天善智能
TIANSHAN SOFT

风控模型中的前沿问题一：标签缺失的处理

讲师：安迪生

课程简介

- 有监督模型是风控建模中的主要方法，它具有覆盖场景丰富、结果精度高等特点。但是建立有监督的模型的同时，模型对样本的标签的质量是有要求的。然而往往在现实工作中，并不能保证样本的标签一定会充足或者准确。此时应该如何处理标签缺失的问题呢？

目录

什么是标签缺失

- 什么是标签缺失

在分类场景中，目标变量是样本所属的类别，通常称之为标签。在开发模型之前需要检查数据的质量，其中一个很重要的检查项目就是检查标签是否完整。实际工作中标签的不完整是经常遇到的情形。常见的标签缺失情形有：

一．极少部分样本的标签是缺失的，且属于完全随机缺失

二．大部分样本的标签是缺失的

三．某些特定类型的标签是缺失的

针对不同的缺失情形，对数据乃至模型的处理是不同的。第一种缺失场景由于占比很少、机制随机，因此可以将标签缺失的样本删除后进行模型开发。我们重点讨论第二、三种场景。

什么是标签缺失

- 标签缺失的机制

在研究解决标签缺失之前，需要梳理一下标签缺失的机制。除了完全随机缺失之外，标签缺失通常有两大类原因：

标签分类工作的成本太高

例如，在申请反欺诈中需要识别申请人是否涉嫌团队欺诈。此时需要对申请人的所有联系人进行识别。由于需要识别的对象太多，带来的时间、人力和经济成本太高，因此一般不会对所有的客群进行识别，而是进行抽样。这种情况下，我们只能对一部分的样本打标签，剩下的样本的标签形成了缺失。

特点：标签非缺失的样本占比通常较少，标签缺失是随机的。

标签分类的策略不完善

例如，在交易反欺诈中，可以通过一定的规则识别信用卡盗刷，也可以通过持卡人报警判断盗刷。但是某些盗刷的行为在没有命中规则的时候是无法识别的。在这种情形下，被打上盗刷标签的就是盗刷行为，而没有被打上盗刷标签的未必都是正常的刷卡交易。

特点：某种类别的标签存在缺失，而另一种不缺失。

什么是标签缺失

- 标签缺失带来的影响

➢ 如果样本集中大量样本的正负标签存在缺失，则对于标签缺失的样本无法建立分类模型。只用标签非缺失的样本建立分类模型时：

若总的样本量较少时，可用样本量就更少，不利于开发出稳定的模型。

若标签缺失与非缺失的样本的分布不一致，则开发出来的模型是“有偏”的。

例如，10000个样本中只有2000个样本的标签是已知的。此时如果只用2000个样本建模势必造成样本量过小从而得到的模型的稳定性受到影响。

➢ 如果样本集中，某一种特定的标签存在缺失则会对剩余的标签产生干扰，从而无法建立正确的模型。

例如，10000个样本中，有1000个样本被识别出是欺诈样本，剩下的样本中无法准确判断是欺诈样本还是正常样本。如果剩余的9000个样本被看成是正常样本，则得到的模型也会产生错误。

标签缺失的处理方法

- 标签缺失的处理方法（一）

上述两种标签缺失的情形有不同的处理方法。对于第一种情形，即大部分样本的正负标签存在缺失，可以使用半监督聚类的方式，将标签缺失的样本进行补全，从而可以近似地得到所有样本的正负标签。

首先介绍无监督K-均值聚类基本概念

一种无监督的分类算法，目的是为了将样本集划分成若干的簇，使得每个簇内的成员尽可能地相近，而簇间的样本尽可能地相远。

对于给定的样本集$D = \{x_1, x_2, \dots, x_m\}$, k-均值算法针对聚类所得的簇划分$C = \{C_1, C_2, \dots, C_k\}$最小化平方误差

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left|\left|x - \mu_i\right|\right|_2^2$$

其中$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$是簇$C_i$的均值向量，也即几何中心。

E越小，表明簇内的成员的相近程度很高。

标签缺失的处理方法

- K-均值聚类的步骤

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇数 k.

过程:

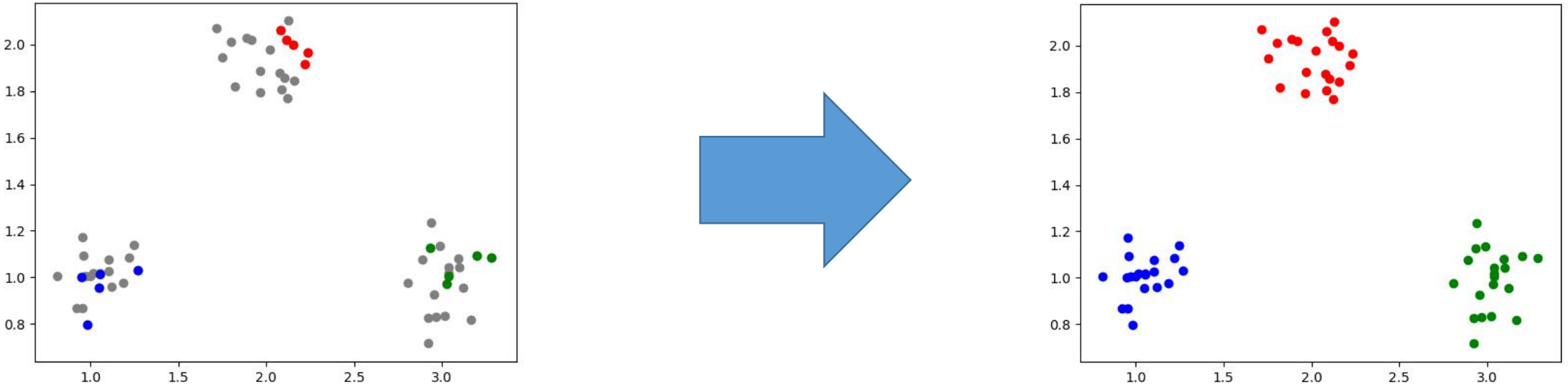
1: 从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$
2: **repeat**
3: 令 $C_i = \varnothing \ (1 \leqslant i \leqslant k)$
4: **for** $j = 1, 2, \ldots, m$ **do**
5: 计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i \ (1 \leqslant i \leqslant k)$ 的距离: $d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||_2$;
6: 根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ji}$;
7: 将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$;
8: **end for**
9: **for** $i = 1, 2, \ldots, k$ **do**
10: 计算新均值向量: $\boldsymbol{\mu}'_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$;
11: **if** $\boldsymbol{\mu}'_i \neq \boldsymbol{\mu}_i$ **then**
12: 将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}'_i$
13: **else**
14: 保持当前均值向量不变
15: **end if**
16: **end for**
17: **until** 当前均值向量均未更新

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

标签缺失的处理方法

- 半监督聚类法

半监督聚类法是基于k-均值聚类的一种拓展。给定样本集$D = \{x_1, x_2, \ldots, x_m\}$, 假设少量样本有标记$S = \cup_{j=1}^{k} S_j$, 其中$S_j \neq \emptyset$为属于第j个标签的样本集。因此可以直接将该部分样本作为种子样本来初始化k-均值算法的中心。在聚类簇迭代更新过程中不改变种子的隶属关系。



标签缺失的处理方法

- 基于k-均值的半监督聚类法算法

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
少量有标记样本 $S = \bigcup_{j=1}^{k} S_j$;
聚类簇数 k.

过程:

```
for j = 1, 2, ..., k do
    μ_j = (1/|S_j|) Σ_{x∈S_j} x
end for
repeat
    C_j = ∅ (1 ⩽ j ⩽ k);
    for j = 1, 2, ..., k do
        for all x ∈ S_j do
            C_j = C_j ⋃ {x}
        end for
    end for
    for all x_i ∈ D \ S do
        计算样本 x_i 与各均值向量 μ_j (1 ⩽ j ⩽ k) 的距离: d_ij = ||x_i − μ_j||_2;
        找出与样本 x_i 距离最近的簇: r = arg min_{j∈{1,2,...,k}} d_ij;
        将样本 x_i 划入相应的簇: C_r = C_r ⋃ {x_i}
    end for
    for j = 1, 2, ..., k do
        μ_j = (1/|C_j|) Σ_{x∈C_j} x;
    end for
until 均值向量均未更新
```

输出: 簇划分 $\{C_1, C_2, \ldots, C_k\}$

标签缺失的处理方法

- 基于半监督k-均值聚类的标签补充方法的优缺点

优点

- 算法简单，易于实施
- 对数据质量要求不高，不对数据的分布做要求

缺点

- 不同类别的样本要尽可能地分开
- 对非平衡样本较敏感（但是可以改进！）
- 精度不高
- 对于有标记的样本，要求包含每一种可能的类别（即每一种类别都需要有样本能观测到其标签）

标签缺失的处理方法

- 标签缺失的处理方法（二）

另一种经常出现的情形是某一种特定的标签存在部分缺失。例如部分坏样本不能被识别出来，造成好坏混合的情形，通常称之为“Positive and Unlabeled Data”。对于PU问题，有多重算法可以进行标签补充。

方法一：Charles E., Keith N. 2008. Learning Classifiers from Only Positive and Unlabeled Data

基本思想：假设正例中每个样本有相同的概率被识别出来。则Charles et al设计的分类器与基于带标签的正例P和不带标签的样本U所训练出来的分类器的关系是：

$$f(x) = \frac{g(x)}{c}$$

其中:

f(x)是真实的以概率表示的分类器(例如逻辑回归)，g(x)是基于P和U训练出来的分类器，c是一个常数且与P、U相关

- 优点：计算简单
- 缺点：“每个正例样本有相同的概率被识别出来”的假设在现实中往往不成立

标签缺失的处理方法

- 标签缺失的处理方法（二）

方法二：Xiao-Li L., Bing. L, Learning from Positive and Unlabeled Examples with Different Data Distributions

基本思想：向未标记样本U中添加大量的负例样本，降低U中的噪声（即正例样本）的比例。再利用Expectation Maximization（EM）算法，构建一些列的朴素贝叶斯分类器（Naïve Bayesian Classifier， NB-C），在迭代过程中不断改变NB-C的参数直至稳定。

- 优点：不依赖于额外的假设，健壮性高
- 缺点：需要添加大量的负例样本；适合文本分类，在其他场景中准确度相对较低

方法三：Wee S.L., Bing L., Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression

基本思想：假设正例样本有相同的概率被识别出来，剩余正例样本和所有负例样本组成未标记样本且其中的正例样本视为噪声。将噪声率作为权重带入逻辑回归模型的损失函数中进行求解。

- 优点：在假设“正例样本有相同的概率被识别出来”成立下，该方法的精度很高
- 缺点：依赖“正例样本有相同的概率被识别出来”的假设

标签缺失的处理方法

- Expectation Maximization算法

隐变量：分布函数中未观测变量。令X表示已观测变量集，Z表示隐变量观测集，Θ表示模型参数。如果对Θ做极大似然估计，则应最大化对数似然$LL(\Theta|X,Z)=\ln P(X,Z|\Theta)$。然而Z未观测，上式无法求解。此时可以对Z求期望从而最大化已观测数据的对数边际似然$LL(\Theta|X)=\ln P(X|\Theta)=ln\sum_Z P(X,Z|\Theta)$

EM算法常采用迭代的方法来估计隐变量：如果参数Θ已知，则根据训练数据推断最优隐变量Z的值（E步）。反之若Z已知则可对参数Θ做极大似然估计（M步）。具体的步骤：

E步（Expectation）：已当前参数Θ^t推断隐变量分布$P(Z|X,\Theta^t)$并计算对数似然函数$LL(\Theta|X,Z)$关于Z的期望$Q(\Theta|\Theta^t)=E_{Z|X,\Theta^t}LL(\Theta|X,Z)$

M步（Maximization）：寻找可最大最大化似然函数的参数，即

$$\Theta^{t+1}=arg\max_{\Theta}Q(\Theta|\Theta^t)$$

标签缺失的处理方法

- 高斯混合分布

一批样本中的每个都服从高斯分布，但是高斯分布的参数不同（但是有限个）。可以定义成：

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

其中p是高斯分布，α_i是混合系数，满足$\sum \alpha_i = 1$

在聚类问题中，如果样本$D = \{x_1, x_2, \dots, x_m\}$服从高斯分布（即$x_j$依概率$\alpha_i$ 服从高斯分布$p(x|\mu_i, \Sigma_i)$ ），令随机变量$z_j \in \{1,2, \dots, k\}$表示生成样本$x_j$的高斯混合成分，且取值未知。则$z_j$的先验概率$p(z_j = i)$对应$\alpha_i$（i=1,2,..,k）。 z_j的后验分布（记为γ_{ji})是：

$$P_M(z_j = i|x_j) = \frac{P(z_j = i)p_M(x_j|z_j = i)}{P_M(x_j)} = \frac{\alpha_i p(x|\mu_i, \Sigma_i)}{\sum_{l=1}^{k} \alpha_i p(x_j|\mu_l, \Sigma_l)}$$

标签缺失的处理方法

- 高斯混合分布（续）

利用极大似然估计对参数$\{(\alpha_i, \mu_i, \Sigma_i)|1 \leq i \leq k\}$进行估计：

$$\mu_i = \frac{\sum_{j=1}^{m} \gamma_{ji} x_j}{\sum_{j=1}^{m} \gamma_{ji}}$$

$$\Sigma_i = \frac{\sum_{j=1}^{m} \gamma_{ji} (x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{m} \gamma_{ji}}$$

$$\alpha_i = \frac{\sum_{j=1}^{m} \gamma_{ji}}{m}$$

可以用上述公式结合EM算法进行更新：在每步迭代中，先根据当前参数来计算每个样本属于每个高斯成分的后验概率γ_{ji}（E步），再根据上述公式更新参数$\{(\alpha_i, \mu_i, \Sigma_i)|1 \leq i \leq k\}$直至收敛。

标签缺失的处理方法

- 如何将EM算法和高斯混合分布应用在PU问题上？

对于有标记的样本集$D_l = \{x_1, x_2, \dots, x_l\}$和未标记的样本集$D_u = \{x_{l+1}, x_{l+2}, \dots, x_{l+u}\}$, $l + u = \mathrm{m}$.假设所有样本独立同分布，且由同一个高斯混合模型生成，包含正例样本服从的高斯分布$p(x|\mu_1, \Sigma_1)$和负例样本服从的高斯分布$p(x|\mu_2, \Sigma_2)$。用EM算法给出$\{\mu_1, \Sigma_1, \mu_2, \Sigma_2\}$和混合样本中，每个样本属于正例的概率$\{\gamma_{l+1}, \gamma_{l+2}, \dots, \gamma_{l+u}\}$.

初始化：令高斯混合成分的系数$\alpha_1 = \alpha_2 = 0.5$，$\{\mu_1, \Sigma_1\}$、$\{\mu_2, \Sigma_2\}$分别由带标记的正例样本$D_l$和无标记样本$D_u$给出。按照下面的步骤进行迭代：

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

$$\gamma_{ji} = \frac{\alpha_i p(x|\mu_i, \Sigma_i)}{\sum_{l=1}^{2} \alpha_i p(x|\mu_l, \Sigma_l)}$$

标签缺失的处理方法

- 如何将EM算法和高斯混合分布应用在PU问题上？（续）

M步：基于γ_{ji}更新模型参数，其中l表示带标签的正例个数

$$\mu_1 = \frac{1}{\sum_{x_j \in D_u} \gamma_{j1} + l}\left(\sum_{x_j \in D_u} \gamma_{j1} x_j + \sum_{x_j \in D_l} x_j\right)$$

$$\mu_2 = \frac{1}{\sum_{x_j \in D_u} \gamma_{j2}} \sum_{x_j \in D_u} \gamma_{j2} x_j$$

$$\Sigma_1 = \frac{1}{\sum_{x_j \in D_u} \gamma_{j1} + l}\left(\sum_{x_j \in D_u} \gamma_{j1}(x_j - u_1)(x_j - u_1)^T + \sum_{x_j \in D_l} (x_j - u_1)(x_j - u_1)^T\right)$$

$$\Sigma_2 = \frac{1}{\sum_{x_j \in D_u} \gamma_{j2}} \sum_{x_j \in D_u} \gamma_{j2}(x_j - u_2)(x_j - u_2)^T$$

$$\alpha_1 = \frac{1}{m}\left(\sum_{x_j \in D_u} \gamma_{j1} + l\right), \qquad \alpha_2 = \frac{1}{m} \sum_{x_j \in D_u} \gamma_{j2}$$

标签缺失的处理方法

- 基于混合高斯分布的标签补充的优缺点

优点

- 当数据的分布比较符合高斯分布时，算法的准确度较高
- GMM给出的是属于某一个类别的概率，使用更灵活

缺点

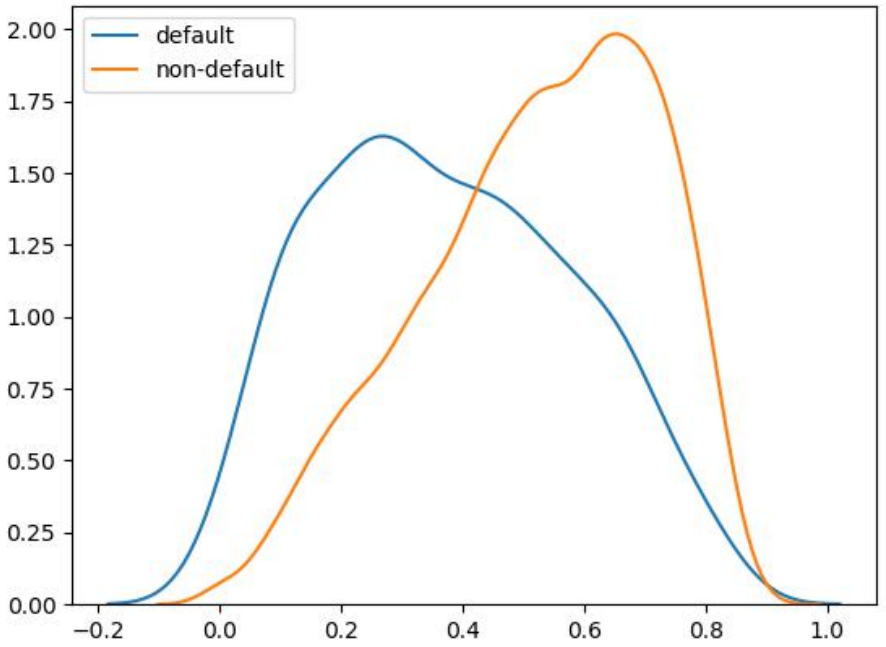
- 计算较为复杂
- 对样本量的要求比较高
- 需要样本（近似）服从高斯分布

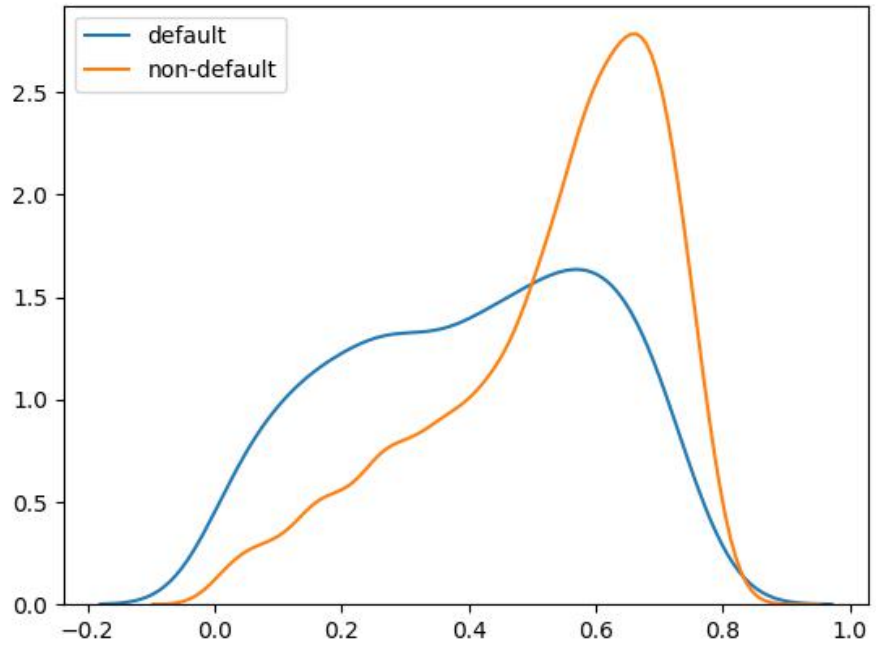
案例

- 案例一：半监督k-均值聚类法补充标签

我们从kaggle上下载一份关于房屋抵押贷款违约预测的数据集，从中挑选了10000个样本用作试验。全部样本的违约与非违约的标签都是已知的。本试验中，我们将随机选择10%的样本（记为A）作为标签已知的样本，剩余90%的样本（记为B）将根据A的标签进行半监督K-均值聚类。算法评估将根据真实标签和预测标签的对比来进行。

排除标签项和身份项后，原数据中有120列特征。我们逐一检查了每个特征在违约与非违约上的分布，发现EXT_SOURCE_2与EXT_SOURCE_3在两批样本上的分布的差异很大：



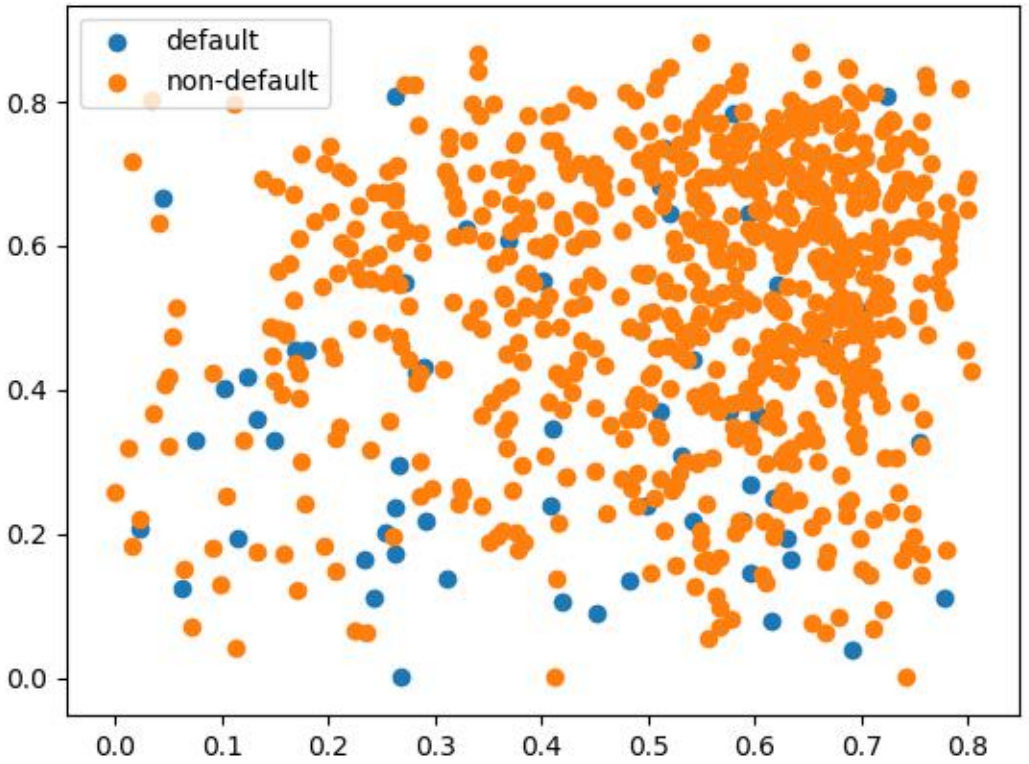


案例

- 案例一：半监督k-均值聚类法补充标签(续)

利用EXT_SOURCE_2与EXT_SOURCE_3检验两种已知标签的样本的分布。可以看出：

- 违约样本的占比很少，且分布较为均匀
- 非违约样本占比较多，但是分布集中在两个变量的较高的值



基于EXT_SOURCE_2与EXT_SOURCE_3使用半监督k-均值聚类后，我们对未标记样本的类别进行指派，得到的聚类结果与真实值的混淆矩阵为：

$$\begin{bmatrix} 362 & 170 \\ 2342 & 4318 \end{bmatrix}$$

可知在未标记样本中，违约样本的识别率为$\frac{362}{362+170} = 68\%$, 同时正常样本的误杀率为$\frac{2342}{2342+4318} = 35\%$

案例

- 案例二：混合高斯法补充标签

我们在案例一中的数据中，随机从正例（即违约样本）中抽取50%的样本与负例（即非违约样本）组成无标记样本，剩余的50%的正例留作有标记样本。此时无标记样本的样本来自于两种（多元）正态分布$Norm_1(\mu_1,\sigma_1^2)$与$Norm_2(\mu_2,\sigma_2^2)$，但是分属两种正态分布的概率是未知的。我们应用混合高斯分布对该概率进行估计。其中μ_1,σ_1^2的初始值由正例样本的均值与方差进行评估，μ_2,σ_2^2的初始值由正例样本的均值与方差进行评估。

当迭代停止后，算法将给出每个为标记样本分属2个正态分布的概率。我们将较大的概率分属的分布，作为该样本所属的类别。对应的混淆矩阵为

$$\begin{bmatrix} 215 & 81 \\ 3096 & 4303 \end{bmatrix}$$

因此我们有：

◆在未标记样本中，有$\frac{215}{215+81}=73\%$的正例（即违约样本）被正确识别出来

◆在未标记样本中，有$\frac{3096}{3096+4303}=41\%$的负例（即非违约样本）被错误地划分为正例

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI

天善智能
TIANSHAN SOFT