第十四章 半监督学习

§ 14. 1 问题引入

§ 14. 2 典型方法

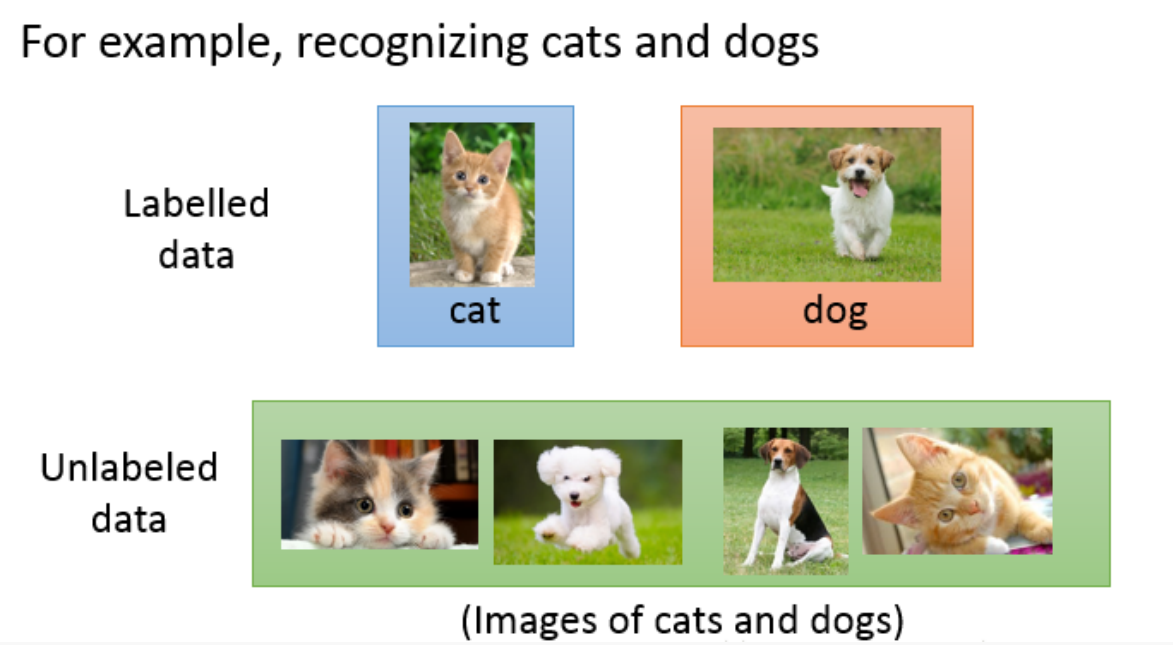
§14.1 问题引入

一、问题背景

二、未标记样本

问题背景

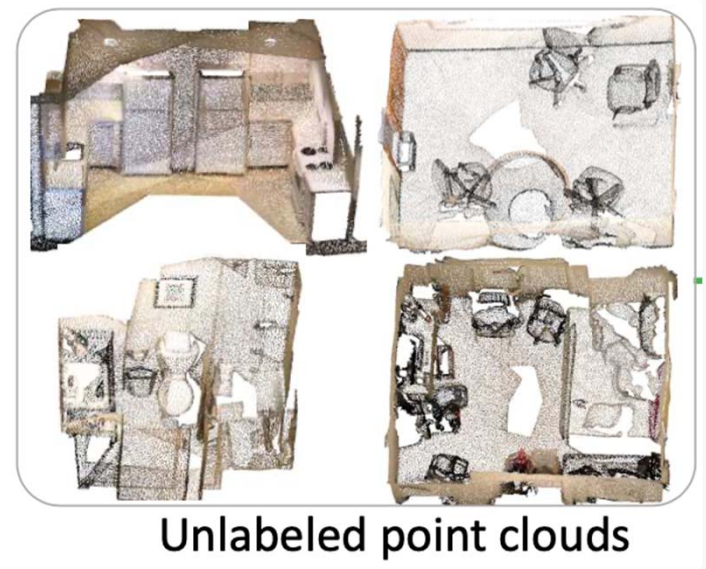
- 标签问题
  - 大多数情况下，**样本的数据量往往都远大于样本的标注量**
  - 以图像分类任务的数据为例：互联网上的海量图片都是数据，但明确标记出类别的图片数量要少得多



问题背景

- 标签问题
  - 如更密集的预测问题（如目标检测），并**不一定所有场景都进行了包围框的标注**
  - 即使进行过标注的场景，可能也没有将所有的物体都标注出来

可能标注并不充分！



未标记样本

未标记样本

- 在**标注样本量**不足的情况下，能否将未标注的数据也利用起来？
- 形式化地看，我们有训练集样本

$$D_l = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$$

- 这l个样本的类别标记已知，称为“**有标记**”样本
- 还有

$$D_u = \{x_{l+1}, x_{l+2}, \dots, x_{l+u}\}, l \ll u$$

- 这u个样本的类别标记未知，称为“**未标记**”样本

未标记样本

未标记样本

- 如果直接使用传统的监督机器学习技术，那么只有D_l可以用于构建模型，D_u所包含的信息则被浪费
- 然而D_l由于样本数较小的缘故，训练样本不足，学出来的模型泛化性能不足，容易过拟合
- 能否在构建模型的过程中利用D_u呢？主动学习！

未标记样本

主动学习

- 将D_u进行人工标记后再训练模型，但这样需要消耗大量时间
- 可以**利用模型和人工标注进行交互**，高效率完成标注任务！
- **先用D_l训练一个模型，再用这个模型去未标注数据D_u中进行预测**，挑选出置信度较低的样本，进行人工标注
- 将新获得的有标记样本加入D_l中重新训练一个模型，再进一步去挑选新的样本进行标注
- 这样**每次都能挑出对改善模型性能帮助大的样本**，标记比较少的次数就可以获得较好的性能

未标记样本

主动学习

- 主动学习的目标是**使用尽量少的“查询”来获得尽量好的性能**，其本质是利用现有模型代替一部分人工进行样本的筛选，辅助标注过程
- 主动学习可以增加标注效率，但仍然需要与外界的交互，依赖人工标注和专家知识
- 如果采用不交互的手段，不额外获取信息，能否利用D_u来提升模型性能呢？

未标记样本

未标记样本

- 事实上，未标记样本虽然没有直接包含标记信息，但如果它们与有标记样本是从同样的数据源独立同分布采样出来，则**其中包含的关于数据分布的信息对建立学习模型帮助很大**
- 如图所示，如果仅仅基于左边的正例和负例，则中间的待判别样本无法确定类别
- 如果有右边的大量未标记样本，则有**很大把握把待判别样本判断为正例**

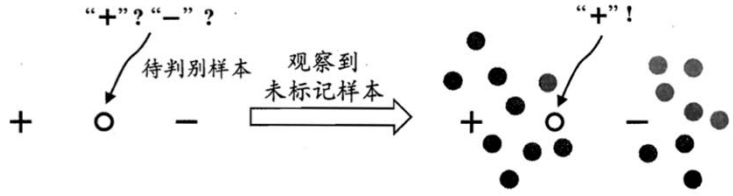


图 13.1 未标记样本效用的例示. 右边的灰色点表示未标记样本

未标记样本

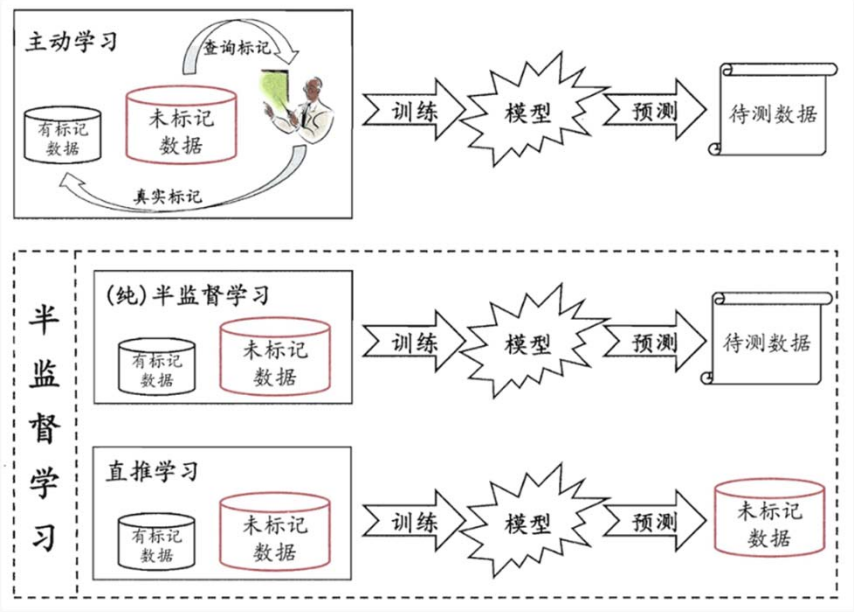
- 半监督学习
  - **定义**：让学习器**不依赖外界交互**、**自动地利用未标记样本来提升模型的学习性能**，就是半监督学习
  - **应用**：半监督学习的应用需求非常强烈，可以极大程度**缓解数据标注的时间成本与人力成本**。例如在进行计算机辅助医学影像分析时，可以从医院获得大量医学影像数据，但请专家把影像中病灶全部标记出来是不现实的

未标记样本

- 半监督学习
  - 要利用未标记样本，必然要**做一些将未标记样本所揭示的数据分布信息与类别标记相联系的假设**
  - 最常见的假设是“**聚类假设**”，即假设数据存在簇结构，同一个簇的样本属于同一个类别
  - 另一种常见假设是“**流形假设**”，即假设数据分布在一个流形结构上，邻近的样本拥有相似的输出值
  - “邻近”程度常用“相似”程度来刻画，因此流形假设可看作聚类假设的推广。但无论是哪种假设，其本质都是“**相似的样本拥有相似的输出**”

未标记样本

- 半监督学习
  - 半监督学习可以进一步划分为**纯半监督学习**和**直推学习**
  - 前者假定训练数据中的未标记样本并非待预测的数据，后者假定学习过程中所考虑的未标记样本恰是带预测数据



§14.2 典型方法

一、生成式半监督方法

二、半监督支持向量机

三、半监督图学习

四、基于分歧的方法

生成式半监督方法

生成式半监督方法

- 生成式半监督方法是直接**基于生成式模型**的方法
- 假设所有数据（无论是否有标记）都是由同一个潜在的模型“生成”的
- 该假设使得我们能够**通过潜在模型的参数将未标记数据与学习目标联系起来**，从而标记数据的标记可以看做模型的缺失参数
- 通常可基于EM算法进行极大似然估计求解

生成式半监督方法

生成式半监督方法

- 给定样本x，其真实类别标记为$y \in Y$，其中$Y = \{1, 2, \ldots, N\}$为所有可能的类别
- 假设样本由高斯混合模型生成，且每个类别对应一个高斯混合分布
- 换言之，数据样本是基于如下概率密度生成：

$$p(x) = \sum_{i=1}^{N} \alpha_i \cdot p(x|\mu_i, \Sigma_i)$$

- 其中，混合系数$\alpha_i \geq 0, \sum_{i=1}^{N} \alpha_i = 1$，$p(x|\mu_i, \Sigma_i)$是样本$x$属于第$i$个高斯混合成分的概率，$\mu_i$和$\Sigma_i$为该高斯混合成分的参数

生成式半监督方法

生成式半监督方法

- 令$f(x) \in Y$表示模型f对x的预测标记，$\Theta \in \{1, 2, \ldots, N\}$表示样本$x$隶属的高斯混合成分，由最大化后验概率可知

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

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

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

- 其中

$$p(\Theta = i|x) = \frac{\alpha_i \cdot p(x|\mu_i, \Sigma_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(x|\mu_i, \Sigma_i)}$$

生成式半监督方法

生成式半监督方法

$$p(\Theta = i|x) = \frac{\alpha_i \cdot p(x|\mu_i, \Sigma_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(x|\mu_i, \Sigma_i)}$$

- 为样本x由第i个高斯混合成分生成的后验概率
- $p(y = j|\Theta = i, x)$为x由第i个高斯混合成分生成且其类别为j的概率
- 由于假设每个类别对应一个高斯混合成分，因此$p(y = j|\Theta = i, x)$仅于样本x所属的高斯混合成分Θ有关，可用$p(y = j|\Theta = i)$代替
- 不失一般性，假定第i个类别对应于第i个高斯混合成分，即

$$p(y = j|\Theta = i) = \begin{cases} 1, & \text{当且仅当} i = j \\ 0, & \text{其他情况} \end{cases}$$

生成式半监督方法

生成式半监督方法

- 不难发现，估计$p(y=j|\Theta=i,x)$需知道样本的标记，因此仅能使用有标记数据；而$p(\Theta=i|x)$不涉及样本标记，因此有标记和未标记数据均可利用
- 通过引入大量的未标记数据，对这一项的估计由于数据量的增长而更为准确，因此整体的估计可能会更准确
- 给定有标记样本集$D_l=\{(x_1,y_1),(x_2,y_2),\dots,(x_l,y_l)\}$和未标记样本集$D_u=\{x_{l+1},x_{l+2},\dots,x_{l+u}\}, l\ll u, l+u=m$. 假设所有样本独立同分布，且均由同一个高斯混合模型生成

生成式半监督方法

生成式半监督方法

- 用极大似然法来估计高斯混合模型的参数$\{(\alpha_i, \mu_i, \Sigma_i)|1 \leq i \leq N\}$，$D_l \cup D_u$的对数似然是

$$LL(D_l \cup D_u) = \sum_{(x_j,y_j)\in D_l} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p(x_j|\mu_i, \Sigma_i) \cdot p(y_j|\Theta = i, x_j)\right)$$

$$+ \sum_{x_j \in D_u} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p(x_j|\mu_i, \Sigma_i)\right)$$

- 上式由两项组成：**基于有标记数据D_l的有监督项和基于未标记数据D_u的无监督项**

生成式半监督方法

生成式半监督方法

- 显然高斯混合模型参数估计可用EM算法求解，迭代更新式如下：
  - E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：

$$\gamma_{ji} = \frac{\alpha_i \cdot p(x_j|\mu_i, \Sigma_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(x_j|\mu_i, \Sigma_i)}$$

  - M步：基于γ_{ji}更新模型参数，其中l_i表示第i类的有标记样本数目：

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

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

$$\alpha_i = \frac{1}{m} \left(\sum_{x_j \in D_u} y_{ji} + l_i \right)$$

生成式半监督方法

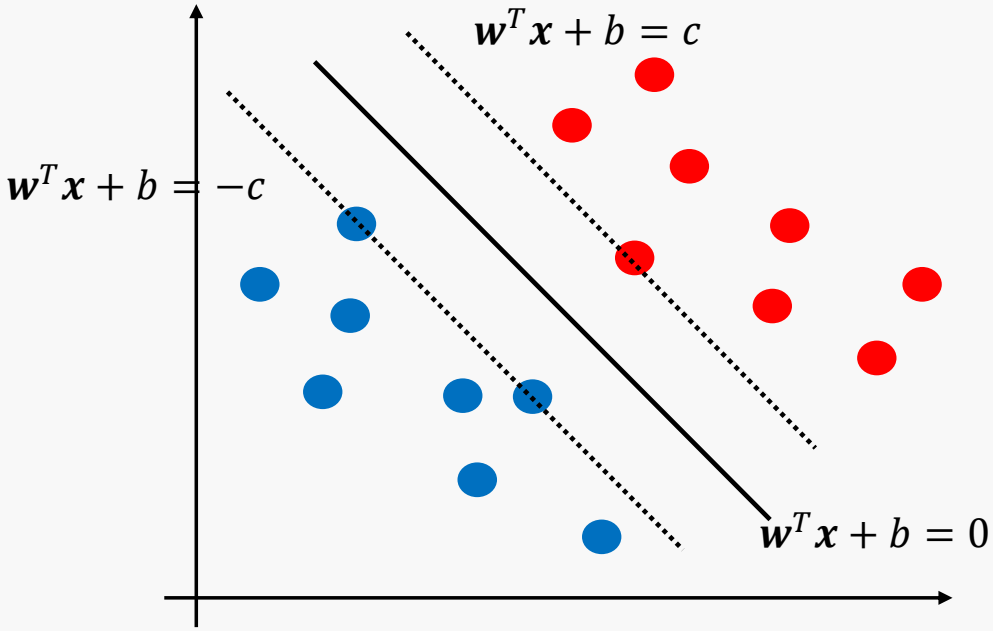
生成式半监督方法

- 以上过程不断迭代直到收敛，即可获得模型参数。然后由最大化后验概率的公式就能对样本进行分类
- 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**方法
- 此类方法在有标记数据极少的情形下往往比其他方法性能更好，但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降低泛化性能

半监督支持向量机

半监督支持向量机

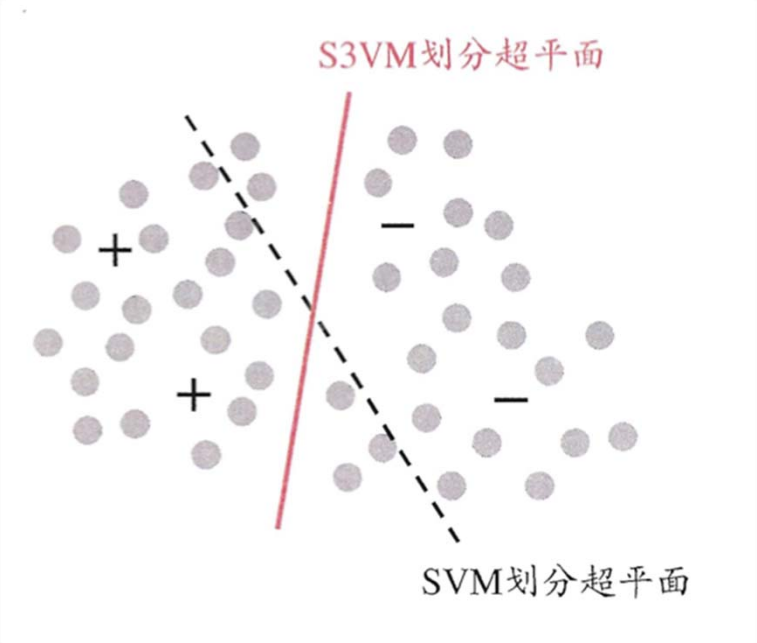
- Semi-Supervised Support Vector Machine，简称 S3VM，是支持向量机在半监督学习上的推广
- 在不考虑未标记样本时，支持向量机试图找到最大间隔划分超平面



半监督支持向量机

半监督支持向量机

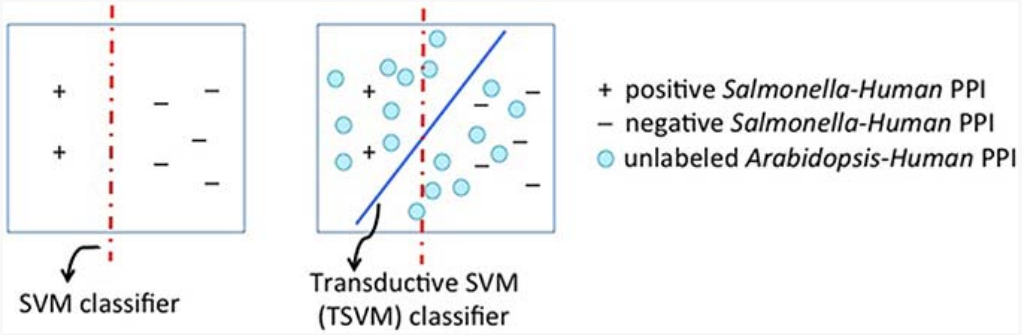
- 在考虑未标记样本后， **S3VM**试图找到能将两类有标记样本分开，且**穿过数据低密度区域**的划分超平面
- **“低密度分隔”**的基本假设，是聚类假设在考虑了线形超平面划分后的推广



半监督支持向量机

TSVM (Transductive Support Vector Machine)

- 最著名的半监督支持向量机，是针对二分类问题的学习方法
- 考虑对未标记样本进行各种可能的**标签指派(label assignment)**，即尝试将每个未标记样本分别作为正例或反例
- 然后在所有这些结果中，寻求一个在所有样本(包括有标记样本和进行了标记指派的未标记样本)上**间隔最大化的划分超平面**
- 一旦划分超平面得以确定，未标记样本的最终标记指派就是其预测结果



半监督支持向量机

TSVM (Transductive Support Vector Machine)

- 给定$D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$和$D_u = \{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$，其中$y_i \in \{-1, +1\}, l \ll u, l + u = m$
- 学习目标：为D_u中的样本给出预测标记$\hat{y} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u}), \hat{y}_i \in \{-1, +1\}$，使得

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

$$\text{s.t. } y_i\left(w^{\mathrm{T}} x_i + b\right) \geqslant 1 - \xi_i, i = 1,2, \ldots, l$$

$$\hat{y}_i\left(w^{\mathrm{T}} x_i + b\right) \geqslant 1 - \xi_i, i = l+1, l+2, \ldots, m,$$

$$\xi_i \geqslant 0, i = 1,2, \ldots, m$$

其中，(w, b)确定了一个划分超平面，ξ为松弛向量，$\xi_i (i = 1, 2, \ldots, l)$对应有标记样本，$\xi_i (i = l + 1, l + 2, \ldots, m)$对应未标记样本

半监督支持向量机

- TSVM (Transductive Support Vector Machine)
  - 采用**局部搜索**来迭代地寻找近似解
  - 利用**有标记样本**学得一个SVM，即忽略关于$\mathbf{D_u}$与$\hat{\mathbf{y}}$的项及约束
  - 利用这个SVM对未标记的数据进行**标记指派**，将SVM预测结果作为“伪标记”赋予未标记样本
  - 此时$\hat{\mathbf{y}}$已知，得到一个**标准SVM问题**，求解出新的划分超平面和松弛向量
  - 注意到此时未标记样本的伪标记很可能不准确，因此$\mathbf{C_u}$要设置为比$\mathbf{C}_l$小的值，使有标记样本所起的作用更大
  - TSVM找出两个标记指派为**异类**且很可能发生**错误**的未标记样本，交换它们的标记，重新求解更新后的划分超平面和松弛向量
  - 逐渐增大$\mathbf{C_u}$以提高未标记样本对优化目标的影响，进行下一轮标记指派调整，直到$\mathbf{C_u} = \mathbf{C}_l$为止

半监督支持向量机

- TSVM (Transductive Support Vector Machine)
  - 采用**局部搜索**来迭代地寻找近似解

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: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

半监督支持向量机

- TSVM (Transductive Support Vector Machine)
  - 类别不平衡问题：某类样本远多于另一类
  - 算法改进：将优化目标中的C_u项拆分为C_u^+与C_u^-两项，分别对应于伪标记而当作正、反例使用的未标记样本，并在初始化时令

$$C_u^+ = \frac{u_-}{u_+} C_u^-$$

其中u_+与u_-为基于伪标记而当作正、反例使用的未标记样本数

  - 若存在一堆未标记样本x_i, x_j，其标记指派$\hat{y}_i, \hat{y}_j$不同，且对应松弛变量满足

$$\xi_i + \xi_j > 2$$

则意味着$\hat{y}_i, \hat{y}_j$很可能错误，需要对二者交换后重新求解，使得每轮迭代后均可使目标函数值下降

半监督图学习

半监督图学习

- 给定一个数据集，我们可将其建模成一个**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)
- 将有标记样本所对应的结点想象为染过色，而未标记样本所对应的结点尚未染色。于是，半监督学习就对应于"颜色"在图上**扩散或传播**的过程
- 由于一个图对应了一个矩阵，这就使得我们能基于**矩阵运算**来进行半监督学习算法的推导与分析

半监督图学习

半监督图学习

- 给定$D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$和$D_u = \{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$，其中$\mathbf{l} \ll \mathbf{u}, \mathbf{l} + \mathbf{u} = \mathbf{m}$
- 基于$D_l \cup D_u$构建一个图$G = (V, E)$，结点集$V = \{x_1, \cdots, x_l, x_{l+1}, \cdots, x_{l+u}\}$，边集$E$可表示为一个亲和矩阵，基于高斯函数定义为

$$(\mathbf{W})_{ij} = \begin{cases} \exp\left(\dfrac{-\|x_i - x_j\|_2^2}{2\sigma^2}\right), & if \quad i \neq j; \\ 0, & otherwise \end{cases}$$

- 其中$i, j \in \{1, 2, \ldots, m\}, \sigma > 0$是用户指定的高斯函数带宽参数

半监督图学习

半监督图学习

- 假定从图$G=(V,E)$将学得一个实值函数$f: V \to \mathbb{R}$，其对应的分类规则为

$$y_i = \mathbf{sign}\left(\mathrm{f}(x_i)\right), y_i \in \{-1, +1\}$$

- 相似的样本应具有相似的标记，定义关于f的能量函数

$$\mathrm{E}(f) = \frac{1}{2}\sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{j}=1}^{m}\left((W)_{ij}\left(\mathrm{f}(x_i)-\mathrm{f}(x_j)\right)\right)^2$$

$$= \frac{1}{2}\left(\sum_{\mathrm{i}=1}^{m} d_i f^2(x_i) + \sum_{\mathrm{j}=1}^{m} d_j f^2(x_j) - 2\sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{j}=1}^{m}(W)_{ij}\mathrm{f}(x_i)\mathrm{f}(x_j)\right)$$

$$= \sum_{\mathrm{i}=1}^{m} d_i f^2(x_i) - \sum_{\mathrm{i}=1}^{m}\sum_{\mathrm{j}=1}^{m}(W)_{ij}\mathrm{f}(x_i)\mathrm{f}(x_j)$$

$$= f^T(\mathrm{D}-W)\mathrm{f},$$

半监督图学习

半监督图学习

- 具有最小能量的函数f在有标记样本上满足$\mathbf{f}(x_i) = y_i (i = 1, 2, \cdots, l)$，
- 在未标记样本上满足$\Delta f = 0$，其中$\Delta = D - W$为拉普拉斯矩阵
- 以第l行与第l列为界，采用分块矩阵表示方式

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} \mathbf{D}_{ll} & 0_{lu} \\ 0_{ul} & \mathbf{D}_{uu} \end{bmatrix}$$

- 能量函数重写为

$$\begin{aligned} E(f) &= \left(\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{f}_u^{\mathrm{T}}\right) \left(\begin{bmatrix} \mathbf{D}_{ll} & 0_{lu} \\ 0_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ul} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \boldsymbol{f}_l^{\mathrm{T}} (\mathbf{D}_{ll} - \mathbf{W}_{ll}) \boldsymbol{f}_l - 2 \boldsymbol{f}_u^{\mathrm{T}} \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} (\mathbf{D}_{uu} - \mathbf{W}_{uu}) \boldsymbol{f}_u. \end{aligned}$$

半监督图学习

半监督图学习

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

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

- 令

$$P = \mathbf{D}^{-1}W = \begin{bmatrix} \mathbf{D}_{ll}^{-1} & 0_{lu} \\ 0_{ul} & \mathbf{D}_{uu}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{D}_{ll}^{-1}\mathbf{W}_{ul} & \mathbf{D}_{ll}^{-1}\mathbf{W}_{lu} \\ \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul} & \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu} \end{bmatrix},$$

- 即$\mathbf{P}_{uu} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}, \mathbf{P}_{ul} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}$

半监督图学习

半监督图学习

- 则$\boldsymbol{f}_u$可重写为

$$\begin{aligned}\boldsymbol{f}_u &= \left(\mathbf{D}_{uu}\left(\mathbf{I}-\mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\right)\right)^{-1}\mathbf{W}_{ul}\boldsymbol{f}_l \\ &= \left(\mathbf{I}-\mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\right)^{-1}\mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}\boldsymbol{f}_l \\ &= \left(\mathbf{I}-\mathbf{P}_{uu}\right)^{-1}\mathbf{P}_{ul}\boldsymbol{f}_l.\end{aligned}$$

- 将$\boldsymbol{D}_l$上的标记信息作为$\boldsymbol{f}_l=(\boldsymbol{y}_1;\boldsymbol{y}_2;\ldots;\boldsymbol{y}_l)$代入，可利用求得的$\boldsymbol{f}_u$对未标记样本进行预测
- 上面描述了一个针对二分类问题的标记传播方法

半监督图学习

- 适用于多分类问题的标记传播方法
  - 假定 $y_i \in \mathcal{Y}$, 仍基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$, 其中结点集 $V = \{x_1, \ldots, x_l, \ldots, x_{l+u}\}$
  - 定义一个非负标记矩阵 $F = \left(F_1^T, F_2^T, \ldots, F_{l+u}^T\right)^T$，其第 i 行元素 $\mathrm{F}_i = \left((\mathrm{F})_{i1}, (\mathrm{F})_{i2}, \ldots, (\mathrm{F})_{\mathrm{i}|\mathcal{Y}|}\right)$ 为实例 x_i 的标记向量
  - 相应的分类规则为

$$y_i = \arg \max_{1 \leqslant \mathrm{j} \leqslant |\mathcal{Y}|} (\mathrm{F})_{ij}$$

  - 将 F 初始化为

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

半监督图学习

- 适用于多分类问题的标记传播方法
  - 基于W构造一个标记传播矩阵

$$\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$$

  - 其中

$$\mathbf{D}^{-\frac{1}{2}} = \mathbf{diag}\left(\frac{1}{\sqrt{d_1}}, \frac{1}{\sqrt{d_2}}, \dots, \frac{1}{\sqrt{d_{\mathrm{i}+u}}}\right)$$

  - 迭代计算式

$$\mathbf{F}(\mathrm{t}+1) = \alpha\mathbf{S}\mathbf{F}(\mathrm{t}) + (1-\alpha)\mathbf{Y}$$

  - 其中$\alpha \in (0,1)$为用户指定的参数，用于对标记传播项 SF(t) 与初始化项 Y 的重要性进行折中，迭代收敛可得

$$\mathbf{F}^* = \lim_{\mathrm{t}\mapsto\infty}\mathbf{F}(\mathrm{t}) = (1-\alpha)(\mathrm{I}-\alpha\mathbf{S})^{-1}\mathbf{Y}$$

半监督图学习

迭代式标记传播方法

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 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
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 $i = l+1, l+2, \ldots, l+u$ **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: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

半监督图学习

- 迭代式标记传播方法
  - 正则化框架

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

  - 其中$\mu > 0$为正则化参数
  - 当

$$\mu = \frac{1-\alpha}{\alpha}$$

  - 上式最优解为迭代式标记传播算法的迭代收敛解

半监督图学习

半监督图学习方法缺陷

- **存储开销**：若样本数为$O(m)$，则算法中所涉及的矩阵规模为$O(m^2)$，这使得此类算法很难直接处理大规模数据
- 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
- 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，例如将D_l和经标记传播后得到标记的D_u包含并作为训练集，另外训练一个学习器例如支持向量机来对新样本进行预测

基于分歧的方法

- 生成式半监督方法、半监督支持向量机、半监督图学习基于**单学习器**
- 基于分歧的方法使用**多学习器**，多学习器之间的**分歧**(disagreement)
  对未标记数据的利用至关重要
  - **协同训练**算法是此类方法的重要代表
  - 学习器之间的分歧由**多视图数据**引入的，而学习器本身是一样的
- 多视图数据的概念
  - 一个数据对象往往同时拥有多个**属性集**
  - 每个属性集就构成了一个**视图**
  - 多个视图一般天然具有**相容性**和**互补性**
  - 举例：一部电影的**画面**和**声音**构成数据的两个视图

基于分歧的方法

- 多视图数据举例（电影数据）

视图1：电影图片画面

视图2：电影音频

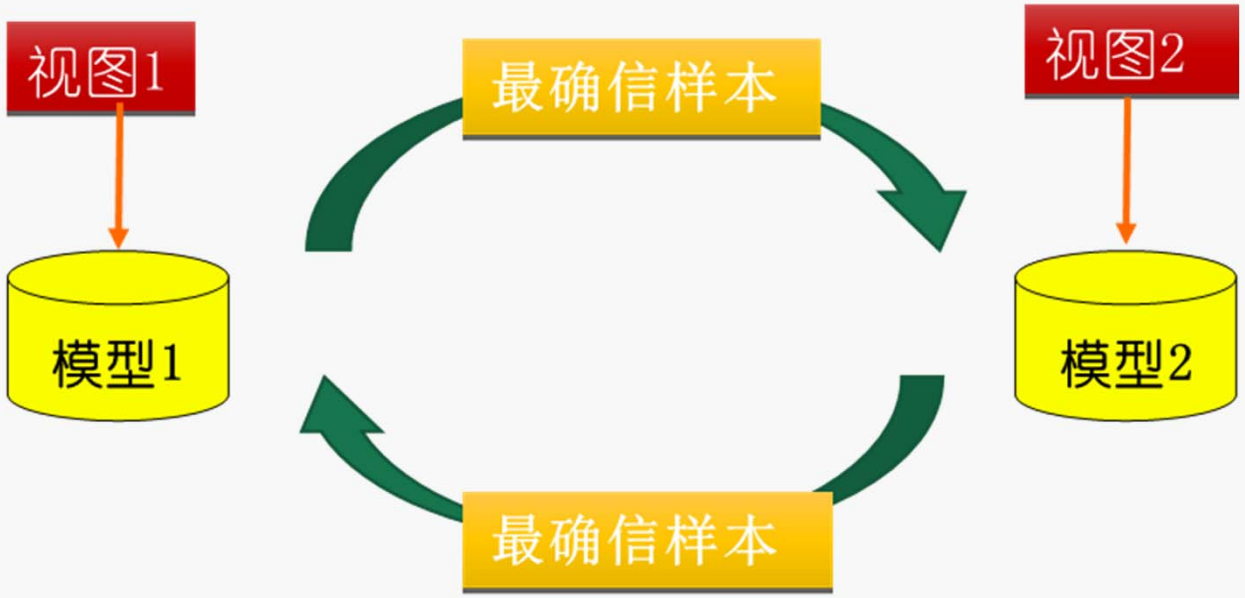
视图3：电影影评

基于分歧的方法

- 多视图数据的性质
  - 记一个**两视图**数据集的一个有标记样本为$(\langle x^1, x^2 \rangle, y)$
  - **相容性**
    - 不同视图所包含的有关标记空间y的信息是一致的
    - y^1和y^2分别表示依据两种视图判别得到的标记空间，应有$y = y^1 = y^2$
    - 例如：从电影的画面和声音得到的有关电影的分类信息应当一致
  - **互补性**
    - 若在学习系统中显示考虑多视图，可以显著减小判别的方差
    - 例如：若仅凭图像画面认为“可能是动作片”，仅凭声音信息也认为“可能是动作片”。则同时考虑两者时，就有很大的把握判别为动作片

基于分歧的方法

- 基于多视图的半监督学习
  - 协同训练正是很好地利用了多视图的**相容性和互补性**
  - 不妨假设数据拥有**充分**且**条件独立**的两个视图



基于分歧的方法

基于多视图的半监督学习

x_i 的上标仅用于指代两个视图，不表示序关系，即 $\langle x_i^1, x_i^2\rangle$ 与 $\langle x_i^2, x_i^1\rangle$ 表示的是同一个样本.

输入: 有标记样本集 $D_l = \{(\langle x_1^1, x_1^2\rangle, y_1), \ldots, (\langle x_l^1, x_l^2\rangle, y_l)\}$;
未标记样本集 $D_u = \{\langle x_{l+1}^1, x_{l+1}^2\rangle, \ldots, \langle x_{l+u}^1, x_{l+u}^2\rangle\}$;
缓冲池大小 s;
每轮挑选的正例数 p;

在视图 j 上用有标记样本训练 h_j.

```
for j = 1,2 do
    h_j ← 𝔏(D_l^j);
    考察 h_j 在 D_s^j = {x_i^j | ⟨x_i^j, x_i^{3-j}⟩ ∈ D_s} 上的分类置信度，挑选 p 个正例置信度最高的样本 D_p ⊂ D_s、n 个反例置信度最高的样本 D_n ⊂ D_s;
    由 D_p^j 生成伪标记正例 D̃_p^{3-j} = {(x_i^{3-j}, +1) | x_i^j ∈ D_p^j};
    由 D_n^j 生成伪标记反例 D̃_n^{3-j} = {(x_i^{3-j}, −1) | x_i^j ∈ D_n^j};
    D_s = D_s \ (D_p ⋃ D_n);
end for
if h_1, h_2 均未发生改变 then
    for j = 1,2 do
        D_l^j = D_l^j ⋃ (D̃_p^j ⋃ D̃_n^j);
    end for
end if
end for
```

扩充有标记数据集.

输出: 分类器 h_1, h_2

图 13.6 协同训练算法

基于分歧的方法

- 基于多视图的半监督学习
  - **充分性**：每个视图都包含足以产生最优学习器的信息
    - 保证各学习器都能在对应的视图上训练成功
  - **条件独立性**：在给定类别标记条件下，两个视图独立
    - “分歧”的来源，如果两个视图完全相关，则学习得到的学习器将一样
  - **未标记样本缓冲池**
    - 如果在每轮学习中都考察分类器在所有未标记样本上的分类置信度，会产生很大的计算开销
  - **分类置信度的估计**
    - 由不同的学习算法相应决定
    - 朴素贝叶斯分类器<->后验概率，支持向量机<->间隔大小

基于分歧的方法

- 协同训练过程虽简单，但令人惊讶的是，理论证明显示出：若两个视图**充分且条件独立**，则可利用未标记样本通过协同训练将弱分类器的泛化性能提升到任意高[Blum and Mitchell，1998]

- 不过视图的条件独立性在现实任务中通常很难满足，因此性能提升幅度不会那么大。但研究表明，即使在更弱的条件下，协同训练仍可有效地提升弱分类器的性能[周志华，2013]
  - 例如，一部电影的画面和声音显然不条件独立

基于分歧的方法

- 基于单视图的半监督学习
  - **协同训练算法**本身是为多视图数据而设计的
  - 此后出现了一些能在**单视图**数据上使用的变体算法
    - 单视图算法不能利用不同视图引入“**分歧**”
    - “分歧”的来源变为**不同的学习器**
  - 产生不同学习器的策略
    - 不同的学习算法[Goldman and Zhou, 2000]
    - 不同的数据采样[Zhou and Li, 2005b]
    - 不同的参数设置[Zhou and Li, 2005a]
  - 即使没有多视图数据，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]

基于分歧的方法

- 优点
  - 基于分歧的方法**只需采用合适的基学习器**，就较少受到模型假设、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
- 局限
  - 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