目录

0.1 生成式方法 **2**
0.2 半监督 SVM **3**
0.3 图半监督学习 **3**
0.4 基于分歧的方法 **4**
0.5 半监督聚类 **4**

第十三章　半监督学习

主动学习：在有标记数据比较少的情况下，主动向专家询问一些样本的标记加入训练集中，再对算法进行训练。

半监督学习 (semi-supervised learning)：学习器不依赖外界交互、自动地利用未标记样本来提升学习性能。半监督学习可分为**纯半监督学习** (pure semi-supervised learning) 和**直推学习** (transductive learning)。纯半监督学习假定训练数据中的未标记样本并非待预测的数据，也就是希望模型能够适用训练过程中未观察到的数据；直推学习假定未标记样本就是待预测数据，学习的目的就是在这些未标记样本上获得最优的泛化性能。

要利用未标记样本，就要对未标记样本做一些关于数据分布信息和类别标记联系的假设。假设分为两种：**聚类假设** (cluster assumption) 和**流形假设** (manifold assumption)。聚类假设考虑的是类别标记，通常用于分类任务，它假设数据存在簇结构，同一个簇的样本属于同一个类别。流形假设对输出值没有限制，它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。无论是聚类假设还是流形假设都是基于“相似的样本拥有相似的输出”这个基本的假设。

半监督学习方法主要有四种：**生成式方法、半监督 SVM、图半监督、基于分歧**的方法。其中前三种针对单学习器，最后一种是多学习器。

0.1　生成式方法

生成式方法 (generative method) 是直接基于生成式模型的方法，即无论是标记数据还是未标记数据都是由同一个模型“生成”的。把未标记数据看作是潜在模型的缺失参数，通过**基于 EM 算法进行极大似然估计**求解。

给定有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_l, y_l)\}$ 和未标记样本集 $D_u = \{(\boldsymbol{x}_{l+1}, y_{l+1}), (\boldsymbol{x}_{l+2}, y_{l+2}), \cdots, (\boldsymbol{x}_{l+u}, y_{l+u})\}, l \ll u, l+u=m$。假设所有样本独立同分布，且都是由同一个**高斯混合模型**生成的，用极大似然估计来估计高斯混合模型的参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) | 1 \le i \le N\}$，其中 N 是样本类别标记个数，$y_i \in \boldsymbol{\mathcal{Y}}, |\boldsymbol{\mathcal{Y}}| = N$。每个类别对应一个高斯混合模型，因此数据样本的概率密度是：

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

$\Theta \in \{1, 2, \cdots, N\}$ 表示样本 $\boldsymbol{x}$ 隶属的高斯混合成分，$D_l \cup D_u$ 的对数似然是：

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

由有标记数据 D_l 的监督项和未标记数据的 D_u 无监督项组成。

用 EM 算法迭代更新求解时和第九章比较：第九章是无监督学习，完全没有标注信息；而此处是半监督学习，有部分样本已标注。

高斯混合模型还可以替换成混合专家模型和朴素贝叶斯模型。

0.2 半监督 SVM

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。其中最著名的是针对二分类问题的 TSVM(Transductive Support Vector Machine)。

给定有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_l, y_l)\}$ 和未标记样本集 $D_u = \{(\boldsymbol{x}_{l+1}, y_{l+1}), (\boldsymbol{x}_{l+2}, y_{l+2}), \cdots, (\boldsymbol{x}_{l+u}, y_{l+u})\}, l \ll u, l+u=m, y_i \in \{-1, +1\}$。TSVM 的学习目标是为 D_u 中的样本给出预测标记 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \cdots, \hat{y}_{l+u}), \hat{y}_i \in \{-1, +1\}$，使得：

$$\begin{aligned} \min_{\boldsymbol{w}, b, \hat{\boldsymbol{y}}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, i = 1, 2, \cdots, l, \\ & \hat{y}_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, i = l+1, l+2, \cdots, m, \\ & \xi_i \geq 0, i = 1, 2, \cdots, m, \end{aligned}$$

其中 $(\boldsymbol{w}, b)$ 确定了划分超平面；$\boldsymbol{\xi}$ 为松弛向量，$\xi_i (i = 1, 2, \cdots, l)$ 对应于有标记样本，$\xi_i (i = l+1, l+2, \cdots, m)$ 对应于未标记样本；C_l, C_u 是指定的用于平衡模型复杂度、有标记样本与未标记样本重要程度的折中参数。相当于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 已知，求 $\boldsymbol{w}, b, \boldsymbol{\xi}$。

为什么说 $\hat{\boldsymbol{y}}$ 是已知呢？因为 TSVM 采用局部搜索来迭代求解上式的近似解。先用有标记样本学得一个 SVM，然后用这个 SVM 来预测未标记样本得到“伪标记”，从而 $\hat{\boldsymbol{y}}$ 已知。此时剩下的就是一个标准的 SVM 问题了。

0.3 图半监督学习

图半监督学习就是将数据集的样本看做图中的节点，把样本的相似程度看成图中节点之间边的强度，把标记过的样本染上相应的颜色，半监督学习就是颜色在图上的扩散传播过程。

给定有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_l, y_l)\}$ 和未标记样本集 $D_u = \{(\boldsymbol{x}_{l+1}, y_{l+1}), (\boldsymbol{x}_{l+2}, y_{l+2}), \cdots, (\boldsymbol{x}_{l+u}, y_{l+u})\}, l \ll u, l+u=m$。基于 $D_l \cup D_u$ 构建图 $G = (V, E)$，其中结点集 $V = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \cdots, \boldsymbol{x}_{l+u}\}$，边集通过高斯函数 (径向基函数) 定义：

$$(\boldsymbol{W})_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

令学习器学的函数 $f: V \to \mathbb{R}$，对应的分类规则为 $y_i = \text{sign}(f(\boldsymbol{x}_i)), y_i \in \{-1, +1\}$，通过最小化能量函数：

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

其中 $\boldsymbol{D} = \text{diag}(d_1, d_2, \cdots, d_{l+u})$ 是一个对角矩阵，$d_i = \sum_{j=1}^{l+u} (\boldsymbol{W})_{ij}$ 为矩阵 $\boldsymbol{W}$ 的第 i 行元素之和，$\boldsymbol{W}$ 是对称矩阵。

通过对 $E(f)$ 求 $\boldsymbol{f}_u$ 的偏导，并令其等于 0，可求得：

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

而 $\boldsymbol{f}_l$ 通过已标记样本已知，所以 $\boldsymbol{f}_u$ 可求。

上述是二分类问题的标记传播，对于多分类问题的标记传播，参考论文[Zhou et al., 2004]。简单来讲就是通过构造标记传播矩阵 $\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W} \boldsymbol{D}^{-\frac{1}{2}}$，然后迭代计算预测函数：

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

0.4 基于分歧的方法

把每个属性集看做一个视图，不同的视图具相同性 (关于输出空间 $\mathcal{Y}$ 是一致的)，协同训练。

首先在每个视图上基于标记样本训练出分类器，然后每个分类器分别挑选最有把握的未标记样本赋予伪标记，并将伪标记样本提供给另一个分类器作为新增的标记样本进行训练更新，以此迭代。

0.5 半监督聚类

聚类任务通过标记样本可以获得必连约束集合 $\mathcal{M}$ 和勿连约束集合 $\mathcal{C}$，前者是指样本属于同一簇中，后者指样本必不属于同一簇中。

半监督聚类和一般聚类的区别在于：在聚类的时候对于 $\boldsymbol{x}_i$，查看它能否在不违背 $\mathcal{M},\mathcal{C}$ 的约束条件下，将其放在最近的某个聚类簇中，若不能则错误；若能，放进相应聚类簇，并查看下一个样本。

还可以通过少量的标记样本，获得初始化的 k-means 算法的 k 个聚类中心，然后再聚类迭代。但是这样显然有一种风险：少量的标记样本未必包含所有的标记信息，那么没被包含的标记信息就可能被忽略。