Author：Liu Jian

Time：2020-07-01

机器学习12-半监督学习

1 基本概念

半监督学习 (semi-supervised learning) 就是不仅利用有标记的数据 (含有示例 $\boldsymbol{x}$ 及其标记 y) 进行学习，还利用大量未标记的示例 (即只有 $\boldsymbol{x}$) 进行学习。显然，若要利用未标记的样本，必须要作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，比如：(1) 聚类假设 (cluster assumption)，即假设数据存在簇结构，同一个簇的样本属于同一个类别；(2) 流形假设 (manifold assumption)，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，等等。可以看到，这些假设都是基于“相似的样本具有相似的输出”这一想法。

按照未标记的样本是否是学习器要预测的样本，半监督学习又可以分为：(1) 纯 (pure) 半监督学习，学习过程中的未标记数据并非待预测数据；(2) 直推学习 (transductive learning)，待预测数据就是学习时的未标记数据。

下面我们记有标记 (label) 的数据为 $D_l = \{(\boldsymbol{x}_i, y_i)\}_1^l$ ，未标记 (unlabel) 的数据为 $D_u = \{\boldsymbol{x}_i\}_{l+1}^N$ ，$N = l + u$ ，总样本为 $D = D_l \cup D_u$

半监督学习的四大范式 (paradigm)：基于分歧的方法、半监督 SVM、图半监督学习、生成式半监督学习。

> 主动学习 (active learning)：先基于有标记数据 D_l 训练学习器，接着不断从未标记数据 D_u 中挑选对改善模型性能帮助最大的样本点 (比如，基于 D_l 训练一个 SVM，挑选距离分类超平面最近的未标记样本，其目的是尽量使用少的查询来获得尽量好的性能)，查询该样本点的标记，更新学习器。

> 可以看到，和半监督学习不同，主动学习中我们是可以得到未标记样本的标记的，虽然使用未标记样本进行训练，但实际上是一个监督学习；但和传统的监督学习不同的是，我们不是使用全部的未标记样本进行学习，而是每次从未标记样本中挑选一个对改善学习器性能帮助最大的样本点进行学习，更新学习器，而这一过程又有点类似于在线学习。

2 生成式方法 - 以 GMM 为例

具有 M 个混合成分的高斯混合模型：

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \Sigma_m)$$

$$\text{with} \quad \sum_{m=1}^{M} \alpha_m = 1, \ \alpha_m > 0 \ \text{ for all } m$$

假设混合成分对应了样本的标记，那么高斯混合模型实际上就是一个如下的生成模型：

$$p(\boldsymbol{x}, y) = p(y)p(\boldsymbol{x}|y)$$
$$\text{with} \ \ p(y = m) = \alpha_m, \ p(\boldsymbol{x}|y = m) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \Sigma_m)$$

对于有标记样本 D_l ，我们最小化对数损失：

$$\min \sum_{i=1}^{l} -\log p(\boldsymbol{x}_i, y_i)$$

对于有标记样本 D_u ，我们最小化对数损失：

$$\min \sum_{i=l+1}^{l+u} -\log p(\boldsymbol{x}_i)$$

因此，半监督学习最小化它们的总体损失：

$$\min \sum_{i=1}^{l} -\log p(\boldsymbol{x}_i, y_i) + \sum_{i=l+1}^{l+u} -\log p(\boldsymbol{x}_i)$$

对于上述优化问题，西瓜书上说采用 EM 算法求解，但个人认为，从 EM 算法的导出过程可知，它并不适用于上述问题的求解，西瓜书上给出的只是一个类似于 EM 算法的迭代计算公式而已，或者说直接借鉴了 GMM 原 EM 迭代公式中的 E 步，并在此基础上构建新的迭代公式，对应了原 EM 算法的 M 步，但其基本原理与 EM 算法无关。

将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。此类方法简单、易于实现，在有标记数量极少的情形下往往比其他方法性能更好，但前提条件是假设的生成式模型必须与真实数据分布吻合。

3 半监督 SVM

半监督支持向量机 (Semi-Supervised Support Vector Machine, S3VM) 是 SVM 在半监督学习上的推广，在不考虑未标记样本时，SVM 试图找到最大间隔划分超平面，而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 (即低密度分隔，low-density separation)。 S3VM 中最著名的是针对二分类问题的 TSVM (Transductive Support Vection Machine)。

TSVM 尝试将每个未标记样本分别作为正例或反例即标记指派 (label assignment)，然后在标记指派的所有结果中，寻找一个在所有样本上间隔最大的划分超平面。对于每一种标记指派 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \cdots, \hat{y}_{l+u}), \ \hat{y}_i \in \{-1, +1\}$ ，求解此时的间隔最大划分超平面在数学上即为求解如下的优化问题：

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

其中， C_l 与 C_u 是由用户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。实际中，我们不可能穷举所有的标记指派并进行考察。为此，TSVM 采用如下的迭代策略：(1) 先基于有标记样本 D_l 训练一个 SVM，再基于此 SVM 对未标记数据 D_u 进行标记指派；(2) 接着，基于第一步得到的标记指派，求解上述优化问题，得到新的 SVM。注意，在生成新的 SVM 时，因为此时的标记指派很可能不准确，因此 C_l 要设置得比 C_u 大，使有标记样本所起作用更大；(3) 迭代计算：在未标记样本此时的样本指派中寻找被当前 SVM 错误分类的一个当前指派为正例的样本 $\boldsymbol{x}_i$ (训练当前 SVM 时，$\hat{y}_i = +1$ ，但反过来使用训练后的 SVM 进行预测时，其标记为 -1 ，即训练时其对应的 $\zeta_i > 0$) 和一个当前指派为负例的样本 $\boldsymbol{x}_j$ (训练当前 SVM 时， $\hat{y}_j = -1$ ，但反过来使用训练后的 SVM 进行预测时，其标记为 $+1$ ，即训练时其对应的 $\zeta_j > 0$)。此外，为了使找到的 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 尽可能是当前指派与真实标记不同的示例 (即它们的当前指派与当前 SVM 给出的结果不同，但当前 SVM 给出的很有可能是它们的真实标记)，我们还需使找的这两个示例满足 $\zeta_i + \zeta_j > 2$ 。找到这两个示例后，交换其当前样本指派结果，得到新的标记指派，并训练新的 SVM；新 SVM 的训练过程中，我们还需增大 C_u 的值以提高未标记样本对优化目标的影响。重复进行上述过程，直至 $C_u = C_l$ 为止。

在对未标记样本进行标记指派及调整过程中，未标记样本有可能出现类别不平衡问题，即标记指派的某类样本远多于另一类。为了减轻类别不平衡造成的不利影响，可将优化目标中的 C_u 项拆分为 C_u^+ 和 C_u^- 两项，它们分别对应未标记样本中指派为正例 (个数记为 u_+) 与负例 (个数记为 u_-) 的项，并在初始化时，令

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

4 图半监督学习

与概率图模型借助图模型表示随机变量间的相关关系与独立关系不同，从下面的描述中我们可以看到，图半监督学习与图模型其实没有什么太大关系，不借助图模型也完全可以。个人认为，图半监督学习倒不如说是原理上基于距离、形式上借助矩阵的半监督学习。

4.1 二分类问题的图半监督学习

样本 $D = D_l \cup D_u$ 中，每个示例 $\boldsymbol{x}_i$ 就对应一个结点，结点间的距离即边长则表征了结点间的相似程度，边长越短越相似。但这里，我们将边长定义为距离的减函数，即边长越大，说明两个结点越相似，结点间的相似程度可用一个亲和矩阵 (affinity matrix) $W_{N\times N}$ 存储，常基于高斯函数定义为：

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

其中， σ 为给定的高斯函数带宽参数。

记我们要学习的函数为 $f: \mathbb{X} \to \mathbb{R}$ ，最终的分类规则为 $y = sign(f(\boldsymbol{x})) \in \{-1, +1\}$ 。记 f 在已标记和未标记样本上的预测向量分别为 $\boldsymbol{f}_l$ 和 $\boldsymbol{f}_u$ ，并记 $\boldsymbol{f} = \langle \boldsymbol{f}_l; \boldsymbol{f}_u \rangle$ 。令 $\boldsymbol{f}_l$ 等于其真实标记 $\langle y_1; \cdots; y_l \rangle$ ，我们的策略是，两个样本点 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 越近似，其预测 $f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)$ 也越接近，即我们要最小化如下的称之为能量函数 (energy function) 的东西：

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

其中， $D = diag(d_1, \cdots, d_N)$ ， $d_i = \sum_{j=1}^{N}[W]_{ij} = \sum_{j=1}^{N}[W]_{ji}$ 。可以看到，其中优化变量为 $\boldsymbol{f}_u$ ，基于分块矩阵的运算法则，求导可推得最优解如下：

$$\boldsymbol{f}_u = (I - P_{uu})^{-1} P_{ul} \boldsymbol{f}_l$$

where

$$P_{uu} = D_{uu}^{-1} W_{uu},\ P_{ul} = D_{uu}^{-1} W_{ul}$$

$$D = \begin{bmatrix} D_{ll} & 0_{l\times u} \\ 0_{u\times l} & D_{uu} \end{bmatrix},\ W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$$

上述由 $\boldsymbol{f}_l$ 得到 $\boldsymbol{f}_u$ 的过程称之为标记传播 (label propagation)，我们也可以将其推广到多分类问题上。

4.1 多分类问题的图半监督学习

假设为 M 分类问题，和二分类问题中不同的是，此时我们要求的不再是一个向量 $\boldsymbol{f}$，而是一个 $N \times M$ 阶的标记矩阵 F，$[F]_{ij}$ 表征了第 i 个样本点属于第 j 类的可能性大小。

西瓜书上直接给出了标记传播的迭代公式：

$$[F^{(0)}]_{ij} = \begin{cases} 1, & (1 \leqslant i \leqslant l) \wedge (y_i = j) \\ 0, & \text{otherwise} \end{cases}$$

$$F^{(t+1)} = \alpha S F^{(t)} + (1-\alpha) F^{(0)}$$

其中，$\alpha \in (0,1)$ 为给定参数，$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$。事实上，上述迭代公式的收敛点为：

$$F^* = \lim_{t \to +\infty} F^{(t)} = (1-\alpha)(I - \alpha S)^{-1} F^{(0)}$$

由此我们可得未标记样本的标记：

$$y_i = \arg \max_{1 \leqslant j \leqslant M} [F^*]_{ij},\ \ i = l+1, \cdots, l+u$$

接着，西瓜书上又指出，上述收敛点是如下的正则化问题的最优解：

$$\min_F \frac{1}{2} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} [W]_{ij} \left\| \frac{1}{\sqrt{d_i}} F_i - \frac{1}{\sqrt{d_j}} F_j \right\|^2 \right) + \mu \sum_{i=1}^{l+u} \| F_i - F_i^{(0)} \|^2$$

其中，F_i 和 $F_i^{(0)}$ 分别为矩阵 F 和 $F^{(0)}$ 的第 i 个行向量，$\mu > 0$ 为正则化参数。当 $\mu = (1-\alpha)/\alpha$ 时，上述正则化问题的最优解即为 F^*。

西瓜书上按照这个顺序介绍的原因可能是因为迭代算法是先于正则化问题被提出的，迭代算法提出后，人们又发现上述正则化问题的解恰为迭代的收敛点。

图半监督学习的两个缺点：

1. 存储开销大，使得此类算法很难直接处理大规模数据
2. 可以看到，我们并没有得到映射 f，只是得到了未标记样本的标记 $sign(\boldsymbol{f}_u)$；而对于新的样本，我们却无法直接对其进行预测，而只能：
   - 将新接受的样本加入原数据集中重新算一遍；
   - 或者，引入额外的预测机制，比如基于图半监督学习的结果训练一个 SVM 以预测新样本。

5 基于分歧的方法

基于分歧 (disagreement，亦称 diversity) 的方法，思路和集成学习有些类似，即基于有标记样本，学习多个学习器，将这些学习器用于未标记样本的预测，选择各学习器中分类置信度高的样本 (比如，若学习器为朴素贝叶斯分类器，则置信度为后验概率；若学习器为 SVM，则置信度为样本点到分类超平面的距离) 和其学习器预测标记，加入到有标记样本中，再基于扩充的有标记样本学习多个学习器，并不断迭

代进行下去。

可以看到，由于各学习器之间的差异性，各学习器在某些样本点上的预测可能比较靠谱，而在另一些样本点上的预测可能就差一点，我们综合多个学习器那些靠谱的结果，扩充有标记样本，不断迭代更新各学习器，直到结束。这一过程中，各学习器取长补短，互相学习，共同进步，因此，个人认为这类方法称之为基于差异或多样性的方法会更贴切，因为我们在意的并不是各学习器对同一样本点的不同预测 (即分歧)，而是各学习器在各自擅长的样本点上的预测，利用的是各学习器性能上的互补，即多样性或者说差异。

6 半监督聚类

半监督聚类 (semi-supervised clustering) 就是利用一些监督信息辅助聚类。监督信息分为如下两种：

1. 必连 (must-link) 与勿连 (cannot-link) 约束。前者指两个样本必属于同一个簇，后者指两个样本必不属于同一个簇 (注意，这些簇所对应的标记是未知的)，代表算法为约束 k 均值 (constrained k-means) 算法。
2. 少量有标记的样本，代表算法为约束种子 k 均值 (constrained seed k-means) 算法。

约束 k 均值算法是 k 均值算法的扩展，实际上就是 k 均值算法在进行时多了判断和保证必连与勿连约束的过程。

约束种子 k 均值算法也是 k 均值算法的扩展，改变在于：(1) k 均值算法的 k 个聚类中心不是随机选取，而是由有标记样本给出；(2) 在聚类的迭代更新过程中，不改变有标记样本的簇隶属关系。
