一个能核实互联网用户身份并利用互联网用户信息的 Credit Risk Model

王益

February 11, 2014

1 问题

金融机构的贷款历史数据可以用来训练一个 credit risk model，比如 logistic regression model，用来预估一个贷款者的信誉（本文中简化为还款率）。但是金融机构了解的贷款者属性不够丰富；我们希望从互联网上找到贷款者的其他信息，让 credit risk model 的特征更加丰富，从而预估更加精准。但是一个难处是，我们不确定一个在金融机构注册的贷款者是不是某个互联网用户。这里我们提出一个具备聚类特点的 credit risk model，它既学习预估一个人的还款率，同时匹配贷款者和互联网用户的身份。

2 数据

考虑金融机构有 N 条贷款记录，每一条记录 $\langle r_i, y_i \rangle$ 中， r_i 是一个布尔变量，如果是 1 则表示还款了，否则用 0 表示没有还款； y_i 是一个贷款者。通常包括姓名、年龄、身份证号等信息。这是训练数据。另外，存在从互联网上爬下来了 M 个互联网用户的信息，记为 $u_j, 1 \leq j \leq M$。每个 u_j 里的信息通常比 y_i 里的丰富，比如包括最近在微博上说了什么，最近和哪些人有过交流等。

3 模型

我们希望核实互联网用户身份。具体的说，想知道每个 y_i 很可能对应哪个 u_j。然后希望利用 u_j 里丰富的信息，帮助我们训练一个精确的 creidit risk model。为此，我们为每一个 y_i 增加一个 hidden variable $z_i \in [1, M]$，用来标记 y_i 对应哪个 u_j。这样，预估一个新的 y 对应的 r 的问题就变成了

这样:
$$P(r|y) = \sum_{1\le z\le M} P(r|u_z)P(z|y) \tag{1}$$
我们得到了一个隐含变量模型，可以用 EM 算法来训练之。

我们假设 $P(r|u_z)$ 用一个 logistic regression model 来描述，但是实际上用什么模型都可以，下面所述算法都能支持。而 $P(z|y)$ 表示成一个 M 维向量 $\boldsymbol{\gamma}_y$。

4 训练

有隐含变量的模型，通常用 EM 算法来学习。EM 算法是一个有收敛性保证的 meta-algorithm——只要我们循环执行一个 E-step 和一个 M-step，就能得到一个单调收敛的模型。其中 E-step 是估计 hidden variable 的概率分布，而 M-step 可以利用 E-step 的结果，通过最大化模型的 log-likelihood 来更新模型参数。为此我们在下文中推导 hidden variable 的分布计算公式和模型的 log-likelihood 的最大化方法。

4.1 初始化

如果我们完全不知道 y_i 应该如何和 u_j 对应，那么我们只能假设 z_i 的分布是 $[1, M]$ 区间上的均匀分布。此时初始化就是把每个 $\boldsymbol{\gamma}_{y_i}$ 的每个元素都设置为 $1/M$。

4.2 E-step

更新每个 γ_{y_i}，其中 $\gamma_{y_i,j} = P(z_i = j \mid r_i, y_i)$，而
$$P(z = j \mid r, y) = \frac{P(z, r \mid y)}{P(r \mid y)} = \frac{P(r|u_j)P(z = j|y)}{\sum_{1\le z\le M} P(r|u_z)P(z|y)} \tag{2}$$

4.3 M-step

如果 $P(r|u)$ 是用 logistic regression model 描述，那么会有一组参数 $\boldsymbol{\beta}$。严格的 M-step 要最大化 log-likelihood:
$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{1\le i\le N} \log P(r_i|y_i; \boldsymbol{\beta}) \tag{3}$$
其中
$$\log P(r_i|y_i; \boldsymbol{\beta}) = \sum_{1\le i\le N} \log \sum_{1\le z\le M} P(r_i|u_z)\gamma_{y_i,z} \tag{4}$$

这个两个 $\sum$ 之间夹着一个 log 的形式很不容易对 $\boldsymbol{\beta}$ 求导。但是我们可以做一个简化，把第二个 $\sum$ 去掉:

$$\log P(r_i|y_i;\boldsymbol{\beta}) \approx \sum_{1 \leq i \leq N} \log P(r_i|u_x) \tag{5}$$

其中 $x = \arg\max_j \gamma_{y_i,j}$。

这个近似和 K-mean 算法对 EM clustering 算法的近似很相近。它的好处是，把 M-step 变成了调用标准 logistic regression model 训练算法，甚至可以给 logistic regression model 加上 L-1/L-2 regularization。

5 改进

从算法原理上，大家都知道 EM 算法会很容易陷入局部最优。对这个问题尤其如此。假设:

1. 有两个贷款者：Bob 和 Alice，其中 Bob 总不还贷，Alice 总还贷;
2. 有两个互联网用户：一男一女;
3. 用户特征只有两个：gender=male 和 gender=female;
4. 随机初始化时，恰好 Bob 和 Alice 都被认为非常像那个男的互联网用户。

那么 M-step 会学得一个 logistic regression model，它认为 gender=male 这个特征的 weight 接近 0，因为从 Alice 和 Bob 的还款记录来看——有的还款有的不还。另外 gender=male 这个特征的 weight 也是 0，因为这个特征根本没有出现在 logistic regression model 的训练数据里。这样一来，这个模型总是认为还款率是 50%。等到 E-step 时，公式 (2) 里的 $P(r|u_j)$ 总是 0.5，那么 $P(z=j \mid r,y)$ 主要受到 $P(z=j|y)$ 的影响，而后者就是随机初始化的结果。换句话说，M-step 对 logistic regression model 的更新没法帮助我们修正随机初始化中的错误。

一个直接的解法是公式 (5) 里的 x 的取法从 $x = \arg\max_j \gamma_{y_i,j}$ 改成 $x \sim Discrete(\boldsymbol{\gamma}_{y_i})$。也就是从“找概率最大的互联网用户”变成“按照分布 $\boldsymbol{\gamma}_{y_i}$ 随机选择一位互联网用户“。这样就把 x 的分布的估计改成了 Gibbs sampling。那么上述 EM 算法也就成了一个 stochastic EM 算法。