弱监督学习

刘家锋

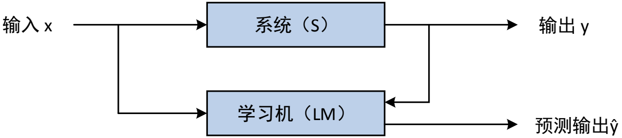
哈尔滨工业大学

弱监督学习

弱监督学习过程

统计学习

- **学习的目标**
  - 希望学习机LM能够模拟被学习的系统S，预测输出$\hat{y}$尽量接近实际输出y
  - 学习机能够在一定程度上代替系统S，完成某些工作



学习的过程

- **学习机**
  - 输出$\hat{y}$与输入$\mathbf{x}$之间可以看作是一个函数关系：

$$\hat{y} = f(\mathbf{x})$$

- **学习的过程**
  - 优化泛函–期望风险：

$$\min_f R(f) = \int L(y, f(\mathbf{x}))dF(\mathbf{x}, y)$$

- **常用的风险函数**
  - 平方误差风险：$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$
  - 似然函数风险：$L(p(\mathbf{x})) = -\ln p(\mathbf{x})$

统计学习的过程

- **经验风险优化**
  - 用一组训练样本$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \sim F(\mathbf{x}, y)$的经验风险近似期望风险：

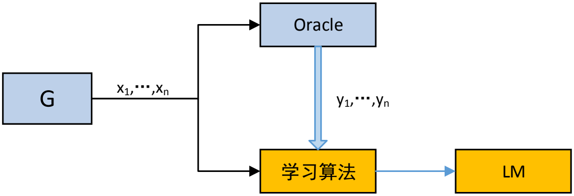
$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$$

- **参数化的经验风险**
  - 函数$f(\mathbf{x})$表示为参数化的形式$f(\mathbf{x}, \mathbf{w})$，泛函优化转化为参数优化：

$$\min_{\mathbf{w}} R_{emp}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w}))$$

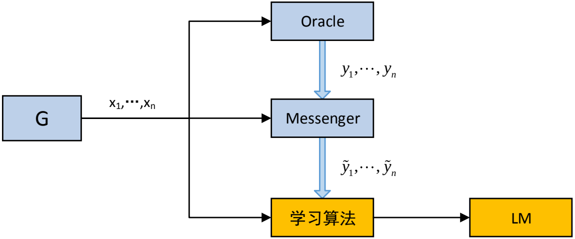
统计学习过程

- **训练样本集的产生**
  - 由学习的环境G产生样本：$\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
  - 由Oracle产生监督信息：$\mathcal{Y} = \{y_1, \cdots, y_n\}$



弱监督学习过程

- **训练样本集的产生**
  - 监督信息不是直接来自于Oracle
  - 监督信息是不完全、不完美的



弱监督学习的风险

- **弱监督学习的期望风险**

$$
\begin{aligned}
R(f) &= \int L(y, f(\mathbf{x}))p(\mathbf{x}, y, \tilde{y})d\mathbf{x}dyd\tilde{y} \\
&= \int L(y, f(\mathbf{x}))p(y|\mathbf{x}, \tilde{y})p(\mathbf{x}, \tilde{y})d\mathbf{x}dyd\tilde{y}
\end{aligned}
$$

  - $p(\mathbf{x}, \tilde{y})$：样本$\mathbf{x}$与弱监督标签的联合概率
  - $p(y|\mathbf{x}, \tilde{y})$：样本$\mathbf{x}$被标记为$\tilde{y}$，而其真实标签为$y$的概率

弱监督学习的风险

- **弱监督学习的经验风险**
  - 弱监督样本集$\{(\mathbf{x}_1, \tilde{y}_1), \cdots, (\mathbf{x}_n, \tilde{y}_n)\} \sim p(\mathbf{x}, \tilde{y})$
  - 对未知的真实监督信息$\{y_1, \cdots, y_n\}$取数学期望：

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}\int L(y_i, f(\mathbf{x}_i))p(y_i|\mathbf{x}_i, \tilde{y}_i)dy_i$$

  - 问题的关键是对$p(y_i|\mathbf{x}_i, \tilde{y}_i)$的估计
  - 可能的方法—EM算法

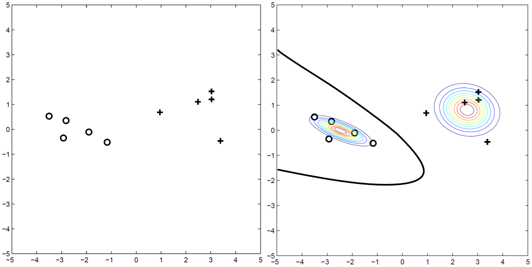
半监督学习

Semi-Supervised Learning

- **半监督学习问题**
  - 训练集包括两部分：
    - 有监督样本集：$L=\{(\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_l,y_l)\}$
    - 无监督样本集：$U=\{\mathbf{x}_{l+1},\cdots,\mathbf{x}_{l+u}\}$
  - 学习的目标：分类器函数$f(\mathbf{x})$

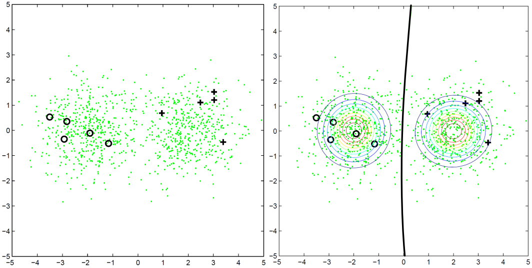
Semi-Supervised Learning

- **半监督学习问题**
  - 训练集包括两部分：
    - 有监督样本集：$L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$
    - 无监督样本集：$U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$
  - 学习的目标：分类器函数$f(\mathbf{x})$



Semi-Supervised Learning

- **半监督学习问题**
  - 训练集包括两部分：
    - 有监督样本集：$L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$
    - 无监督样本集：$U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$
  - 学习的目标：分类器函数$f(\mathbf{x})$



Self-Training

Algorithm 1 Self-Training

Input:

有监督样本集：$L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$
无监督样本集：$U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$

1: **repeat**
2: 利用选定的有监督学习算法在L上训练分类器f
3: 分类器f对U中无标记样本进行类别标注
4: 从U中选出子集S，$L = L \cup \{(\mathbf{x}, f(\mathbf{x})) | \mathbf{x} \in S\}$
5: **until** 收敛

Output: 分类器函数$f(\mathbf{x})$

Self-Training

- **Self-Training的EM算法**
  - 估计无监督样本的后验概率：

$$p(y_i|\mathbf{x}_i, \tilde{y}_i) = \delta(y_i = f(\mathbf{x}_i))$$

  - 重新学习分类器函数$f(\mathbf{x})$
- **子集S的选择**
  - 一般选择分类器$f(\mathbf{x})$认为可信度最高的样本
  - 例如，距离分类界面最远的样本
- **自学习的缺点**
  - 初始的分类器错误，容易被放大

Co-Training

Algorithm 2 Co-Training

Input:

有监督样本集：$L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$

无监督样本集：$U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$

1: 在自然分割的特征集$X = X^{(1)} \cup X^{(2)}$上用有监督样本集L学习两个分类器f_1和f_2

2: **repeat**

3: 分类器f_1和f_2分别从U中挑选样本子集E_1和E_2，并标注类别

4: $L_1 = L_1 \cup E_2, L_2 = L_2 \cup E_1$

5: 分别根据L_1和L_2重新学习分类器f_1和f_2

6: **until** 收敛

Output: 分类器$f(\mathbf{x})$为f_1和f_2的组合

Co-Training

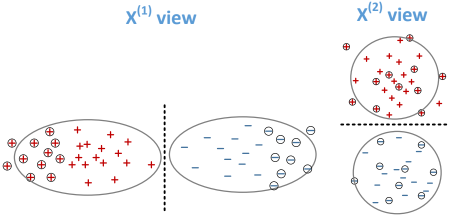
- **Co-Training的EM算法**
  - 估计无监督样本的后验概率：

$$p_1(y_i|\mathbf{x}_i^{(1)}, \tilde{y}_i) = \delta(y_i = f_2(\mathbf{x}_i^{(2)}))$$
$$p_2(y_i|\mathbf{x}_i^{(2)}, \tilde{y}_i) = \delta(y_i = f_1(\mathbf{x}_i^{(1)}))$$

  - 重新学习分类器函数$f_1(\mathbf{x}^{(1)}), f_2(\mathbf{x}^{(2)})$
- **子集的选择**
  - 子集E_1和E_2一般选择各自分类器认为可信度最高的样本给对方学习
  - 可以避免单个分类器错误的累积

Co-Training的特征分割

- **Co-Training的假设**
  - 存在特征集的一个分割$X = X^{(1)} \cup X^{(2)}$
  - $X^{(1)}$和$X^{(2)}$对于每个类别来说是条件独立的
  - 单独使用每个特征子集，都可以学习一个好的分类器
- **特征分割**
  - 例如，网页分类中图像特征和文本特征



Tri-Training

Algorithm 3 Tri-Training

Input:
有监督样本集：$L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$
无监督样本集：$U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$

1: Bootstrap抽样L，得到3个样本集分别学习三个分类器f_1, f_2, f_3

2: **repeat**

3: 挑选分类器f_i的样本子集：

$$L_i = \{\mathbf{x} | \mathbf{x} \in U, f_j(\mathbf{x}) = f_k(\mathbf{x})\} \quad j, k \neq i$$

4: 分别根据$L \cup L_i$重新学习分类器f_i

5: **until** 收敛

Output: 分类器$f(\mathbf{x})$为f_1、f_2和f_3的组合

Multiview learning

- **Tri-Training可以看作是一种多视学习**
  - 不做特征分割，使用全部特征学习
  - 使用不同的训练集，或者不同的模型学习多个分类器
  - majority vote融合多个分类器
- **多视学习的一般优化目标**

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

其中：$L(y, f(\mathbf{x}))$为损失函数；$||f||^2$为正则项

无监督样本的作用

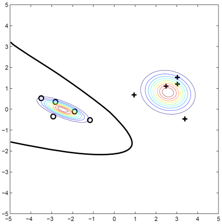
- **Generative Model**
  - 条件概率密度：$p(\mathbf{x}|y,\theta)$
  - 先验概率：$P(y|\boldsymbol{\pi})$
- **对数似然函数**

$$
\begin{aligned}
l(\theta,\boldsymbol{\pi}) &= \ln p(L,U|\theta,\boldsymbol{\pi}) \\
&= \sum_{i=1}^{l} \ln P(y_i)p(\mathbf{x}_i|y_i,\theta) + \sum_{i=l+1}^{l+u}\left[\sum_{y_i=1}^{C} \ln P(y_i)p(\mathbf{x}_i|y_i,\theta)\right]
\end{aligned}
$$

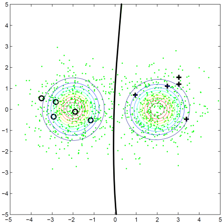
其中：$\boldsymbol{\pi} = (P(y=1),\cdots,P(y=C))^t$

产生式模型的学习—方法1

- **cluster-and-label**
  - 混合密度模型描述半监督样本的产生式模型
  - 使用数据集$L \cup U$估计混合密度模型的参数
  - 或者对数据集$L \cup U$聚类
  - 使用有标签样本集L标记每一个分量分布（或聚类）



$p(X_L|\theta)$



$p(X_L, X_U|\theta)$

产生式模型的学习—方法2

- **EM**
  - E步：估计所有样本的标签分布$q(y|\mathbf{x}_i)$

    有监督样本：$q(y|\mathbf{x}_i) = \delta(y = y_i)$

    无监督样本：$q(y|\mathbf{x}_i) \propto P(y)p(\mathbf{x}_i|y, \boldsymbol{\theta})$
  - M步：重新估计参数$\boldsymbol{\theta}$和$\boldsymbol{\pi}$

$$\max_{\boldsymbol{\theta},\boldsymbol{\pi}} l(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{l+u} \sum_{y=1}^{C} q(y|\mathbf{x}_i) \ln[P(y)p(\mathbf{x}_i|y, \boldsymbol{\theta})]$$

无监督样本的作用

- **Discriminative Model**
  - 样本分布：$p(\mathbf{x}|\boldsymbol{\mu})$
  - 后验分布：$P(y|\mathbf{x},\boldsymbol{\theta})$
  - 参数的独立性：$p(\boldsymbol{\theta},\boldsymbol{\mu}) = p(\boldsymbol{\theta})p(\boldsymbol{\mu})$
- **似然函数**

$$
\begin{aligned}
p(L,U|\boldsymbol{\theta},\boldsymbol{\mu}) &= p(X_L,Y_L,X_U|\boldsymbol{\theta},\boldsymbol{\mu}) \\
&= P(Y_L|X_L,X_U,\boldsymbol{\theta},\boldsymbol{\mu})p(X_L,X_U|\boldsymbol{\theta},\boldsymbol{\mu}) \\
&= P(Y_L|X_L,\boldsymbol{\theta})p(X_L,X_U|\boldsymbol{\mu})
\end{aligned}
$$

无监督样本对判别模型$P(y|\mathbf{x},\boldsymbol{\theta})$的学习，没有作用

正则化模型

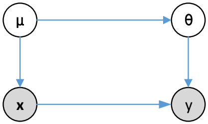
- **参数的先验依赖**
  - 引入参数μ和θ之间的先验依赖关系

$$p(\theta, \mu) = p(\theta|\mu)p(\mu)$$

- **θ关于U的后验分布**

$$p(\theta|U) = \int p(\theta|\mu)p(\mu|U)d\mu$$

- **无监督样本的作用**
  - 无监督样本U是通过$\mathbf{x}$的分布$p(\mathbf{x}|\mu)$，间接地影响后验概率$p(y|\mathbf{x})$的估计
  - 需要对依赖关系$p(\theta|\mu)$做出一定的假设



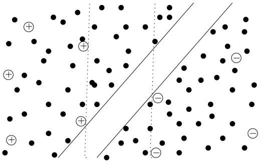
Transductive SVM

- **基本假设：** Low density separation
  - 判别界面通过样本分布的低密度区域
- **优化问题**

$$\min_f \sum_{i=1}^{l} (1 - y_i f(\mathbf{x}_i))_+ + \lambda_1 ||h||^2_{\mathcal{H}_k} + \lambda_2 \sum_{i=l+1}^{l+u} (1 - |f(\mathbf{x}_i)|)_+$$

其中：

$$f(\mathbf{x}) = h(\mathbf{x}) + b, \quad h \in \mathcal{H}_k$$



Graph-Based Method

- **两分类半监督学习**
  - 有标签样本：$X_L = \{\mathbf{x}_1, \cdots, \mathbf{x}_l\}, Y_L = \{y_1, \cdots, y_l\} \in \{0, 1\}^l$
  - 无标签样本：$X_U = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$
  - 全部样本$X_L \cup X_U$构造相似图，计算相似矩阵W
- **优化问题**
  - 令$f_i = P(y_i = 1)$，表示样本$\mathbf{x}_i$的标签为1的概率
  - 半监督的优化问题：

$$\min_{\mathbf{f}} E(\mathbf{f}) = \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2 = \mathbf{f}^t L \mathbf{f}$$

subject to

$$f_i = y_i, \quad i = 1, \cdots, l$$

优化问题求解

写成分块矩阵形式

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}_l \\ \mathbf{f}_u \end{bmatrix} \quad W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix} \quad L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$$

优化目标函数

$$\begin{aligned} E(\mathbf{f}) &= \begin{bmatrix} \mathbf{f}_l^t & \mathbf{f}_u^t \end{bmatrix} \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \begin{bmatrix} \mathbf{f}_l \\ \mathbf{f}_u \end{bmatrix} \\ &= \mathbf{f}_l^t L_{ll} \mathbf{f}_l + \mathbf{f}_u^t L_{ul} \mathbf{f}_l + \mathbf{f}_l^t L_{lu} \mathbf{f}_u + \mathbf{f}_u^t L_{uu} \mathbf{f}_u \end{aligned}$$

优化问题求解

目标函数对$\mathbf{f}_u$求极值

$$\begin{aligned}\frac{\partial E(\mathbf{f})}{\partial \mathbf{f}_u} &= (L_{ul} + L_{lu}^t)\mathbf{f}_l + 2L_{uu}\mathbf{f}_u \\ &= 2L_{ul}\mathbf{f}_l + 2L_{uu}\mathbf{f}_u \\ &= -2W_{ul}\mathbf{f}_l + 2L_{uu}\mathbf{f}_u \\ &= \mathbf{0}\end{aligned}$$

得到优化问题的解

$$\mathbf{f}_u = L_{uu}^{-1}W_{ul}\mathbf{f}_l$$

图方法的特点

- **转导学习方法**
  - 诱导学习（Inductive Learning）：目标是优化期望风险

$$R(f) = \int L(y, f(\mathbf{x}))dF(\mathbf{x}, y)$$

  - 转导学习（Transductive Learning）：目标是优化测试集上的风险

$$R_t\left(\{\hat{y}_{l+1}, \cdots, \hat{y}_{l+u}\}\right) = \frac{1}{u}\sum_{i=1}^{u} L(y_{l+i}, \hat{y}_{l+i})$$

- **Harmonic性**
  - 可以证明优化问题的解具有Harmonic特性

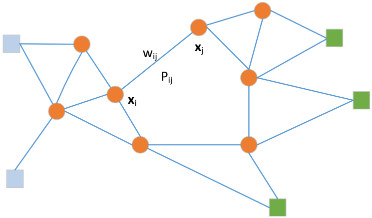
$$f_j = \frac{\sum_{i\sim j} w_{ij} f_i}{\sum_{i\sim j} w_{ij}}$$

图上的随机游走

- **随机游走过程**
  - 开始于一个无标签节点$\mathbf{x}_i$
  - 依据概率P_{ij}转移到节点$\mathbf{x}_j$，直到一个有标签节点为止

$$P_{ij} = \frac{w_{ij}}{\sum_{i \sim j} w_{ij}}$$

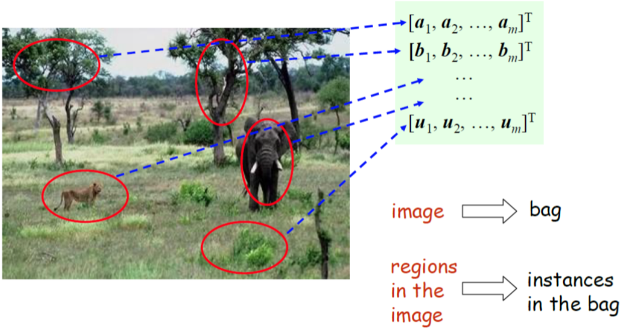
  - 将节点标签赋予$\mathbf{x}_i$
- **可以证明：** 解f_i是随机游走过程节点$\mathbf{x}_i$获得标签1的概率



其它弱监督学习问题

Multi-Instance Learning

- **问题的提出**
  - 识别对象以“示例包”的形式描述，而非示例
  - 示例包是示例的集合，每个示例以特征矢量描述
  - 已知示例包的类别，而每个示例的类别未知



MIL问题的表示

- **训练集**

$$D=\{\mathbf{B}_1^+,\cdots,\mathbf{B}_{m^+}^+,\mathbf{B}_1^-,\cdots,\mathbf{B}_{m^-}^-\}$$

  - 示例包：

$$\mathbf{B}_i=\{\mathbf{x}_{i1},\cdots,\mathbf{x}_{in_i}\}$$

  - 正例包：至少有一个示例是正例
  - 反例包：所有示例都是反例

- **分类问题**
  - 判别示例包A是正例包还是反例包
  - 判别示例$\mathbf{x}$是正例还是反例

MIL算法

- **Bag Based Methods**
  - 将示例包作为一个整体，看作是示例包空间中一个点
  - 方法1：将示例包空间视为度量空间，直接定义示例包之间的距离度量
  - 方法2：采用某种方法将示例包空间映射为欧氏空间，采用单示例分类器分类
- **Instance Based Methods**
  - 按照多示例的定义，利用示例包学习一个示例的分类器
  - 分类时，对每个示例进行分类
  - 根据示例的类别属性，分类示例包

Citation k-NN

Algorithm 4 Citation k-NN

1: 定义：示例包$\mathbf{A}$和$\mathbf{B}$之间的距离为Hausdoff距离

$$\mathsf{Dist}(\mathbf{A},\mathbf{B}) = \min_{\mathbf{a}\in\mathbf{A},\mathbf{b}\in\mathbf{B}} \mathsf{Dist}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{a}\in\mathbf{A},\mathbf{b}\in\mathbf{B}} ||\mathbf{a}-\mathbf{b}||$$

2: 在所有训练样本中计算待识别样本（示例包）$\mathbf{X}$的R近邻示例包

3: 将$\mathbf{X}$加入训练样本集，计算所有以$\mathbf{X}$为C近邻的示例包

4: 计算上述所有示例包中正例包和反例包的数量，选择数量多者作为$\mathbf{X}$的类别

MI-SVM

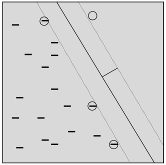
- 求解优化问题

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_I \xi_I$$

subject to

$$\forall I: Y_I \max_{i \in I} (\mathbf{w}^t \mathbf{x}_i + b) \geq 1 - \xi_I, \quad \xi_I \geq 0$$

  - Y_I为示例包I的类别标签
  - 正例包：与分类界面之间的“间隔”为最远示例的间隔
  - 反例包：与分类界面之间的“间隔”为最近示例的间隔



其它的弱监督学习问题

- **Multi-Label Learning**
  - 每个示例有多个标签
- **Multi-Instance Multi-Label Learning**
  - 每个示例包有多个标签
- **Multi-Instance Semi-Supervised Learning**
  - 半监督的多示例学习问题
- **Imperfect Oracle**
  - 每个示例由多个标注者给出标签
  - 每一个标注者对不同的示例给出标签