半监督生成式方法介绍

【参考资料】

周志华《机器学习》

南瓜书 半监督学习

李宏毅 机器学习课程 半监督学习

半监督生成式方法假设所有样本独立同分布，标记样本和未标记样本都是由同一个生成模型生成的。以高斯混合模型为例，给定样本x，其真实类别标记$y \in \mathcal{Y}$，其中$\mathcal{Y} = \{1, 2, \ldots, N\}$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成分，即

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

其中，混合系数$\alpha_i \geqslant 0$，$\sum_{i=1}^{N} \alpha_i = 1$；$p\left(\boldsymbol{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$是样本$x$属于第$i$个高斯混合成分的概率；$\mu_i$和$\boldsymbol{\Sigma}_i$为该高斯混合成分的参数。

现在给定标记样本集合$D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$和未标记样本集合$D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$，$l \ll u$，$l + u = m$，并且假设它们都是独立同分布的。用极大似然法来估计高斯混合模型的参数$\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) | 1 \leqslant i \leqslant N\}$，$D_l \cup D_u$的对数似然是

$$\begin{aligned} LL\left(D_l \cup D_u\right) = & \sum_{(x_j, y_j) \in D_l} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot p\left(y_j | \Theta = i, \boldsymbol{x}_j\right)\right) \\ & + \sum_{x_j \in D_u} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right) \end{aligned} \tag{2}$$

式（2）中的第一项是有标记数据的特征和标签的联合概率分布$P(x, y)$，第二项是无标记数据特征的概率分布$P(x)$。

高斯混合模型的参数估计需要使用EM算法求解，迭代更新过程如下：

- E步：根据当前模型参数计算未标记样本$\boldsymbol{x}_j$属于各高斯混合成分的后验概率

$$\gamma_{ji} = \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \tag{3}$$

- 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

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j\right) \tag{4}$$

$$\begin{aligned}\boldsymbol{\Sigma}_i = & \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i \right) \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i \right)^{\mathrm{T}} \right. \\ & \left. + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i \right) \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i \right)^{\mathrm{T}} \right)\end{aligned} \tag{5}$$

$$\alpha_i = \frac{1}{m} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) \tag{6}$$

参数更新的过程相当直观，实际上是用无标记样本对原来的、仅包含有标记样本的参数更新公式进行修正：

式（4）求均值μ，括号内的第二项是对带标记的样本中，类别为i的样本特征x_j求和，即$\sum_{(x_j,y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j$，第一项则是无标记样本特征$x_j$的加权和，而权重就是无标记样本属于类别$i$的概率，即$\sum_{x_i \in D_y} \gamma_{ji} \boldsymbol{x}_j$。然后除以的是类别为$i$的有标记样本数目，加上无标记样本属于类别$i$的概率的和。

式（5）同理。

式（6）求解的α即为类别的先验概率，类似地，用类别为i的有标记样本数目，加上无标记样本属于类别i的概率的和，去除以样本的总数m。

要求解后验概率，即式（3），可以通过有标记数据来对模型参数进行初始化，具体来说：

$$\begin{aligned}\alpha_i &= \frac{l_i}{|D_l|}, \text{ where } |D_l| = \sum_{i=1}^{N} l_i \\ \mu_i &= \frac{1}{l_i} \sum_{(x_j,y_j) \in D_l \wedge y_i = i} \left(x_j - \mu_j \right) \left(x_j - \mu_j \right)^T \\ \Sigma_i &= \frac{1}{l_i} \sum_{(x_j,y_j) \in Dl \wedge yj = i} \left(x_j - \mu_i \right) \left(x_j - \mu_i \right)^\top\end{aligned}$$

注意上面三个式子与在有无标记数据参与下的参数求解式（4）、（5）、（6）的对比。

详细的推导过程请见：

https://datawhalechina.github.io/pumpkin-book/#/chapter13/chapter13

如果使用其他生成模型来对数据分布进行建模，那么就会导出不同的半监督生成式方法。

半监督的生成式方法在**有标记数据极少**的情形下往往会比其他方法性能更好。然而，这类方法的关键在于，**模型假设必须准确**，即假设的生成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。但是在实际中很难作出正确的模型假设，所以生成式方法具有很大的局限性。