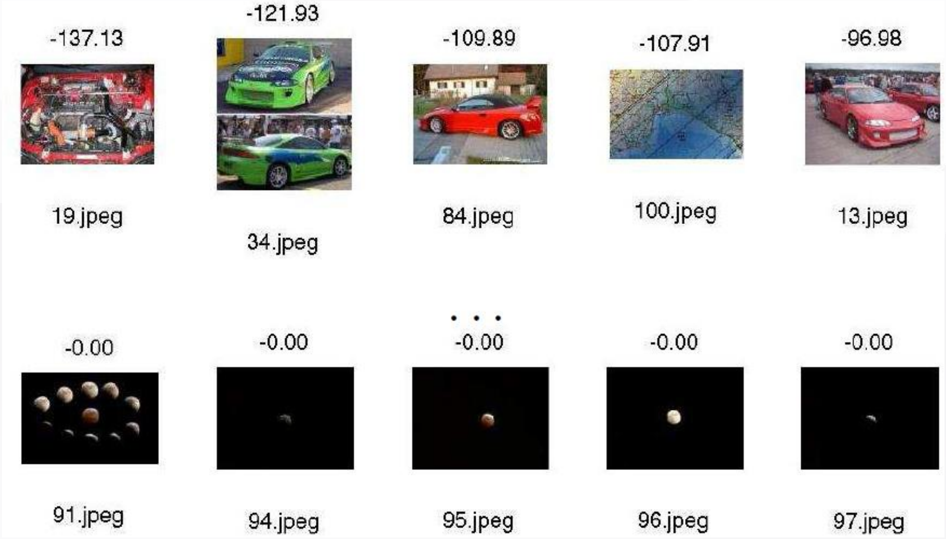
- 加入一些最置信的 $(x, f(x))$到标注数据集
  - 置信度的估计依据基分类器决定
  - 例如：朴素贝叶斯可将后验概率转化为分类置信度，支持向量机可将间隔大小转化为分类置信度
- 把所有 $(x, f(x))$ 加到标注数据
- 把所有$(x, f(x))$ 加到标注数据, 为每条数据按置信度赋予权重

自学习的例子: 图像分类

- 在两个初始图像上训练朴素贝叶斯分类器

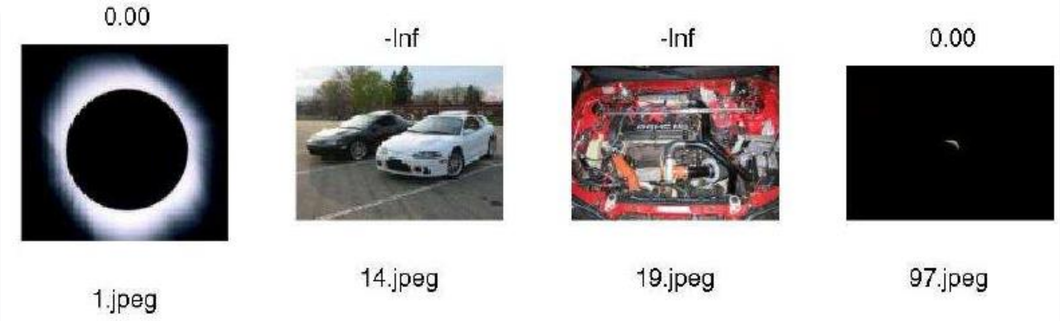


- 对无标记的数据分类, 根据置信度 $\log p(y = astronomy|x)$排序

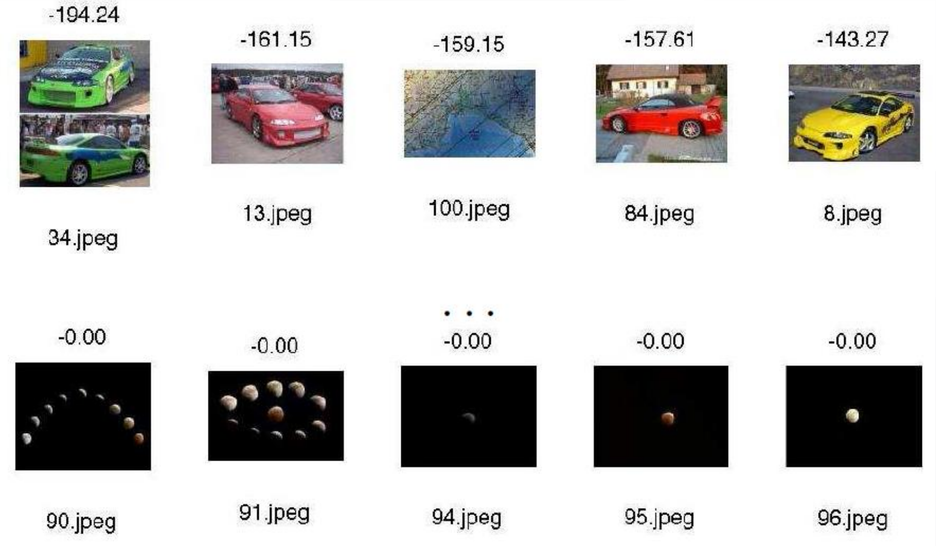


自学习的例子: 图像分类

- 将最置信的图像及其预测标签加入到标注数据

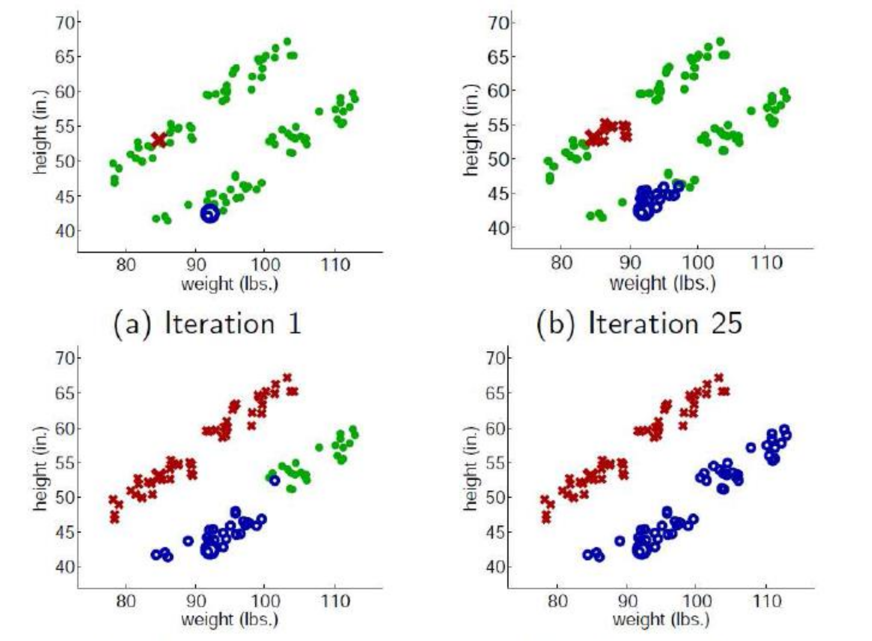


- 重新训练分类器，重复上述过程



自学习的优势

- 最简单的半监督学习方法
- 这是一种wrapper方法,可以应用到已有的（复杂）分类器上
- 经常被用到实际任务中，例如自然语言处理任务中

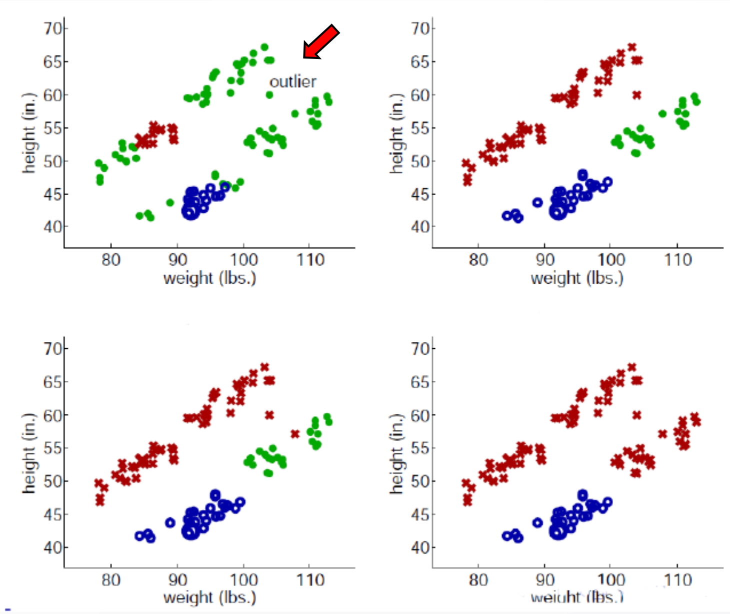


(a) Iteration 1 (b) Iteration 25

(c) Iteration 74 (d) Final labeling of all instances

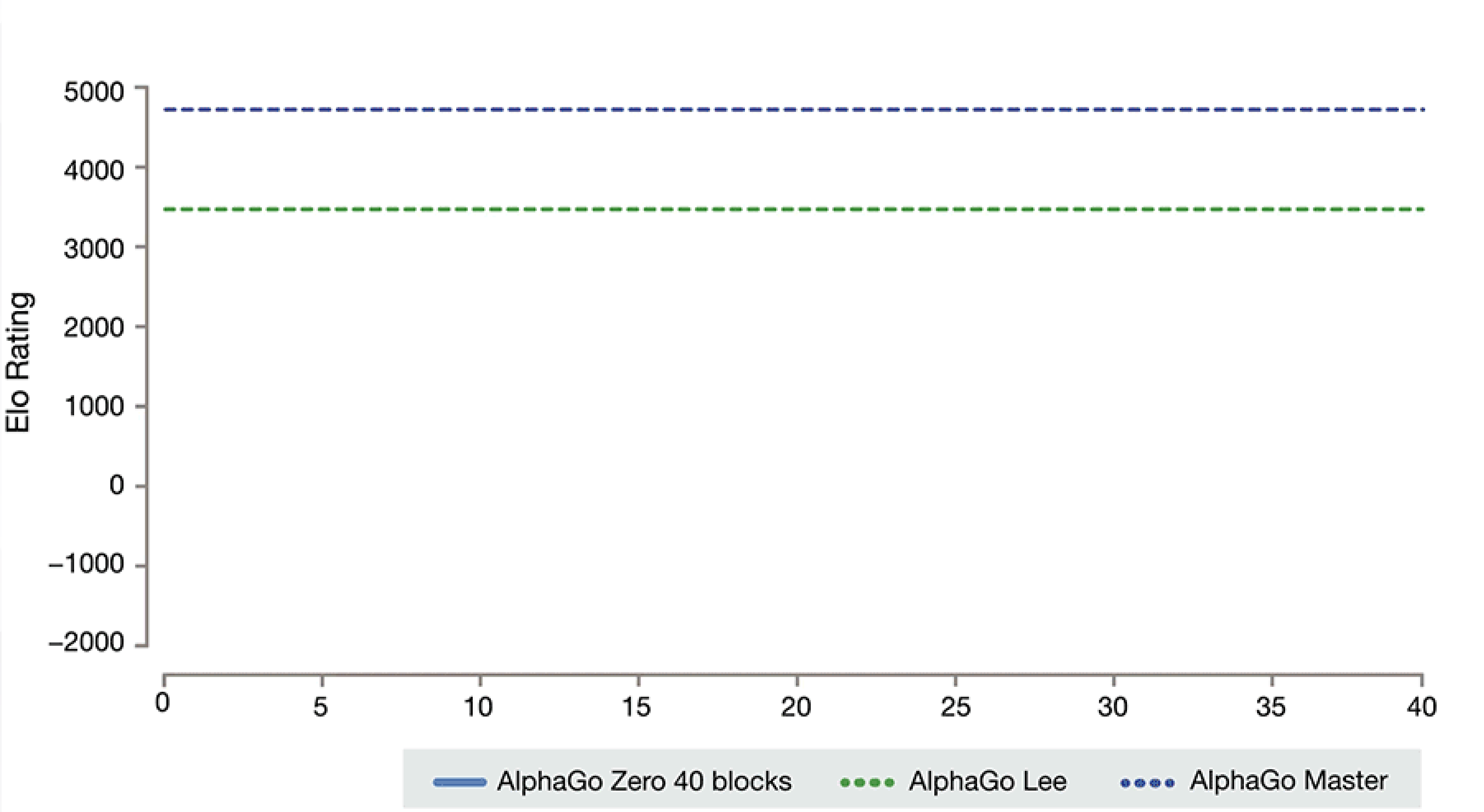
一个好的例子
基学习器: KNN

自学习的劣势



一个坏的例子
基学习器: KNN

知名的自学习例子（AlphaGo Zero）

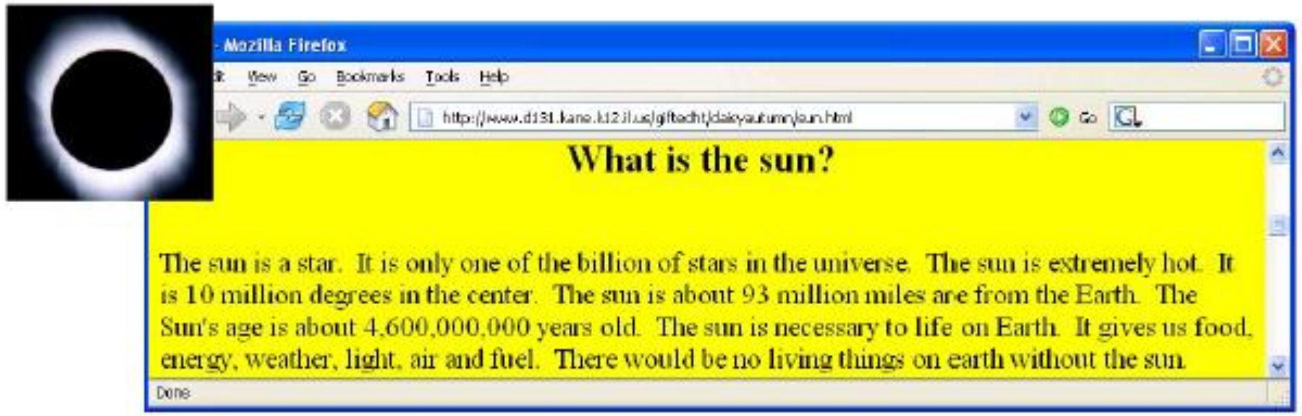


大纲

- 简介
- 半监督学习
  - 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
- 前沿进展

协同训练（co-training）

- 一个对象的两个视角: 图像和HTML文本

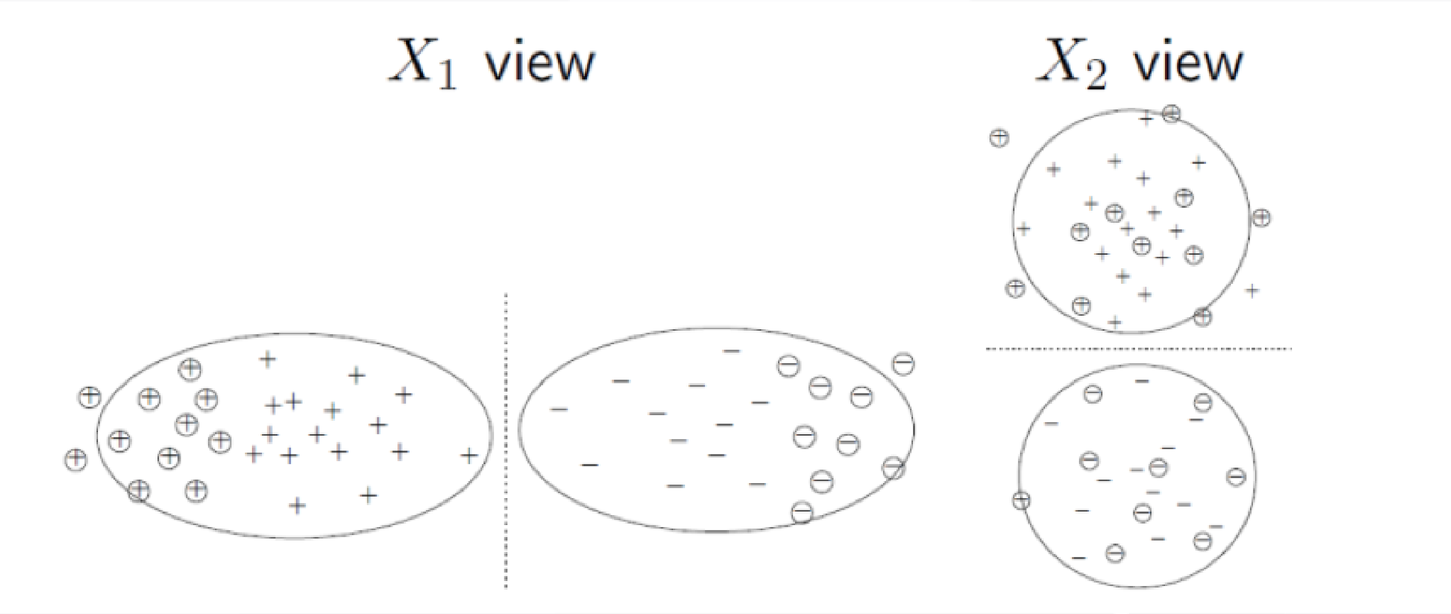


特征分裂

- 每个实例由两个特征集合$x = [x^{(1)}; x^{(2)}]$表示
  - $x^{(1)}$ = 图像特征
  - $x^{(2)}$ = 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: 数据拥有两个充分且条件独立的视图（相容互补性）
  - 特征集合可分裂$x = [x^{(1)}; x^{(2)}]$
  - 充分：$x^{(1)}$ 或 $x^{(2)}$单独对于训练一个最优分类器是充分的
  - 条件独立：$x^{(1)}$ 和 $x^{(2)}$ 在给定类别后是条件独立的



协同训练

- 协同训练
  - 训练两个分类器: 从$\left(X_l^{(1)}, Y_l\right)$学习$f^{(1)}$, 从$(X_l^{(2)}, Y_l)$学习$f^{(2)}$.
  - 用$f^{(1)}$ and $f^{(2)}$分别对X_u分类.
  - 把$f^{(1)}$ 的k个最置信的预测结果 $(x, f^{(1)}(x))$ 当做 $f^{(2)}$ 的标注数据
  - 把$f^{(2)}$ 的k个最置信的预测结果 $(x, f^{(2)}(x))$ 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

协同训练的变体

- Co-EM: 不止是top k, 全部预测数据当做标注数据
  - 每个分类器有一个概率标签X_u
  - 每个(x, y) 加上权重$P(y|x)$
- 假的特征集分裂
  - 构造随机的、人工的特征分裂
  - 再应用协同训练

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)
  - 基于分歧的方法disagreement-based methods
  - 协同训练是多视角学习中一个特例
- 通用的想法:
  - 训练多个分类器，每个分类器使用不同的视角
  - 多个分类器在无标签数据上应该达成一致

一个正则化风险最小化框架，鼓励多个学习器的一致性:

$$\min_f \sum_{v=1}^{M} \left(\sum_{i=1}^{l} c(y_i, f_v(x_i)) + \lambda_1 \|f\|_K^2 \right) + \lambda_2 \sum_{u,v=1}^{M} \sum_{i=l+1}^{n} (f_u(x_i) - f_v(x_i))^2$$

M 个学习器. C() 是原来的损失函数, 例如：铰链损失(hinge loss)、平方损失

多视角学习

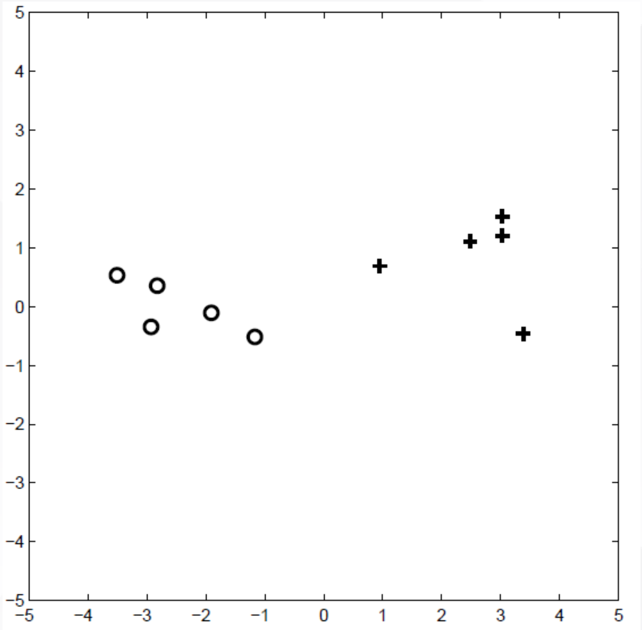
- 为什么多视角学习能学得更好?
  - 学习过程实质上搜索最好的分类器
  - 通过强迫多个分类器的预测一致性, 我们减少了搜索空间
  - 希望在较少的训练数据能够找到最好的分类器
- 对于测试数据，多个分类器被结合
  - 例如, 投票, 共识等.
- 得到了一些理论结果的支持[Blum and Mitchell, 1998, 周志华 2013]
- 基于多视角的半监督学习是半监督学习和集成学习的自然过渡
  - 使用不同的学习算法，使用不同的数据采样，甚至使用不同的参数设置，来产生不同的学习器
  - 无需数据拥有多视图，仅需弱学习器之间有显著的分歧，不同视图、不同算法、不同数据采样、不同参数设置等，都是产生差异的渠道，而非必备条件

大纲

- 简介
- 半监督学习
  - 自我训练
  - 多视角学习
  - 生成式模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类
- 前沿进展

生成式模型的例子

- 带标签的数据(X_l, 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 $\theta = \{\alpha_1, \alpha_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$

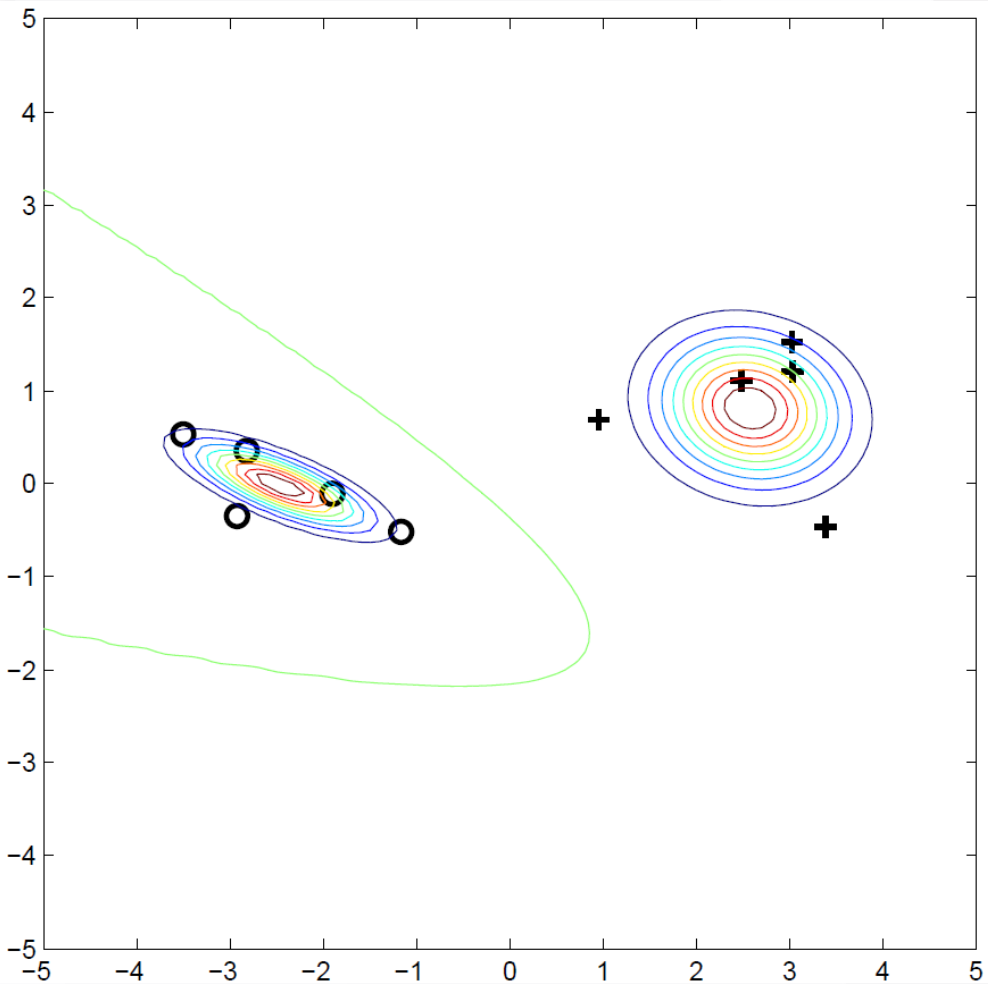
高斯混合模型:

$$p(x, y|\theta) = p(y|\theta)p(x|y, \theta) = \alpha_y \mathcal{N}(x; \mu_y, \Sigma_y)$$

分类: $P(y|x, \theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)} \underset{<}{>} 1/2$

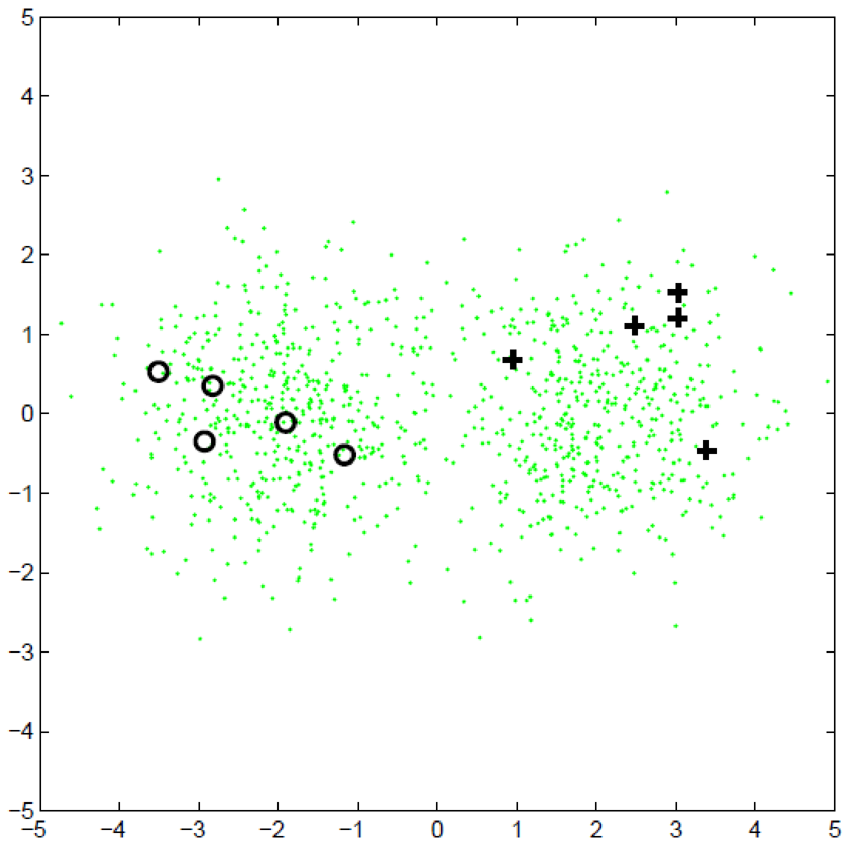
生成模型的例子

最可能的模型和它的决策边界



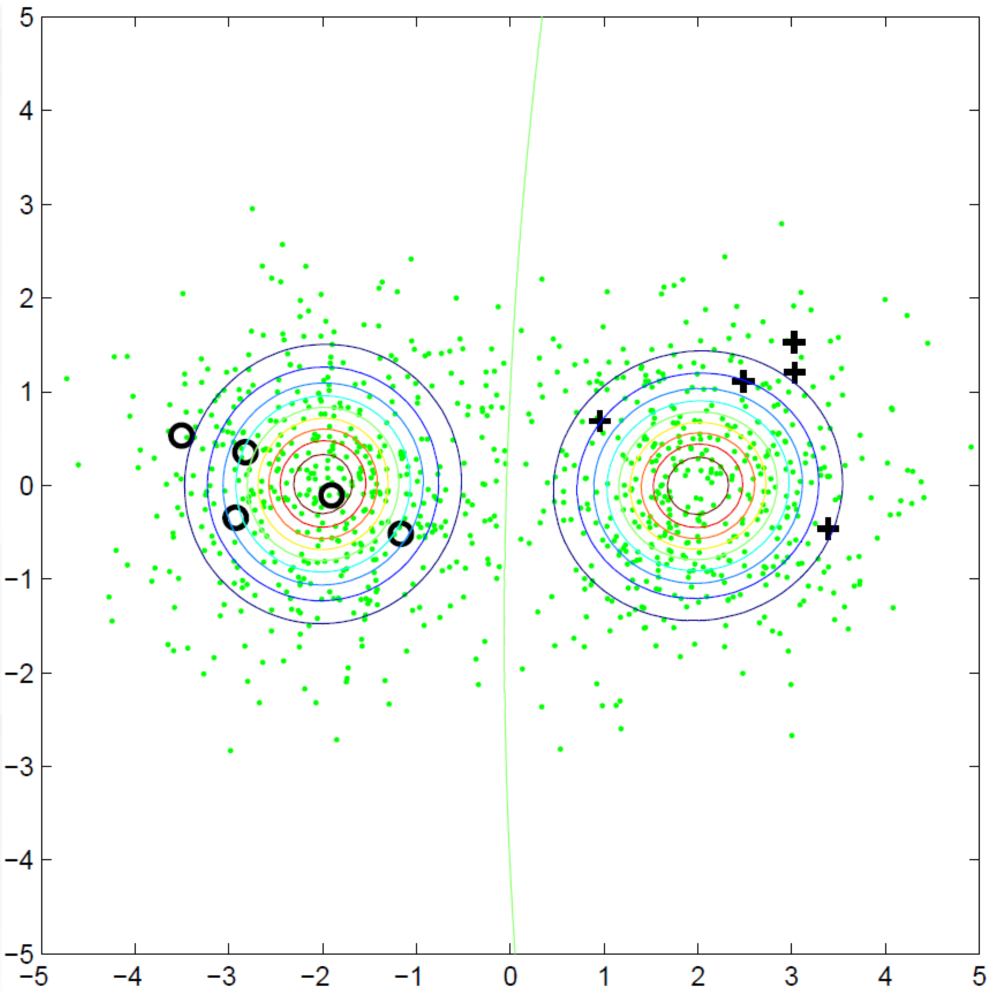
生成模型的例子

加入无标签数据



生成模型的例子

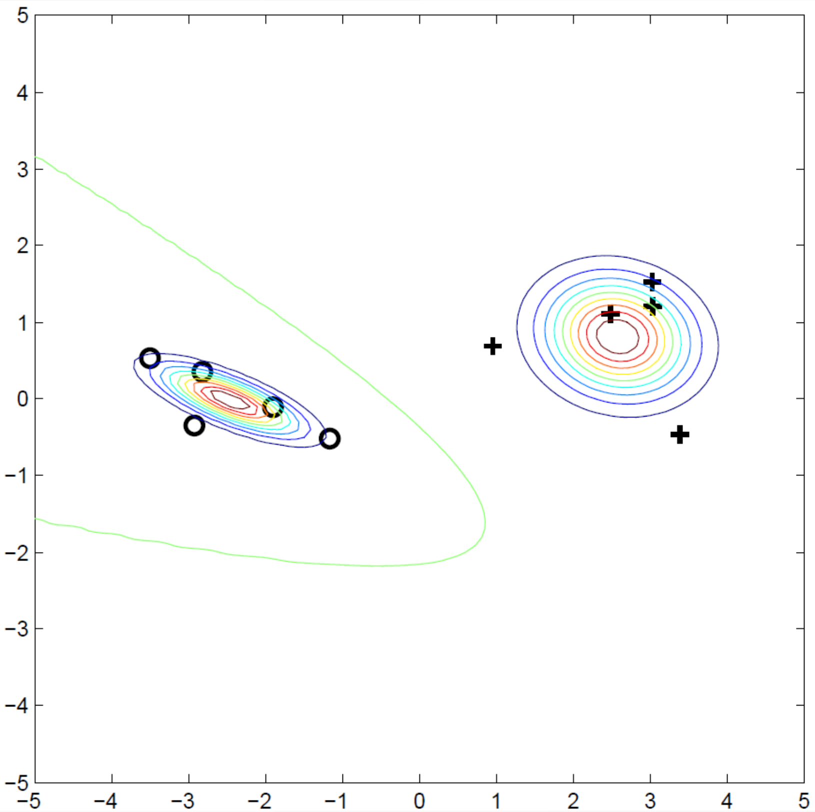
加入无标签数据, 最可能的模型和它的决策边界



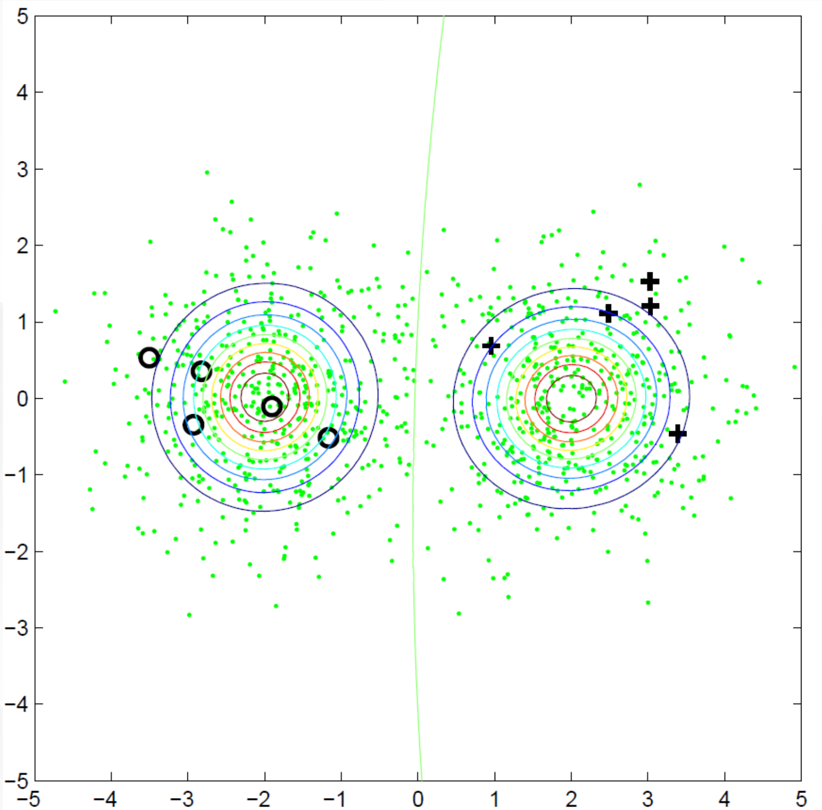
生成模型的例子

决策边界的不同，是由于模型最大化的目标不同

$p(X_l, Y_l|\theta)$



$p(X_l, Y_l, X_u|\theta)$



生成模型用于半监督学习

- 生成模型假设
  - 完全的生成模型（只考虑有标注数据） $P(X,Y|\theta)$
- 半监督学习生成模型:
  - 所有数据（无论标注与否）都是由同一个潜在的模型“生成”的
  - 我们所感兴趣的量: $p(X_l,Y_l,X_u|\theta)=\sum_{Y_u}p(X_l,Y_l,X_u,Y_u|\theta)$
  - 寻找θ的极大似然估计，或者最大后验估计（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)
  - 图像分类
  - EM算法
- 混合多项式分布 (朴素贝叶斯)
  - 文本归类
  - EM算法
- 隐马尔科夫模型 (HMM)
  - 语音识别
  - Baum-Velch 算法

案例分析: 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
  - $\log p(X_l, Y_l|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta) p(x_i|y_i, \theta)$
  - 利用MLE 计算 θ (频率, 采样均值, 采样协方差)
- 同时考虑有标注和无标注数据

$$\log p(X_l, Y_l, X_u|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta) p(x_i|y_i, \theta)$$

$$+\sum_{i=l+1}^{l+u} \log \sum_{y=1}^{2} p(y|\theta) p(x_i|y, \theta)$$

  - 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. 在(X_l, Y_l)上用MLE估计 $\theta = \{\alpha, \mu, \sum\}_{1:2}$
   - α_c=类别 c的比例
   - μ_c=类别c采样的均值
   - $\sum_c$=类别c采样的协方差

重复：

2. E步:对所有$x\epsilon X_u$，计算类别的期望 $p(y|x, \theta)$
   - 将x以$p(y = 1|x, \theta)$ 的比例标记为类别1
   - 将x以$p(y = 2|x, \theta)$ 的比例标记为类别2
3. M步: 用有标签数据X_l和预测标签的数据X_u MLE估计θ

可以被看作自训练的一种特殊形式

EM算法用于高斯混合模型

初始化: $\theta = \{\alpha, \mu, \Sigma\}_{1:2}$

E步: $$\gamma_{ji} = \frac{p(y|\theta)p(x|y,\theta)}{\sum_{y'} p(y'|\theta)p(x|y',\theta)} = \frac{\alpha_i \cdot p(x_j|\mu_i, \Sigma_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(x_j|\mu_i, \Sigma_i)}$$

第j个样本属于第i个类别的概率

M步:

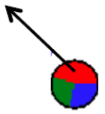
$$\alpha_i = \frac{\sum_{x_j \in D_u} \gamma_{ji} + l_i}{m}$$

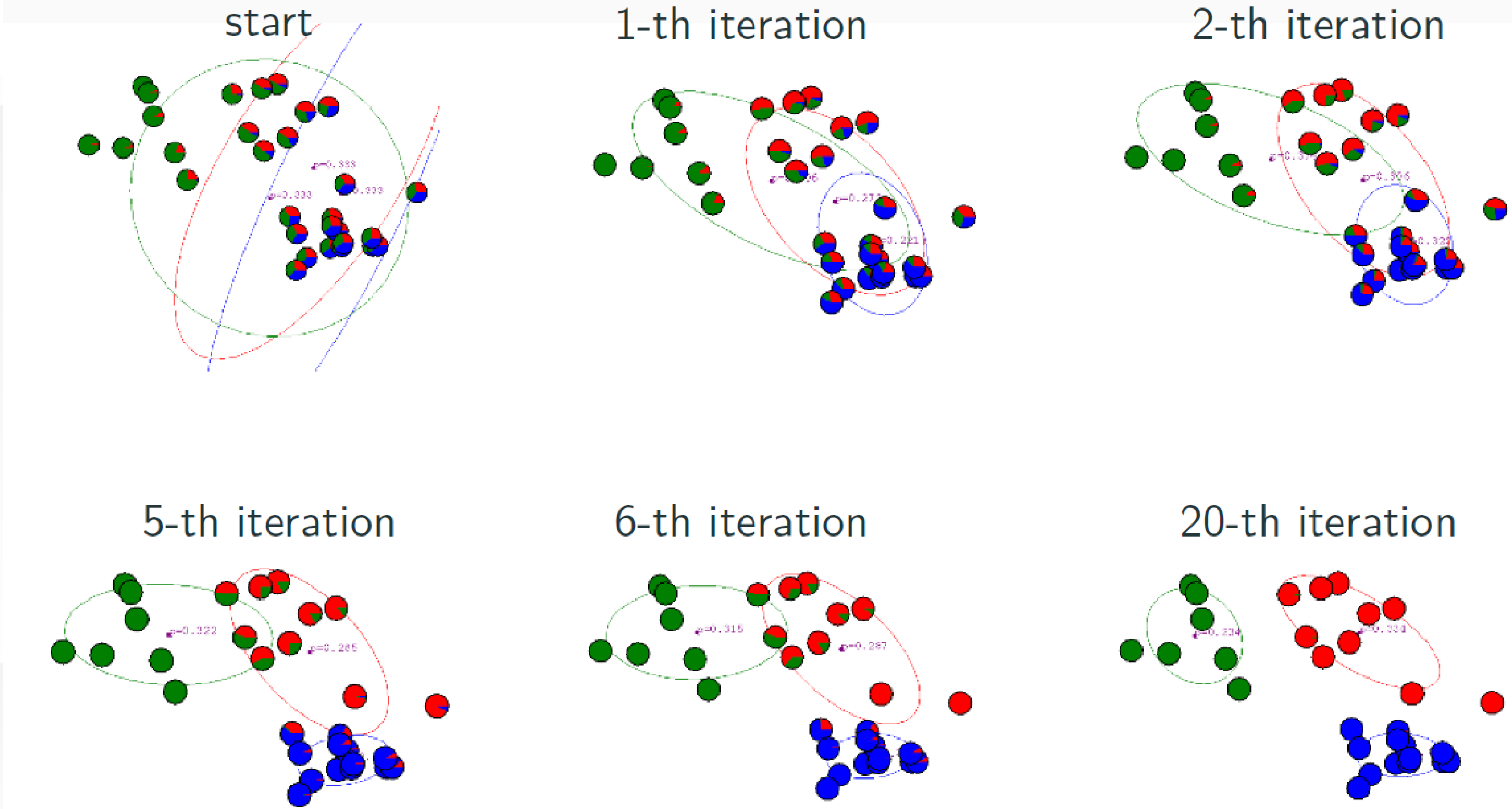
$$\mu_i = \frac{1}{\sum_{x_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{x_j \in D_u} \gamma_{ji} x_j + \sum_{(x_j, y_j) \in D_l \wedge y_j = i} x_j \right)$$

$$\Sigma_i = \frac{1}{\sum_{x_j \in D_u} \gamma_{ji} + l_i} \sum_{x_j \in D_u} \gamma_{ji} (x_j - \mu_i)(x_j - \mu_i)^T + \frac{1}{\sum_{x_j \in D_u} \gamma_{ji} + l_i} \sum_{(x_j, y_j) \in D_l \wedge y_j = i} (x_j - \mu_i)(x_j - \mu_i)^T$$

EM

$P(y = ● \mid x_j, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3, w_1, w_2, w_3)$





生成模型用于半监督学习: 除了EM之外

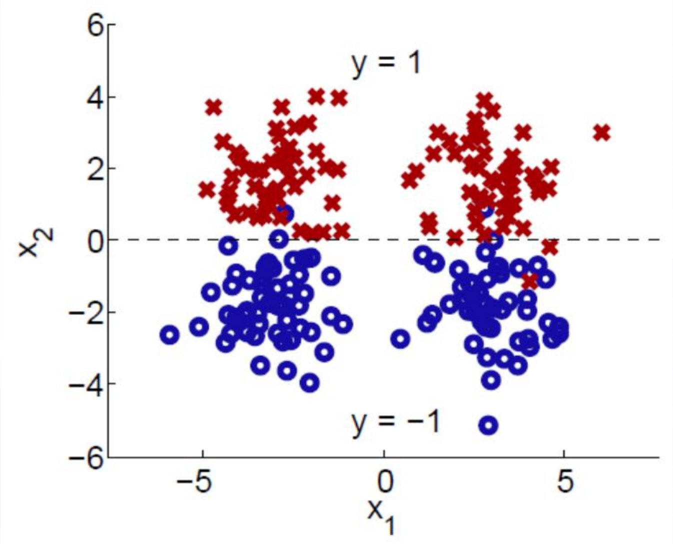
- 核心是最大化$p(X_l, Y_l, X_u|\theta)$
- EM 只是最大化该概率的一种方式
- 其他能计算出使其最大化参数的方法也是可行的，如，变分近似，或者直接优化

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

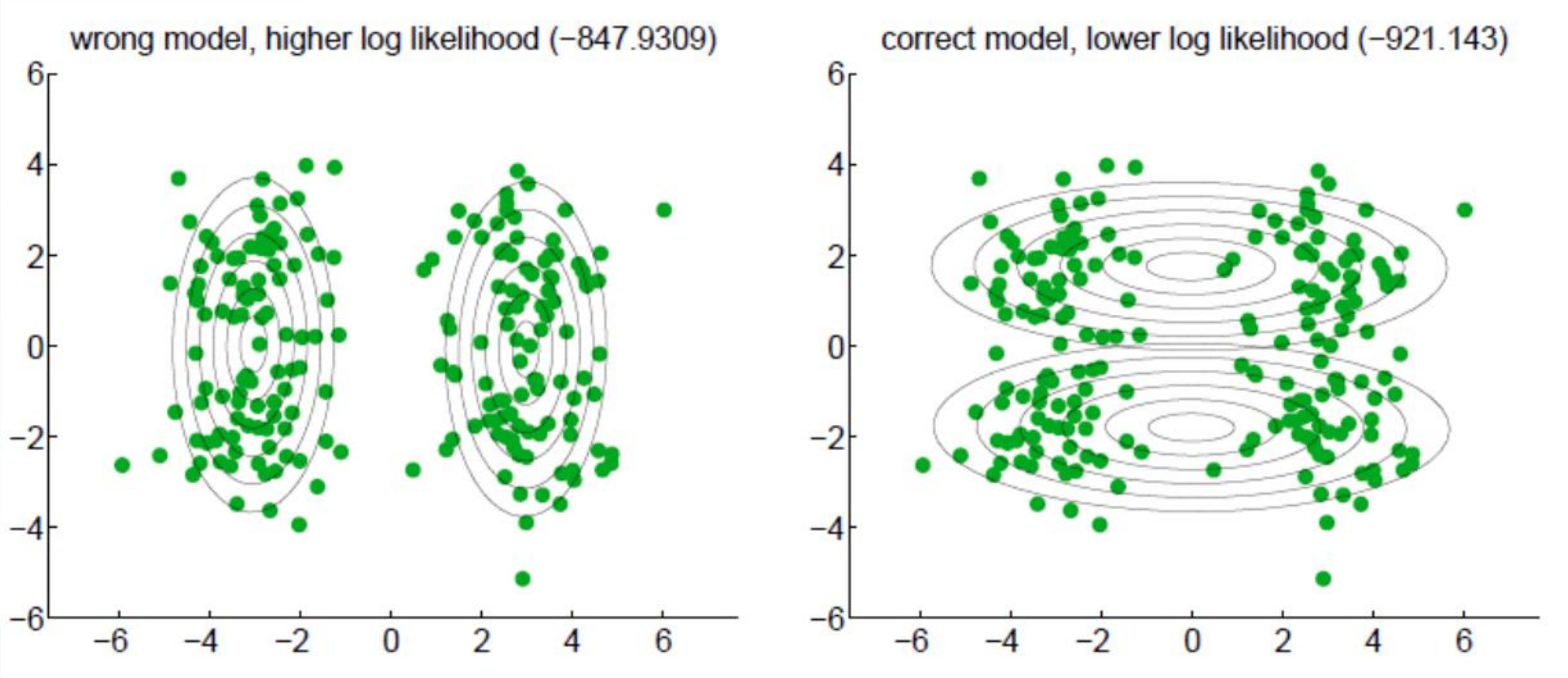
- 验证模型的正确性比较困难
- EM局部最优
- 如果生成模型是错误，无监督数据会加重错误



例如, 对文本进行题材分类

无标注数据可能伤害半监督学习

- 如果生成模型是错误的:



减少风险的启发式方法

- 需要我们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

$$\log p\,(X_l, Y_l|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i,\theta) + \lambda \sum_{i=l+1}^{l+u} \log \sum_{y=1}^{2} p(y|\theta)p(x_i|y_i,\theta)$$

相关方法: 聚类标签法（Cluster-and-label）

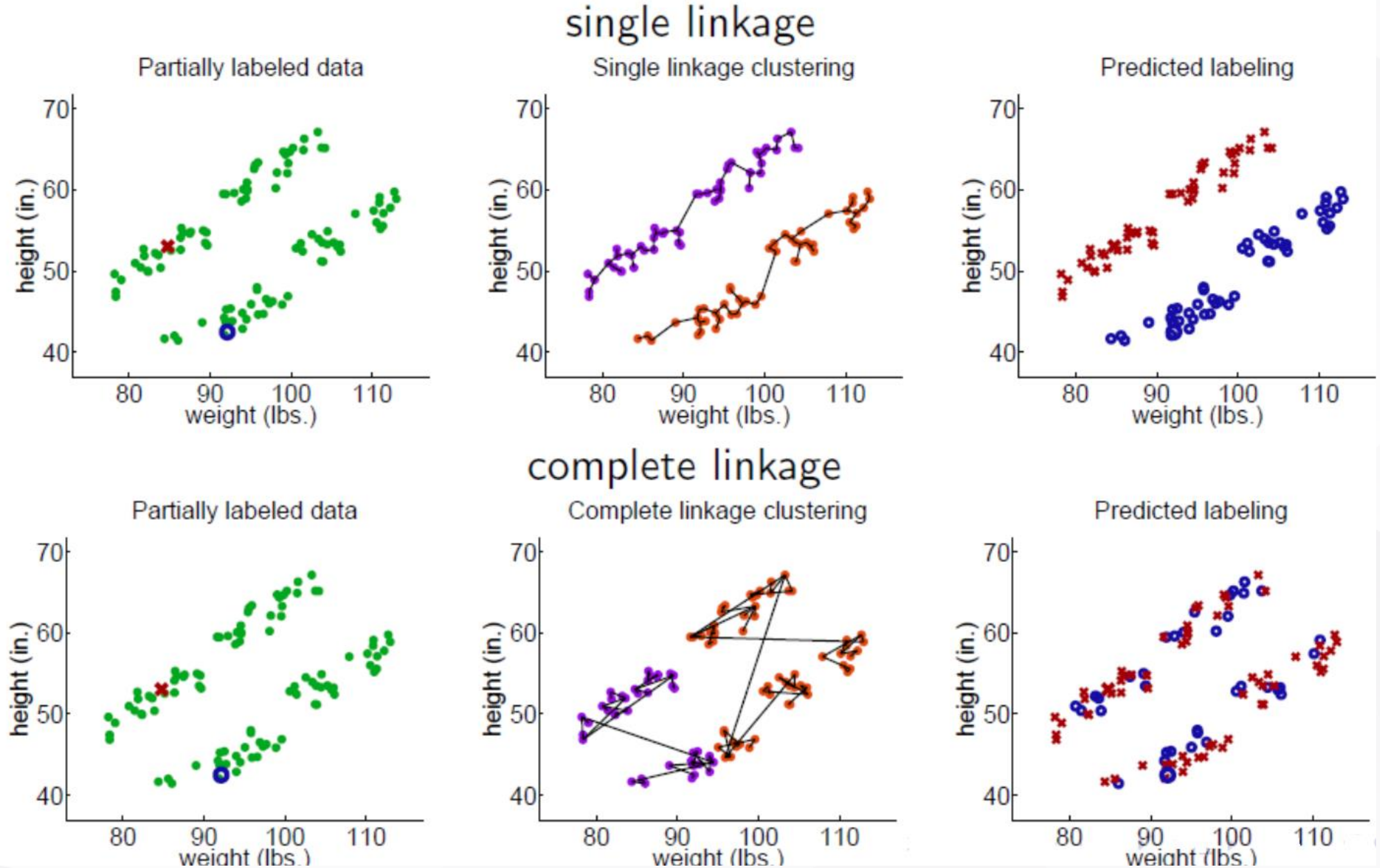
除了使用概率生成模型，任何聚类算法都可以被用于半监督学习:

- 在$X_1 \dots X_u$上运行某种聚类算法
- 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

簇和标签: 有效和无效的例子

Example: $\mathcal{A}$=Hierarchical Clustering, $\mathcal{L}$=majority vote.

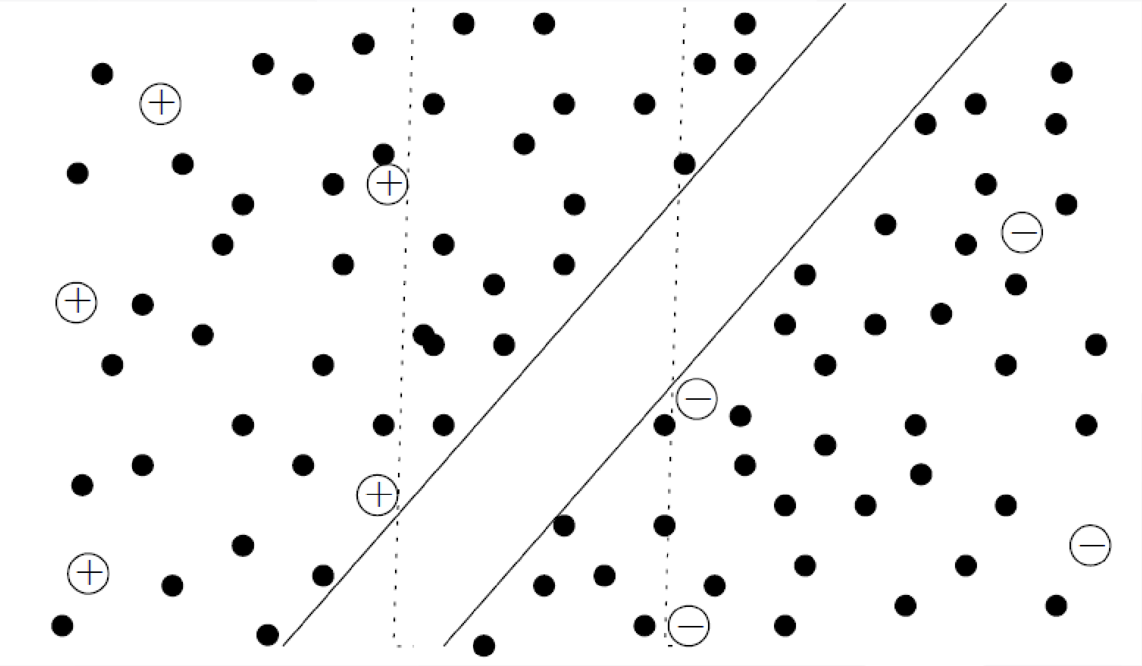


大纲

- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成式模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类
- 前沿进展

半监督支持向量机

- 半监督支持向量机(Semi-supervised SVMs, 简称S3VMs) = 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
  - 是什么假设?

- S3VMs 基本思想:
  - 遍历所有 2^u 种可能的标注X_u
  - 为每一种标注构建一个标准的SVM (包含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$\{X_l, Y_l\}$
  - 权重 w
- SVM 寻找一个函数 $f(x) = w^{\mathsf{T}}x + b$
- 通过$sign(f(x))$分类x

标准软间隔SVM

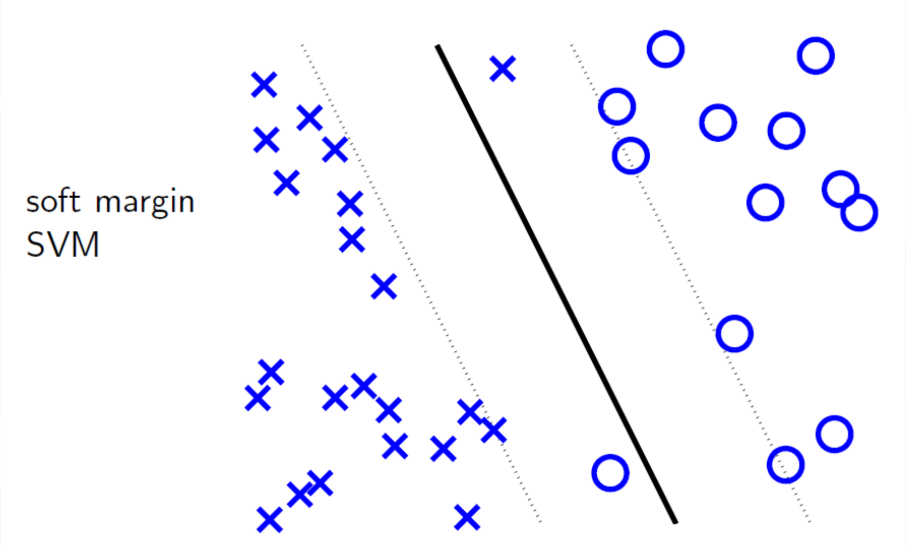
- 尝试去保持标注的点远离边界, 同时最大化间隔:

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

$$s.t. \quad y_i(w^\intercal x_i + b) \geq 1 - \xi_i, \forall i = 1 \dots l$$

$$\xi_i \geq 0$$

- ξ_i是松弛变量



Hinge 函数

$$\min_{\xi} \xi$$

$$\text{subject to } \xi \geq z$$

$$\xi \geq 0$$

如果 $z \leq 0, \min \xi = 0$

如果 $z > 0$, min $\xi = z$

因此带有限定条件的优化问题等价于hinge函数

$$(z)_+ = \max(z, 0)$$

使用hinge函数的SVM

令$z_i = 1 - y_i\left(w^\top x_i + b\right) = 1 - y_i f(x_i)$, 目标函数

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

$$\text{subject to } y_i(w^\top x_i + b) \geq 1 - \xi_i\,, \forall i = 1 \cdots l$$

$$\xi_i \geq 0$$

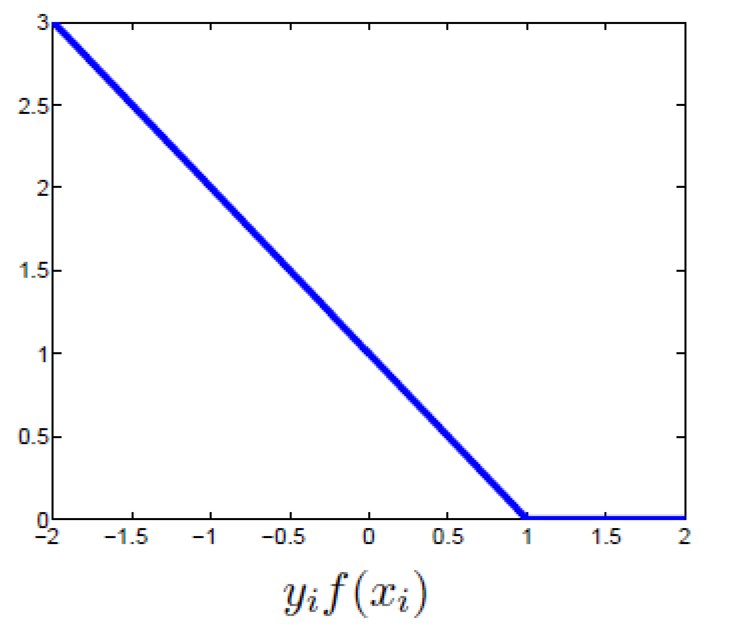
等价于

$$\min_{f} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} (1 - y_i f(x_i))_+$$

标准软间隔SVM中的hinge损失

$$\min_f \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(1 - y_i f(x_i))_+$$

$(1 - y_i f(x_i))_+$ hinge损失



倾向于让有标注的点在“正确”的一边

S3VM 目标函数

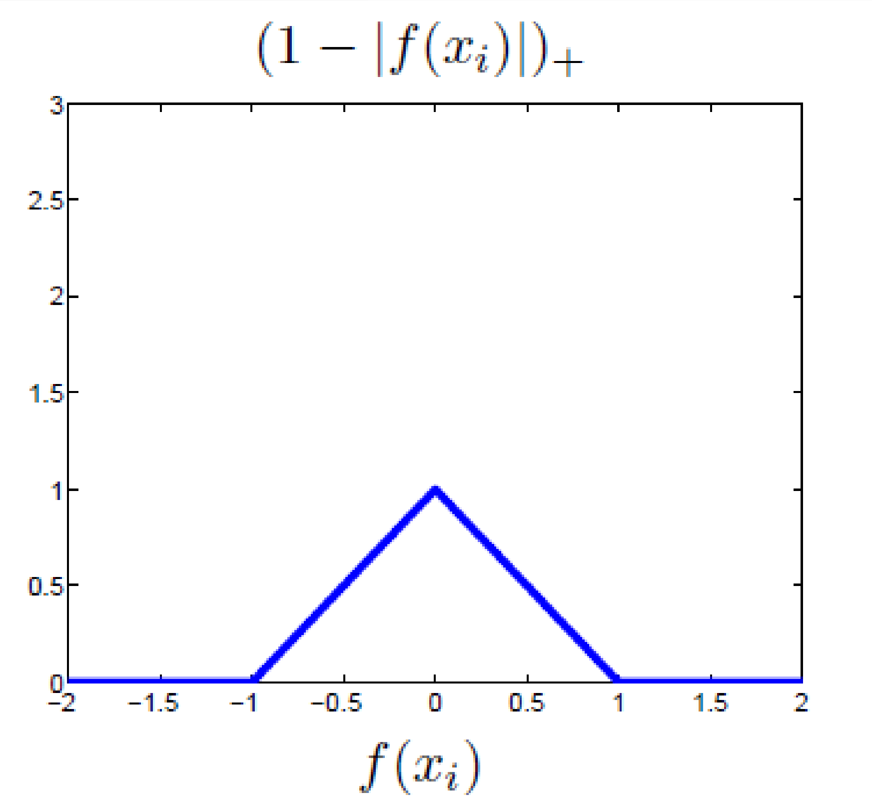
- 如何利用没有标注的点?
  - 分配标签sign($f(x)$)给 $x \in X_u$
  - sign($f(x)$) $f(x) = |f(x)|$
  - 无标注上的hinge损失为

    $(1 - y_i f(x_i))_+ = (1 - |f(x_i)|)_+$

- S3VM 目标函数:

$$\min_f \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + C_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

无标注数据上的帽形损失 (hat loss)



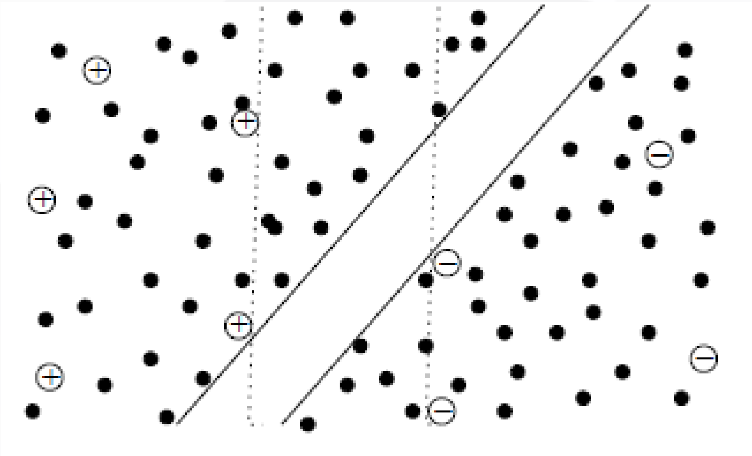
偏向$f(x) \geq 1$ or $f(x) \leq -1$, 即无标注数据远离决策边界$f(x) = 0$.

避免无标签数据落在间隔内

S3VM 目标函数

$$\min_{f} \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + C_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

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S3VM 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: $\frac{1}{n-l}\sum_{i=l+1}^{n} f(x_i) = \frac{1}{l}\sum_{i=1}^{l} y_i$

S3VM算法

- 输入: 权重w, C_1, C_2, (X_l, Y_l), X_u
- 求解优化问题求 $f(x) = w^\top x + b$

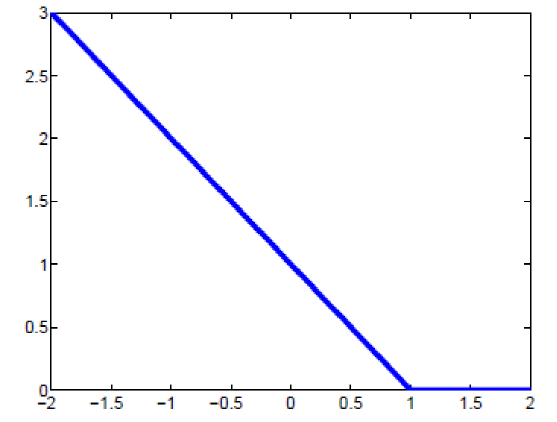
$$\min_f \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + C_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

$$\text{s.t.} \quad \frac{1}{n-l} \sum_{i=l+1}^{n} f(x_i) = \frac{1}{l} \sum_{i=1}^{l} y_i$$

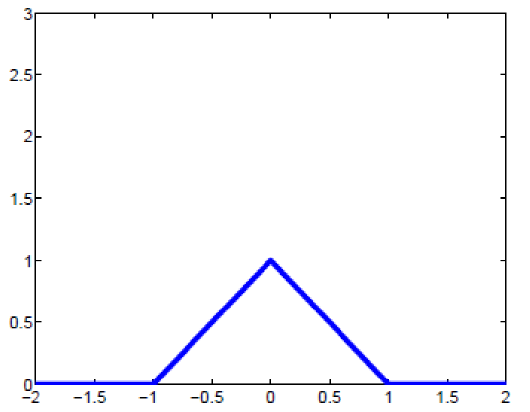
- 通过sign($f(x)$)分类新的测试点x

S3VM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S3VM研究主要关注的点

用于S3VM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(X_l, Y_l)训练一个SVM.
- 根据$f(X_u)$排序X_u . 以合适的比例标注$y = 1, -1$
- FOR $C_2 \leftarrow 10^{-5}C_2 \ldots C_2$
  - REPEAT:
  - $\min_f \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{l}(1 - y_i f(x_i)) + C_2 \sum_{i=l+1}^{n}(1 - |f(x_i)|)_+$
  - IF $\exists(i, j)$ 可交换 THEN 交换 y_i, y_j
  - UNTIL 没有标签可交换

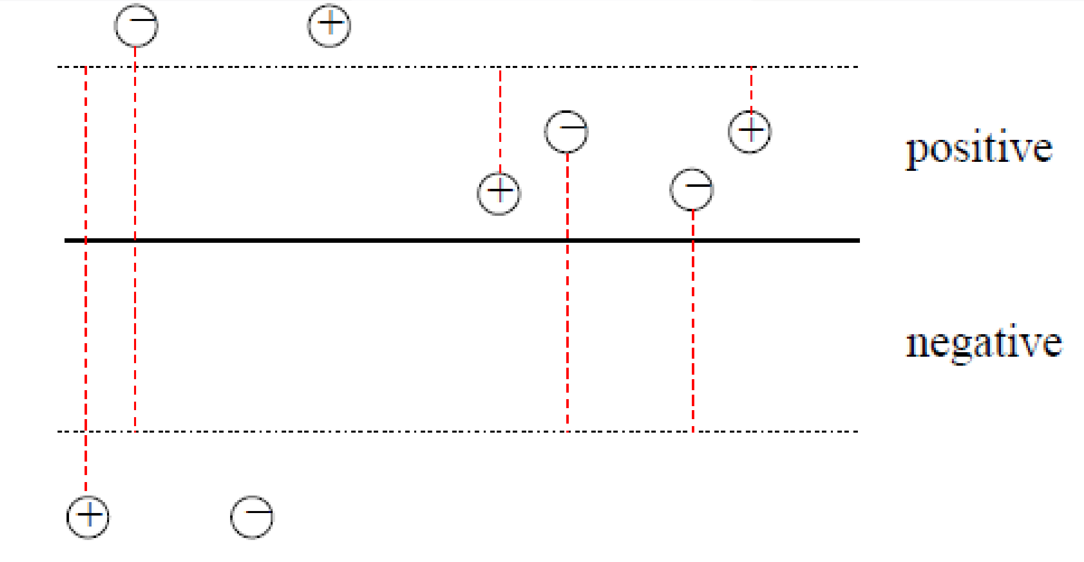
S3VM 实现1: SVM^{light}

$i,j \in X_u$ 可交换if $y_i = 1, y_j = -1$ and

$$loss(y_i = 1, f(x_i)) + loss\left(y_j = -1, f(x_j)\right)$$

$$> loss(y_i = -1, f(x_i)) + loss(y_j = 1, f(x_j))$$

Hinge损失 $loss(y, f) = (1 - yf)_+$



S3VM 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.
- 它使用AI中经典的分支定界搜索技术.
- 不足是它只能处理数百个无标注的点.

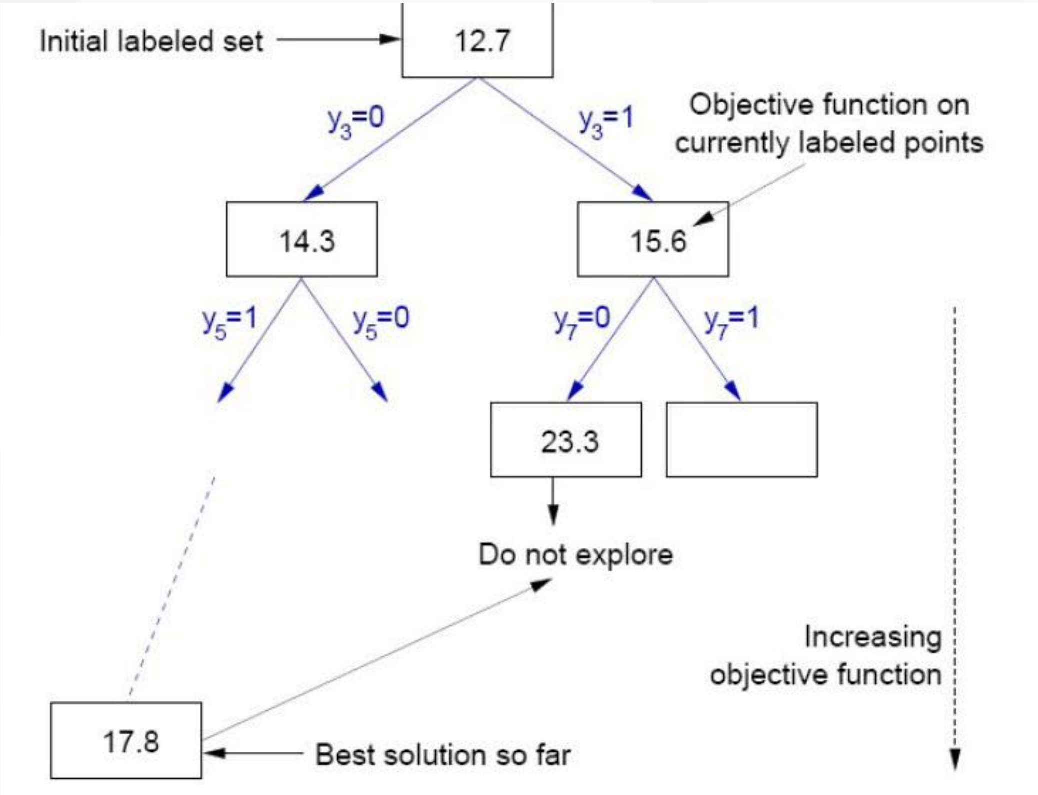
S3VM 实现2: 分支定界

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

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

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)



S3VMs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
  - 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
- 前沿进展

例子:文本分类

- 分类 天文学 vs. 旅行 文章
- 相似性是通过文档中词的重叠度度量的

	d_1	d_3	d_4	d_2
asteroid	●	●		
bright	●	●		
comet		●		
year				
zodiac				
⋮				
airport				
bike				
camp			●	
yellowstone			●	●
zion				●

No overlapping words!

	d_1	d_3	d_4	d_2
asteroid	●			
bright	●			
comet				
year				
zodiac		●		
⋮				
airport			●	
bike			●	
camp				
yellowstone				●
zion				●

当只有标注数据没法使用时

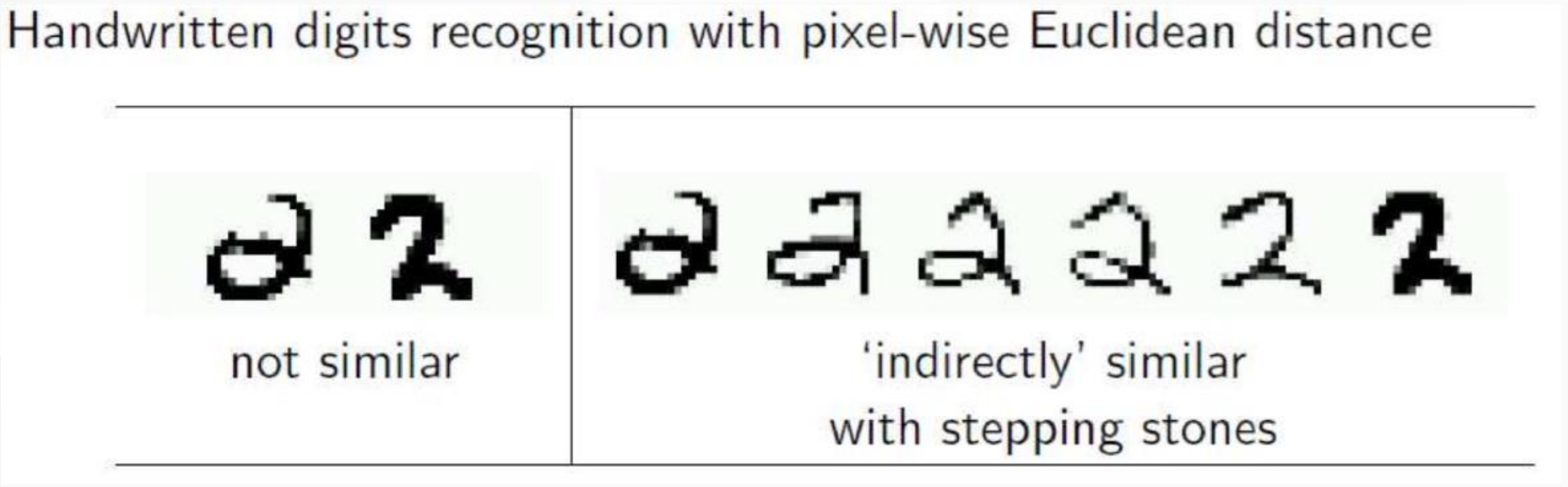
无监督数据作为跳板

标签通过相似的无标注文章进行传播

	d_1	d_5	d_6	d_7	d_3	d_4	d_8	d_9	d_2
asteroid	●								
bright	●	●							
comet		●	●						
year			●	●					
zodiac				●	●				
⋮									
airport						●			
bike						●	●		
camp							●	●	
yellowstone								●	●
zion									●

基于图的半监督学习

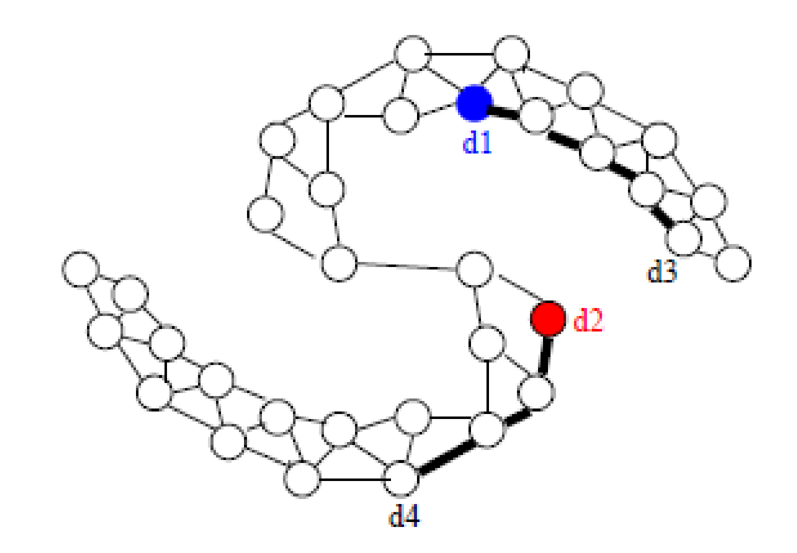
- 假设
  - 假定在有标注和无标注数据上存在一个图. 被“紧密”连接的点趋向于有相同的标签. （什么假设？）
- 在图上标签的变化应该是平滑的
  - 临近节点应该有相似的标签



我们称之为标签传播

图

- 节点: $X_l \cup X_u$
- 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$w = \exp(-\|x_i - x_j\|^2/\sigma^2)$
  - ε–半径(ε-radius) 图
- 想要的结果: 通过所有的路径来推导相似度



一些基于图的算法

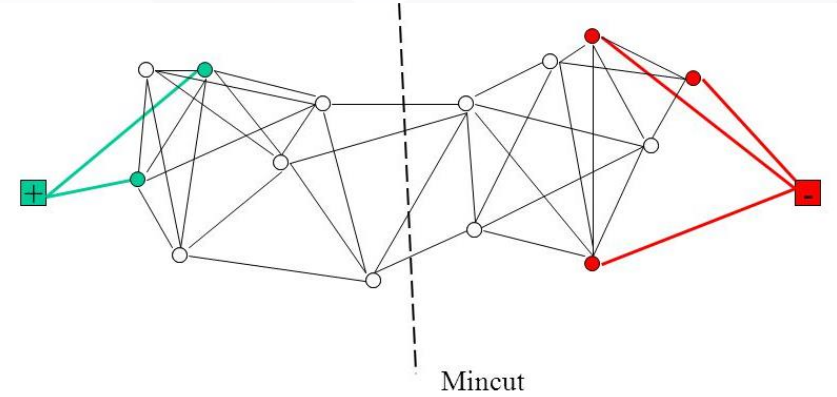
- 最小割（Mincut）
- 调和函数法(Harmonic)
- 局部和全局一致性法(Local and global consistency)
- 流形正则化方法(Manifold regularization)

最小割算法

- 图的最小割问题:
  - 固定Y_l, 寻找$Y_u \in \{0,1\}^{n-l}$ 使得最小化$\sum_{ij} w_{ij}|y_i - y_j|$
  - 等价地, 解决如下的优化问题

$$\min_f \infty \sum_{i=1}^{l} (y_i - Y_{li})^2 + \sum_{ij} w_{ij}(y_i - y_j)^2$$

  - 组合问题，但是有多项式时间的解



最小割算法

Karger‘s algorithm（随机算法）

While there are more than 2 vertices:

- Pick a remaining edge (u, v) uniformly at random
- Merge (or “contract”) u and v into a single vertex
- Remove self-loops

Return cut represented by final 2 vertices

Stoer–Wagner algorithm（确定性算法）

```
function: MinCutPhase(Graph G, Weights W, Vertex a):
        A <- {a}
        while A != V:
                add tightly connected vertex to A
        store cut_of_the_phase and shrink G by merging the two vertices added last

minimum = INF
function: MinCut(Graph G, Weights W, Vertex a):
        while |V| > 1:
                MinCutPhase(G,W,a)
                if cut_of_the_phase < minimum:
                        minimum = cut_of_the_phase
return minimum
```

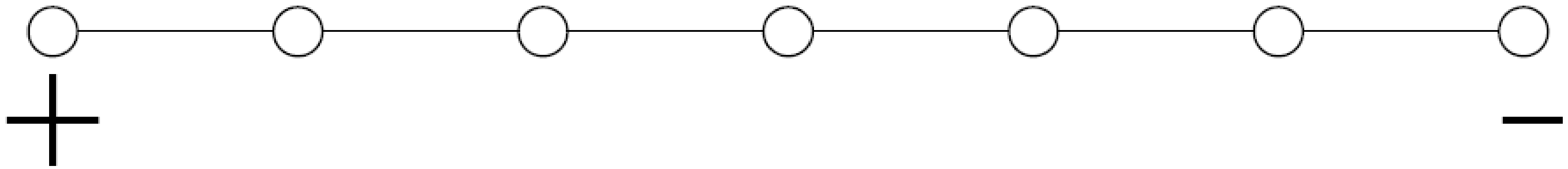
最小割算法

- 最小割计算了玻尔兹曼机的modes（峰值）

$$p_\beta(f) = \frac{1}{Z}\exp\left(-\beta E(f)\right)$$

$$E(f) = \sum_{ij} w_{ij}(f(i) - f(j))^2 \qquad f(i) \in \{0,1\}$$

- 可能存在多种模式
- 一个方法对权重做一些随机扰动，平均结果



半监督学习中的Randomized Mincut算法

- 构建一个图G
- 随机给边加上一些噪声，然后求解最小割
- 移除那些极度不平衡的解（小于5%的解在一边的）
- 用多数投票获得最后的分割

Avrim Blum et al. Semi-Supervised Learning Using Randomized Mincuts

调和函数法（Ten Year Best Paper Award, ICML'2013）

- 放松离散的标签值到连续值$\mathbb{R}$, 调和函数f满足
  - $f(x_i) = y_i \in \mathbb{R}$对于$i = 1 \cdots l$
  - f 最小化能量

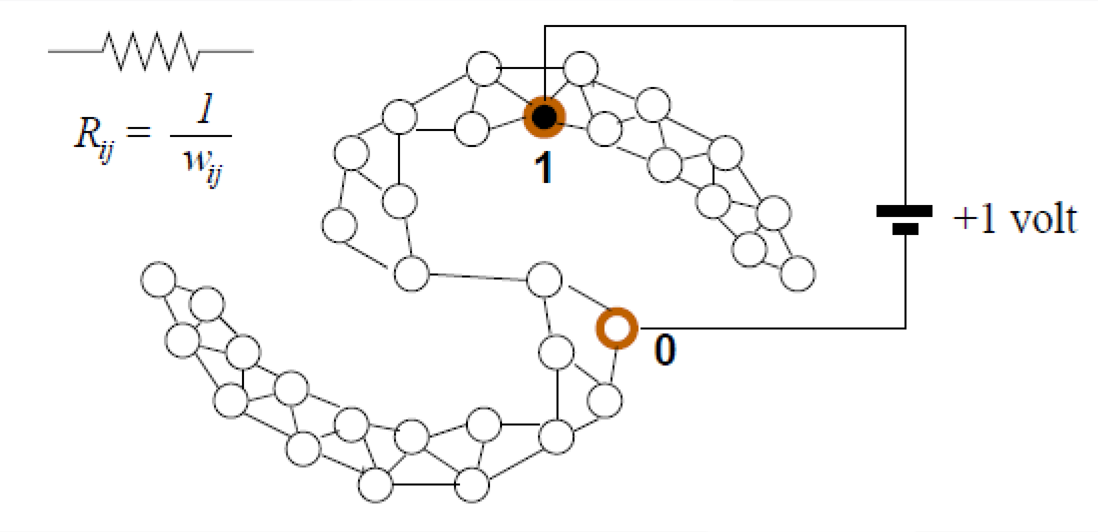
$$E(f) = \sum_{i \sim j} w_{ij}(f(x_i) - f(x_j))^2$$

  - 高斯随机场的均值
  - 邻居的均值 $f(x_i) = \frac{\sum_{j \sim i} w_{ij} f(x_j)}{\sum_{j \sim i} w_{ij}}, \forall x_i \in X_u$

电子网络的解释

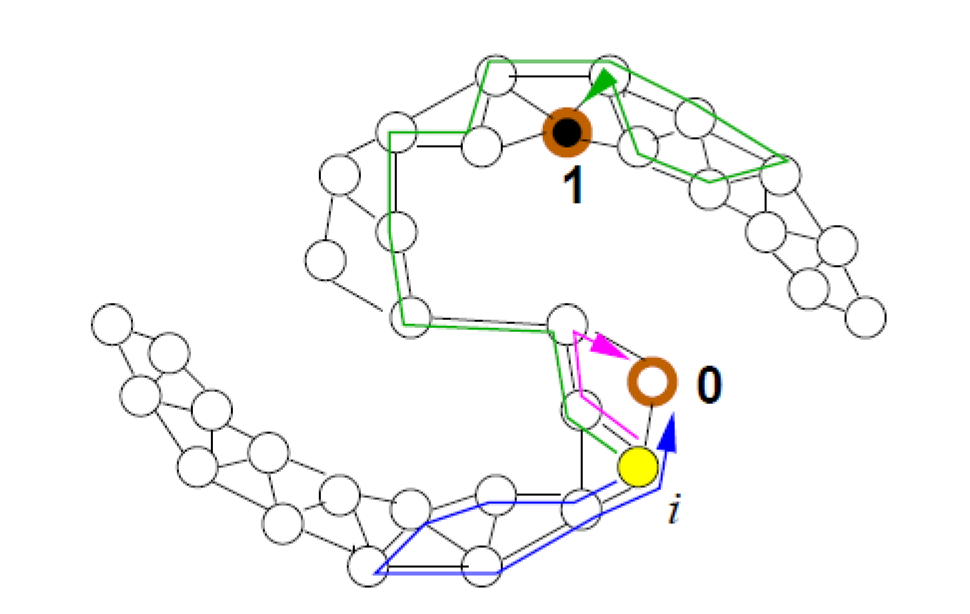
- 边看作是电导系数为w_{ij}的电阻
- 1 v电压被连接到有标注的点$y = 0,1$
- 节点上的电压是调和函数f

相似性推导: 如果两点之间有许多路径存在，那么电压相同



随机游走解释

- 从节点 i 以概率$\frac{w_{ij}}{\sum_k w_{ik}}$随机游走到$j$
- 如果遇到有标注节点就停止
- 调和函数$f = Pr$(遇到标签1|从节点 i出发)



计算调和函数的算法

- 计算调和函数的一种方式是:
  - 初始, 设置 $f(x_i) = y_i$ 对于 $i = 1 \cdots l$, 对于$x_j \in X_u$ $f(x_j)$ 设为任意值 (例如, 0)
  - 重复这个步骤直到收敛: Set $f(x_i) = \frac{\sum_{j\sim i} w_{ij} f(x_j)}{\sum_{j\sim i} w_{ij}}, \forall x_i \in X_u$, 即邻接点的加权平均值. 注意 $f(X_l)$ 是固定的
- 这也可以看成是自学习的一种特殊形式.

图拉普拉斯

我们也使用图拉普拉斯(graph Laplacian)计算f的闭式解

- 在 $X_l \cup X_u$上$n \times n$ 权重矩阵 W
  - 对称, 非负
- 对角的度矩阵(Diagonal degree matrix)D: $D_{ii} = \sum_{j=1}^{n} W_{ij}$
- 图拉普拉斯矩阵Δ

$$\Delta = D - W$$

- 能量函数可以重写为

$$\frac{1}{2}\sum_{i \sim j} w_{ij}\left(f(x_i) - f\left(x_j\right)\right)^2 = f^{\top}\Delta f$$

利用拉普拉斯计算调和解

调和函数解最小化给定标注情况下的能量

$$\min_f \infty \sum_{i=1}^{l} (f(x_i) - y_i)^2 + f^\top \Delta f$$

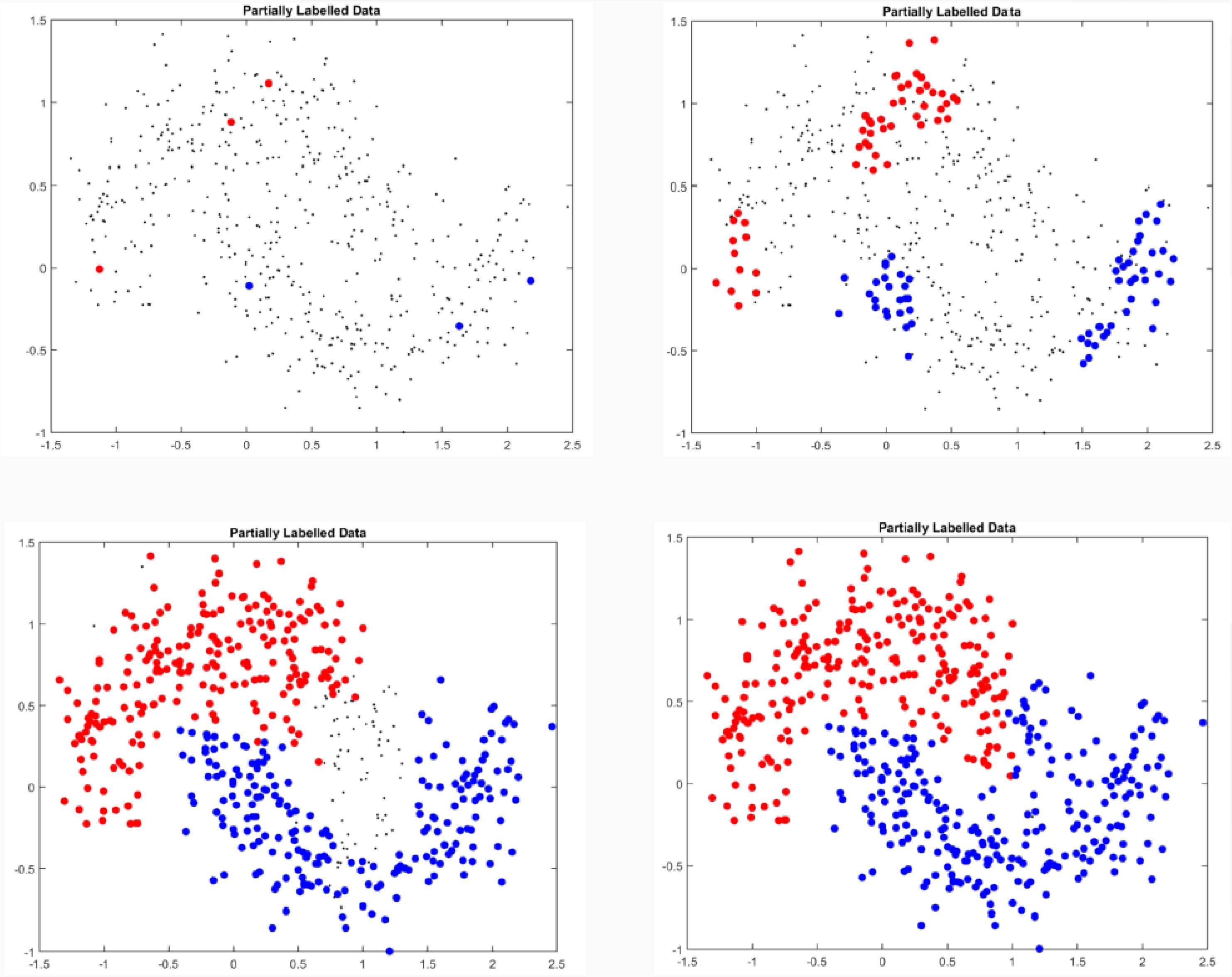
拉普拉斯矩阵分割 $\Delta = \begin{bmatrix} \Delta_{ll} & \Delta_{lu} \\ \Delta_{ul} & \Delta_{uu} \end{bmatrix}$

调和解 Δf =0

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} Y_l = (I - P_{uu})^{-1} P_{ul} Y_l \quad \text{其中} P = D^{-1} W$$

归一化拉普拉斯 $\mathcal{L} = D^{-1/2} \Delta D^{-1/2} = I - D^{-1/2} W D^{-1/2}$, 或$\Delta^p, \mathcal{L}^p$ 也经常使用$(p > 0)$.

标签传播


Partially Labelled Data
Partially Labelled Data
Partially Labelled Data
Partially Labelled Data
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

调和函数法存在的问题

调和函数存在两个问题

- 它固定已有的标注 Y_l
  - 标注存在错误怎么办?
  - 想要更加灵活，有时希望偶尔不服从给定的标注
- 它不能够直接处理新的测试数据
  - f 仅定义在X_u上
  - 我们需要把新的测试点加到图上，重新求解调和函数

局部和全局一致性方法（Local and Global Consistency）

(1) 邻近的点有可能有相同的标签;

(2) 在相同结构（簇或者流形）上的点可能有相同的标签;

将标签拓展到多分类任务，假定 $y_i \in \mathcal{Y}$,定义非负的$(l+u)\times|\mathcal{Y}|$的标记矩阵$F=(F_1^T, F_2^T, \dots, F_{l+u}^T)^T$,其中第$i$行的元素$F_i=((F)_{i1},(F)_{i2},\dots,(F)_{i|\mathcal{Y}|})$为示例$x_i$的标记向量，相应的分类规则为

$$y_i = argmax_{1\le j\le|\mathcal{Y}|}(F)_{ij}$$

标记矩阵初始化为

$$F(0)=(Y)_{ij}=\begin{cases}1, & if\ (1\le i\le l)\wedge(y_i=j)\\ 0, & otherwise\end{cases}$$

局部和全局一致性方法（Local and Global Consistency）

基于W构造一个标记传播矩阵$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$,其中$D^{-\frac{1}{2}} = \mathrm{diag}\left(\frac{1}{\sqrt{d_1}}, \frac{1}{\sqrt{d_2}}, \dots, \frac{1}{\sqrt{d_{l+u}}}\right)$

于是有迭代算式 $F(t+1) = \alpha S F(t) + (1-\alpha) Y$

收敛后$F^* = \lim_{t \to \infty} F(t) = (1-\alpha)(I - \alpha S)^{-1} Y$

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

过程:

基于式(13.11)和参数 σ 得到 $\mathbf{W}$;
基于 $\mathbf{W}$ 构造标记传播矩阵 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$;
根据式(13.18)初始化 $\mathbf{F}(0)$;
$t = 0$;
repeat
  $\mathbf{F}(t+1) = \alpha \mathbf{S} \mathbf{F}(t) + (1-\alpha) \mathbf{Y}$;
  $t = t + 1$
until 迭代收敛至 $\mathbf{F}^*$
for $i = l+1, l+2, \dots, l+u$ **do**
  $y_i = \arg\max_{1 \leqslant j \leqslant |\mathcal{Y}|} (\mathbf{F}^*)_{ij}$
end for

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

图 13.5 迭代式标记传播算法

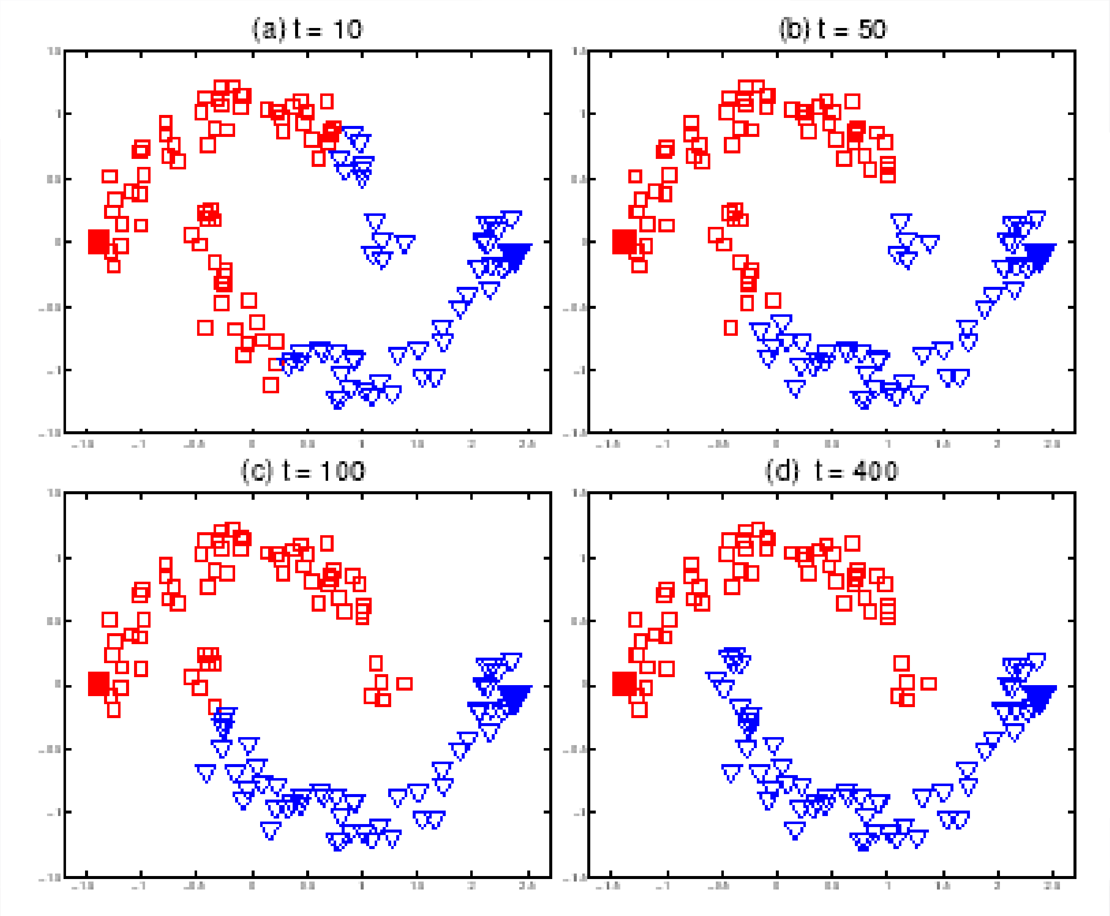
局部和全局一致性方法（Local and Global Consistency）

该算法对应于正则化框架

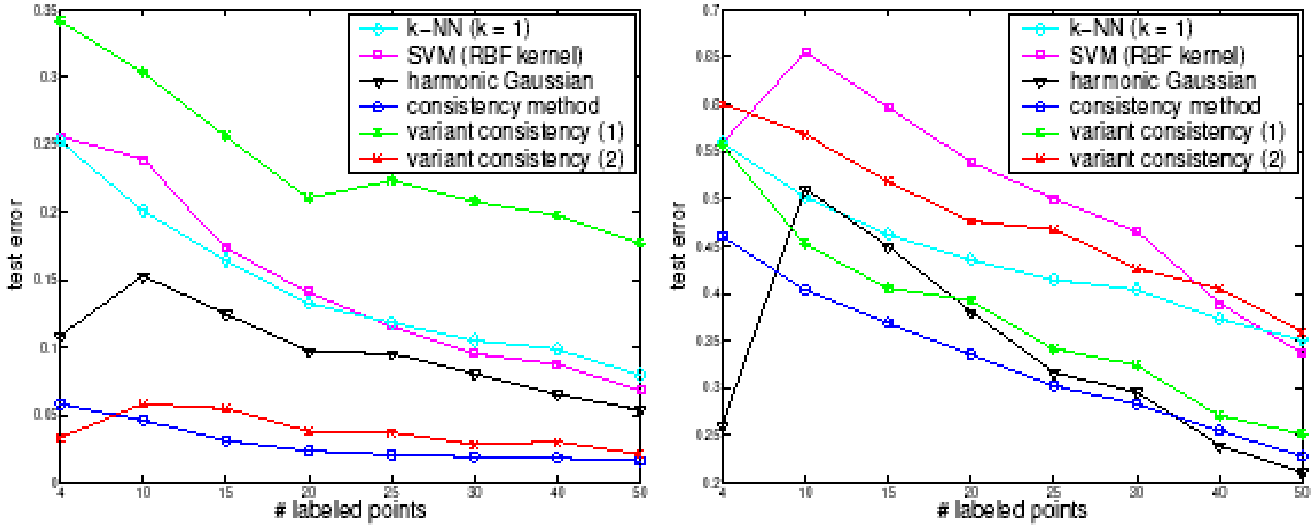
$$min_F \frac{1}{2}\left(\sum_{i,j=1}(W)_{ij}\left\|\frac{1}{\sqrt{d_i}}F_i-\frac{1}{\sqrt{d_j}}F_j\right\|^2\right)+\mu\sum_{i=1}^{l}\|F_i-Y_i\|^2$$

- 允许$f(X_l)$不同于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

局部和全局一致性方法（Local and Global Consistency）



迭代收敛过程



数字识别（USPS）　　文本分类（20newsgroup）

基于图的算法的总结

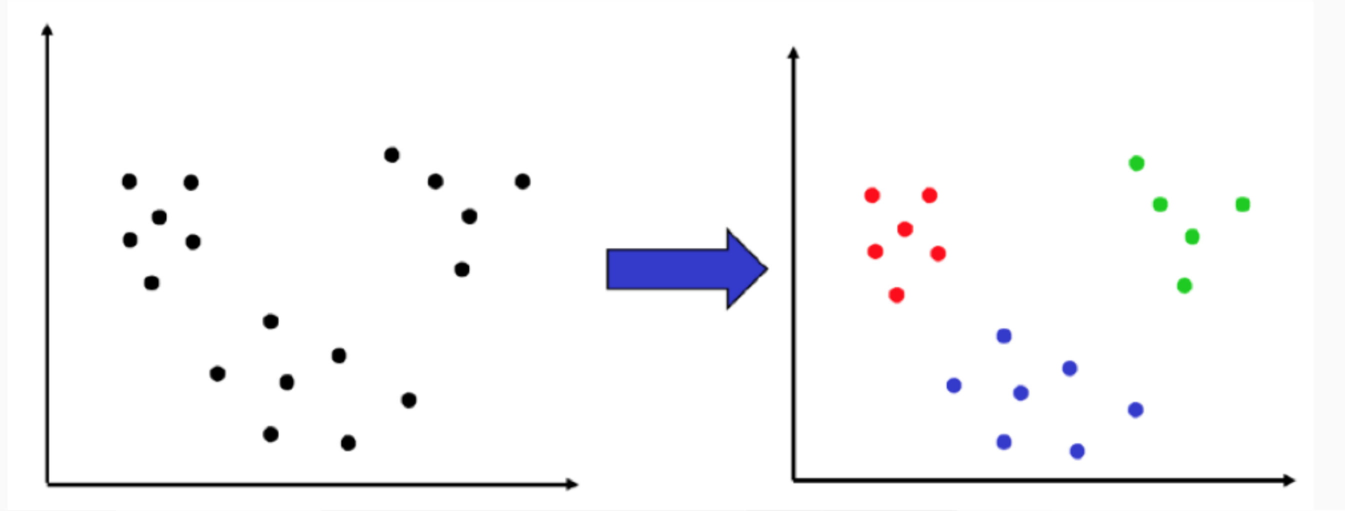
- 优点:
  - 清晰的数学框架
  - 当图恰好拟合该任务时，性能强大
  - 能够被扩展到有向图
- 缺点:
  - 图质量差的时候性能差
  - 对图的结构和权重敏感
  - 存储开销比较大

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
- 前沿进展

半监督聚类

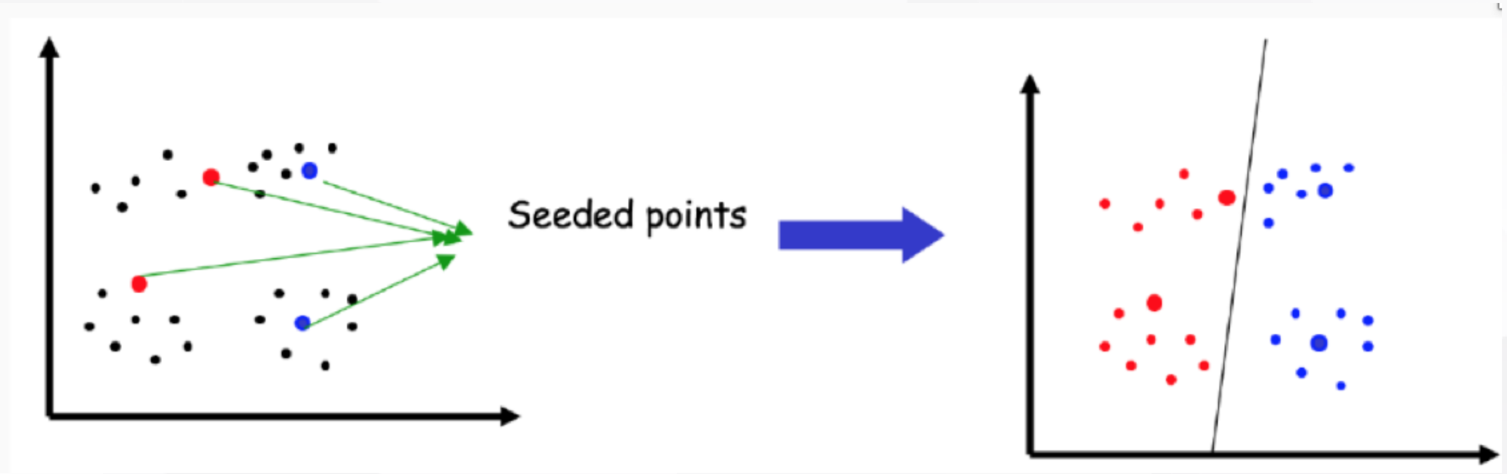
- 聚类是无监督学习的一种算法



- 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文档的类别标签



半监督聚类

- 已知少量的标记信息

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

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.

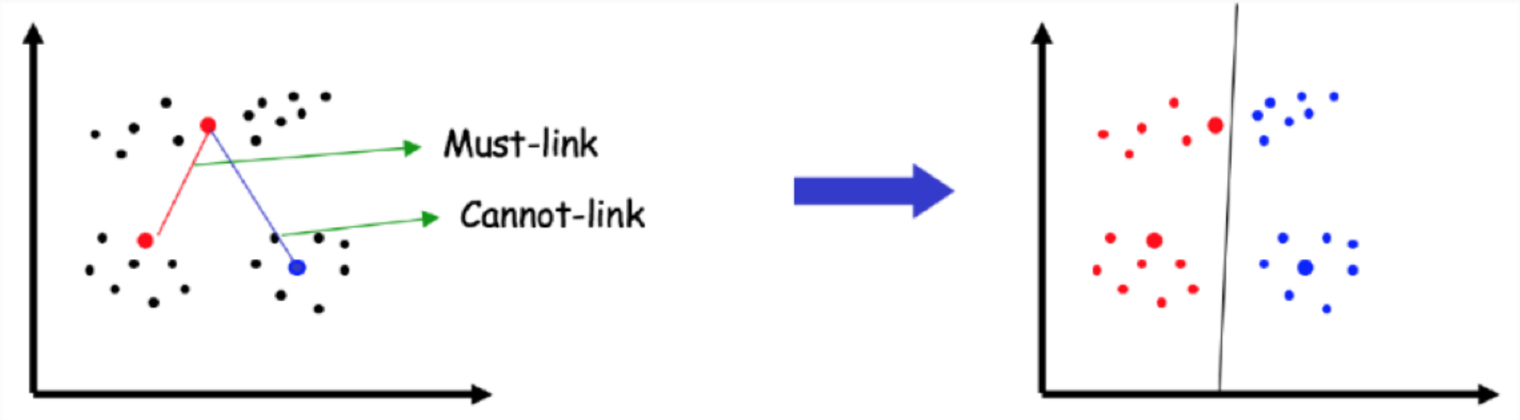
过程：

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

输出： 簇划分 $\{C_1,C_2,\ldots,C_k\}$

半监督聚类

- 根据给定的不同的领域知识:
  - 用户知道其中一些文档是相关 (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)

检查是否违背约束条件

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:

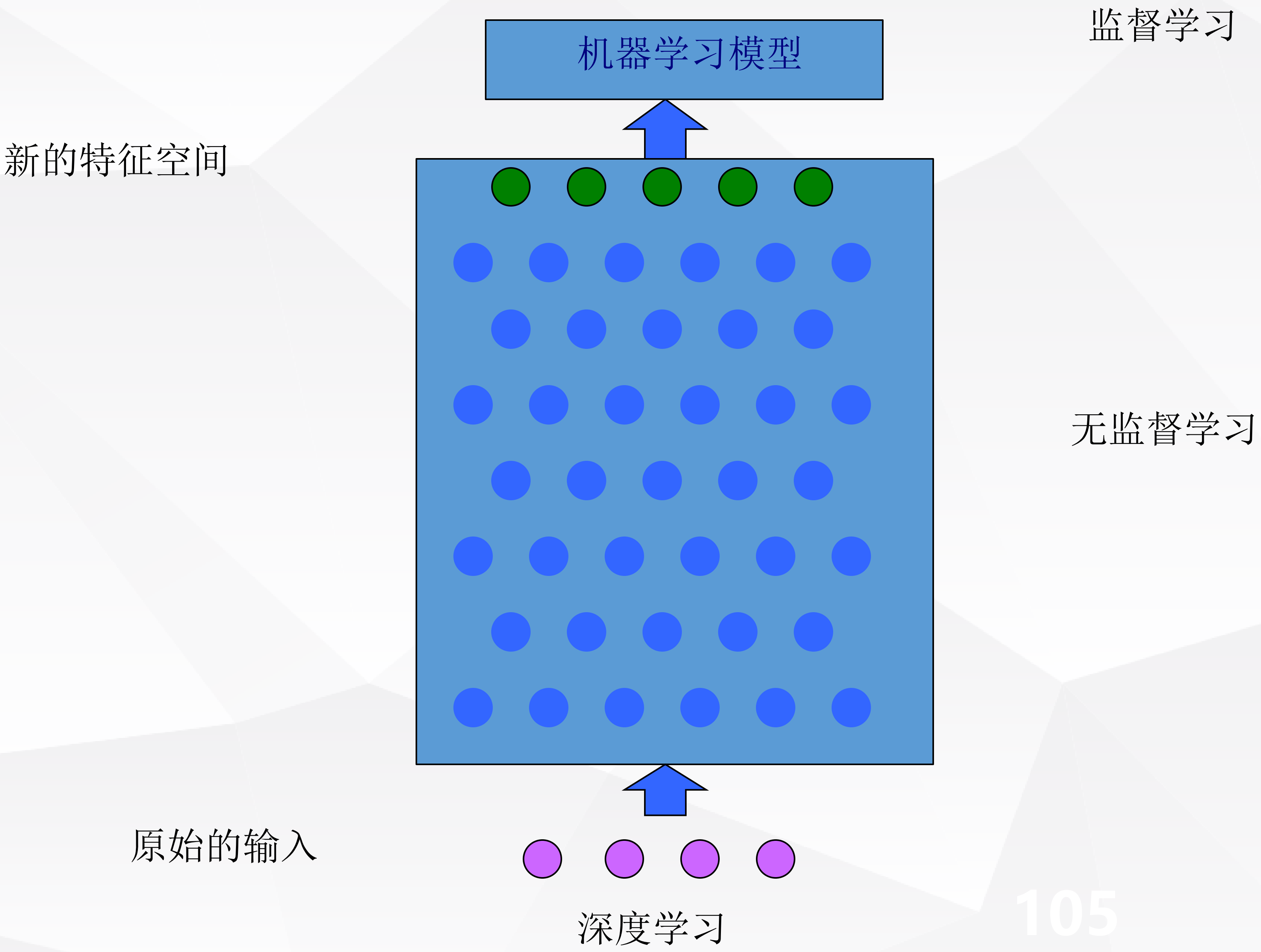
从 D 中随机选取 k 个样本作为初始均值向量$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$;
repeat
　$C_j = \varnothing \; (1 \leqslant j \leqslant k)$;
　for $i = 1, 2, \ldots, m$ **do**
　　计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j \; (1 \leqslant j \leqslant k)$ 的距离: $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2$;
　　$\mathcal{K} = \{1, 2, \ldots, k\}$;
　　is_merged=false;
　　while ¬ is_merged **do**
　　　基于 $\mathcal{K}$ 找出与样本 $\boldsymbol{x}_i$ 距离最近的簇: $r = \arg\min_{j \in \mathcal{K}} d_{ij}$;
　　　检测将 $\boldsymbol{x}_i$ 划入聚类簇 C_r 是否会违背 $\mathcal{M}$ 与 $\mathcal{C}$ 中的约束;
　　　if ¬ is_voilated **then**
　　　　$C_r = C_r \bigcup \{\boldsymbol{x}_i\}$;
　　　　is_merged=true
　　　else
　　　　$\mathcal{K} = \mathcal{K} \setminus \{r\}$;
　　　　if $\mathcal{K} = \varnothing$ **then**
　　　　　break并返回错误提示
　　　　end if
　　　end if
　　end while
　end for
　for $j = 1, 2, \ldots, k$ **do**
　　$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$;
　end for
until 均值向量均未更新

输出: 簇划分 $\{C_1, C_2, \ldots, C_k\}$

大纲

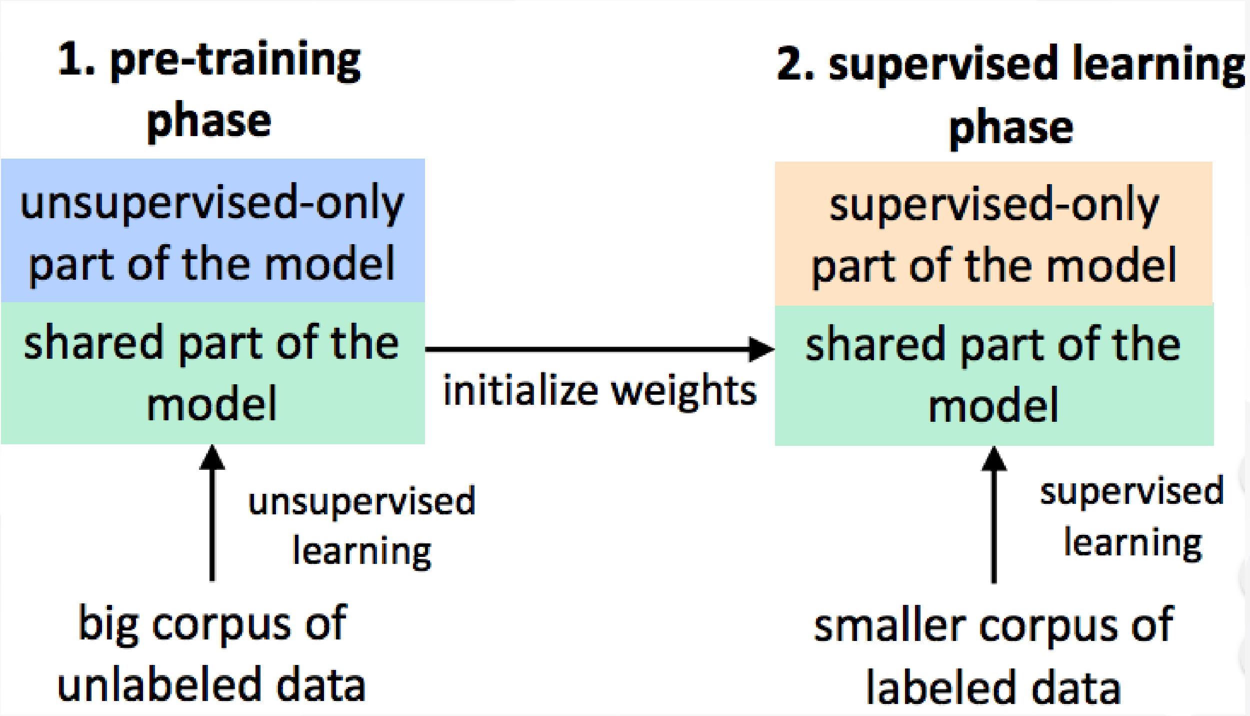
- 简介
- 半监督学习算法
  - 自我训练
  - 多视角学习
  - 生成模型
  - S3VMs
  - 基于图的算法
  - 半监督聚类
- 前沿进展

深度学习中的半监督学习



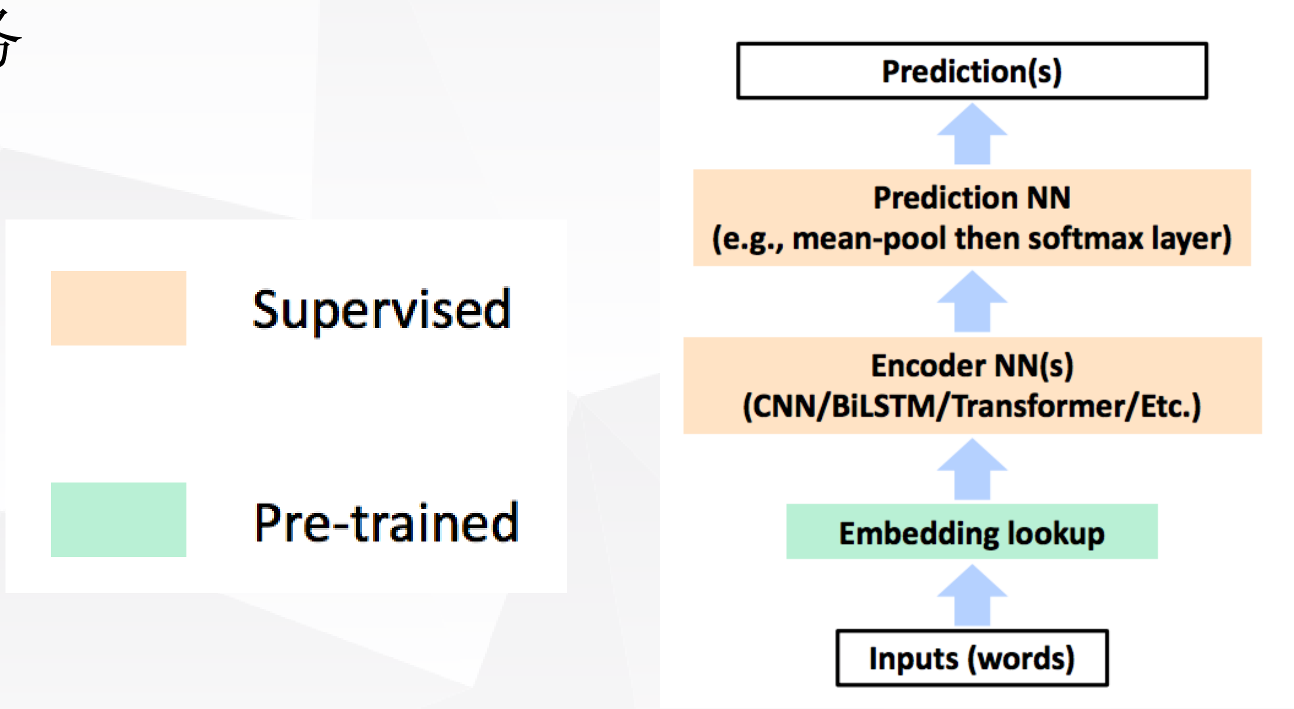
深度学习中的半监督学习

- 自然语言处理中的预训练
  - 在无标注数据上训练模型
  - 学习到的权重放到监督任务的模型中



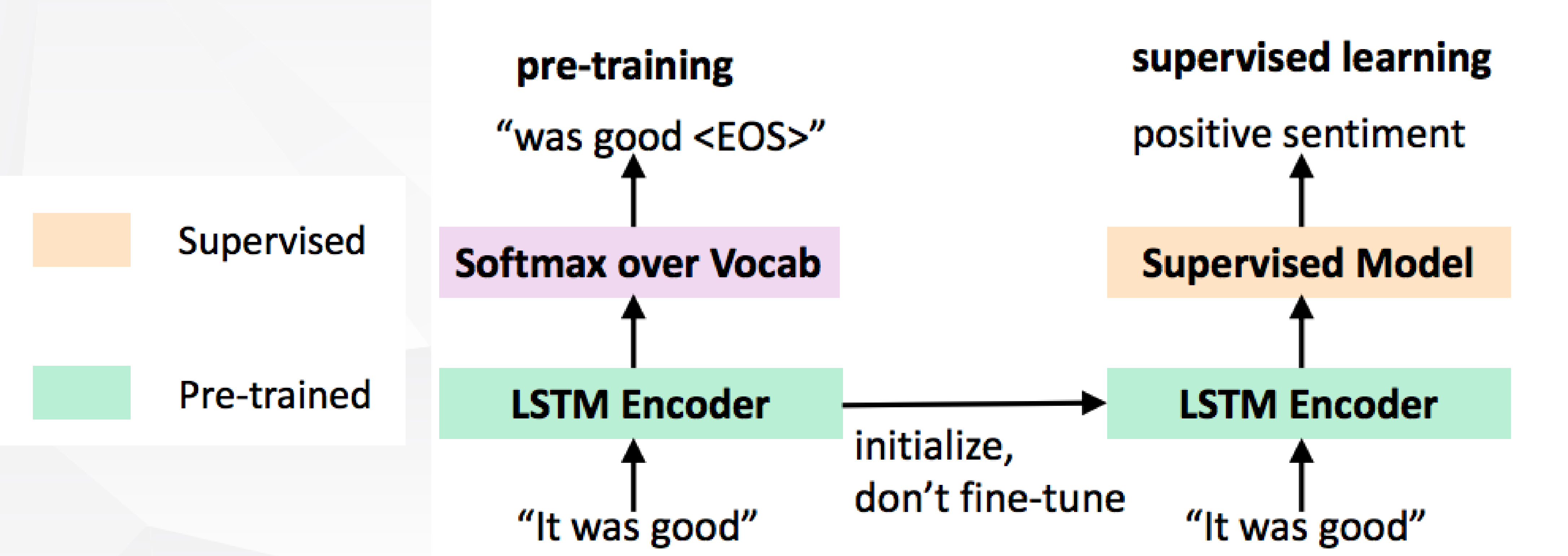
预训练

- Word2vec
  - 共享词向量部分
  - 无监督学习： skip-gram/cbow/glove
  - 有监督学习：一些NLP任务



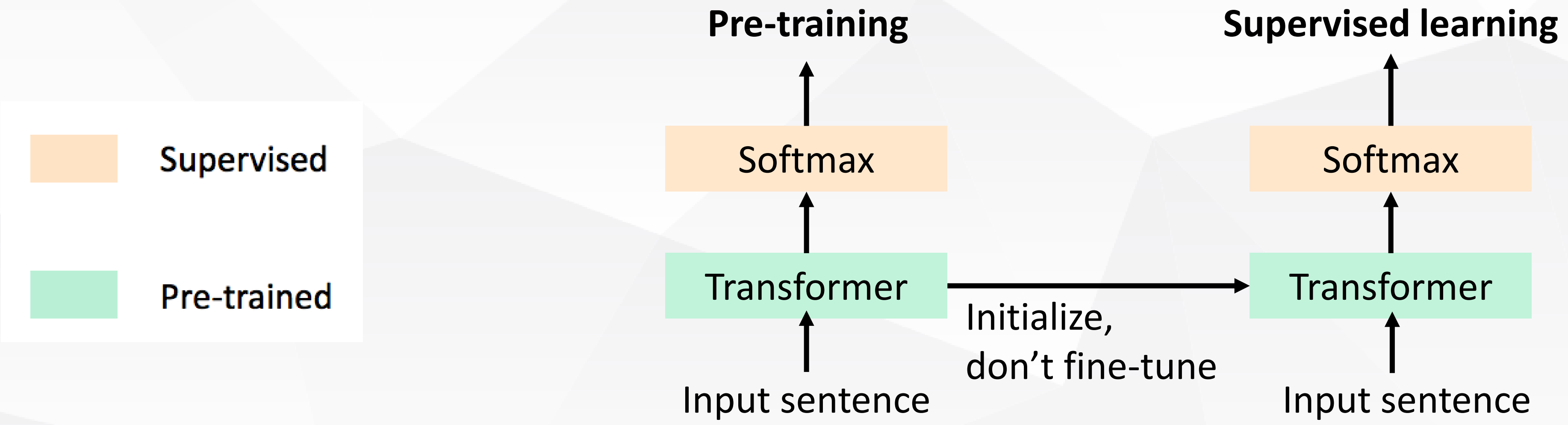
预训练

- ELMo
  - 在双向语言模型任务上预训练模型
  - 共享词向量和上下文编码部分(LSTM)



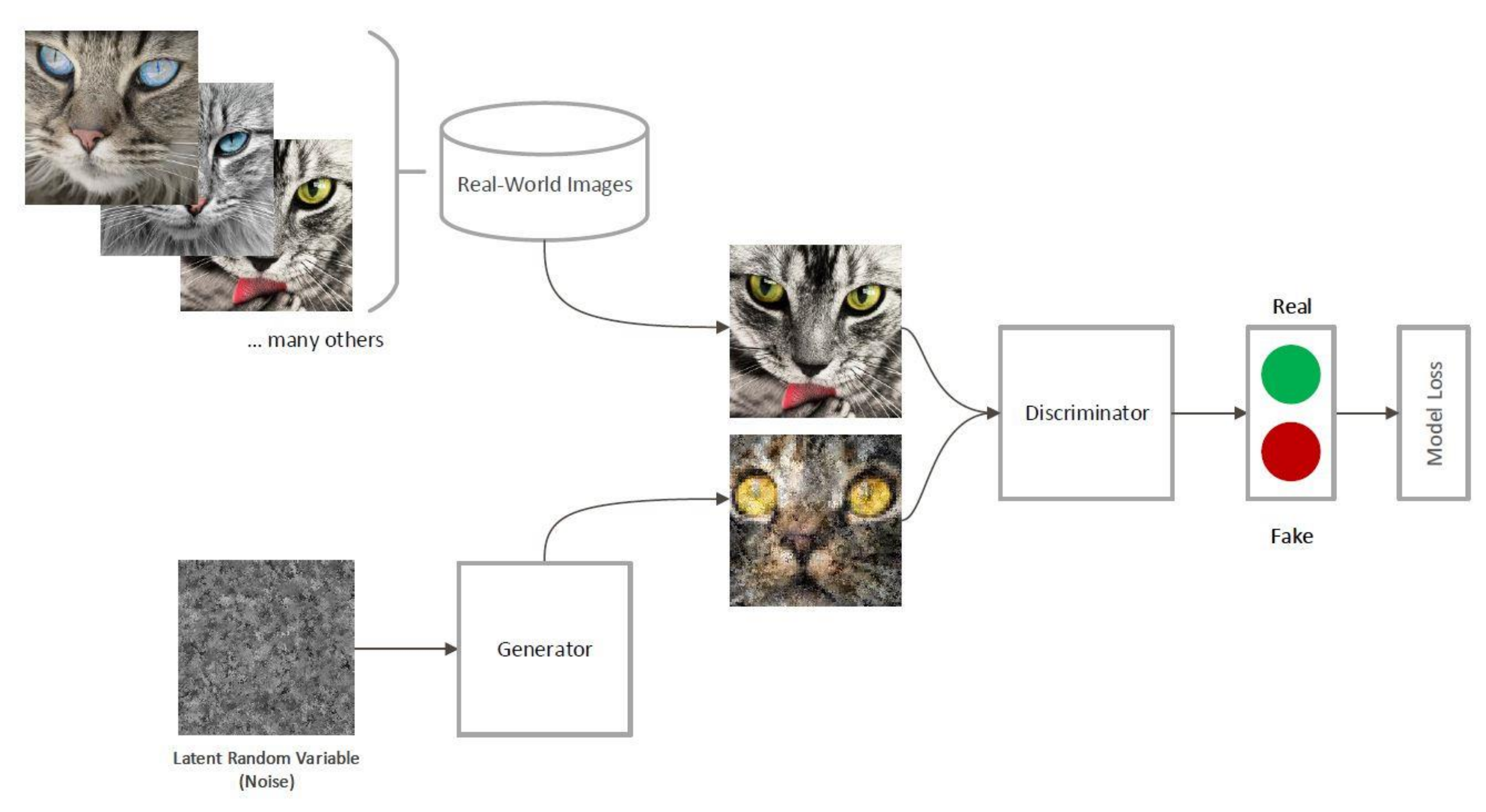
预训练

- BERT
  - 共享全部的编码部分
  - 预训练的任务是屏蔽(mask)语言模型和上下句关系预测
    - 随机屏蔽一些词，无监督模型根据上下文预测该词
    - 判断两句话是不是连续的两句话，例如，随机将部分下一句换成其他句子
  - 监督模型只保留最后的任务特定的输出层不预训练，例如，分类任务非预训练参数仅一层softmax



semi-supervised GAN

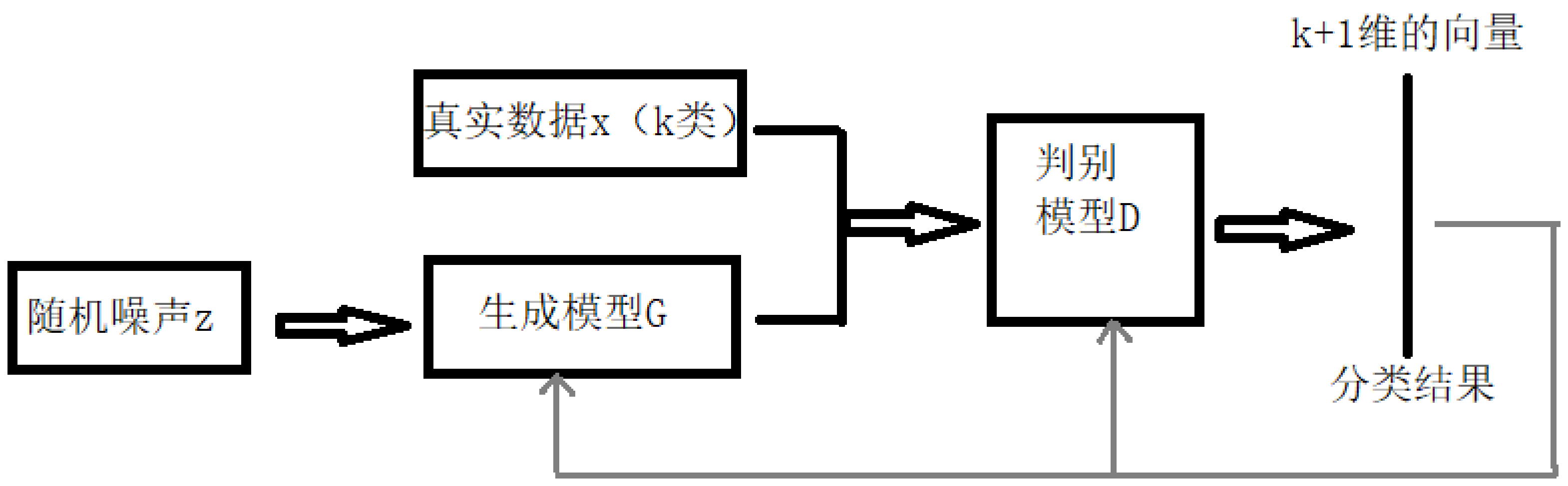
对抗生成网络（GAN）



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))].$$

semi-supervised GAN

- 将GAN用在半监督学习领域的时候需要做一些改变
  - 生成器不做改变，仍然负责从输入噪声数据中生成图像
  - 判别器D不再是一个简单的真假分类（二分类）器，假设输入数据有K类，D就是K+1的分类器，多出的那一类是判别输入是否是生成器G生成的图像



总结

- 蓬勃发展的领域（2008年开始ICML “Ten Years’ Best Paper” 6年3次获奖）
- 通用想法: 从有标注和无标注数据学习
- 假设:
  - 平滑假设 (生成式)
  - 聚类假设 (S3VM)
  - 流形假设 (基于图)
  - 独立假设 (联合训练)
- 挑战:
  - 其他假设?
  - 效率
- 使用无标注数据的两种方式:
  - 在损失函数中 (如S3VM)

    非凸 – 优化方法很重要!
  - 正则化 (如图方法)

    凸问题, 但是图的构建很关键

参考文献

- 课程代码：

https://github.com/lixinsu/tutorials2018/blob/master/semi_supervise.ipynb

- 课后作业：
  - 思考传统机器学习和深度学习中半监督学习思路的不同
- 参考资料

- 周志华，《机器学习》
- 常虹《Semi-supervised Learning》
- 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.

The End