<Counterfactual fairness> 代码复现实验报告

徐鑫鑫

2020 年 11 月 4 日

1 问题设定

已知：21790 法学院申请者的入学考试分数（LSAT）、之前平均绩点（GPA）、性别（sex）和种族（race）和标签（FYA）

预测：申请者第一年入学后是否将取得高的平均成绩（FYA）

要求：预测值不会因申请者个人的种族和性别存在偏见

2 算法

文章提出了 4 种模型：unfair full model(基准模型)，unfair unaware model(基准模型)，fair deterministic model（Fair add）和 fair nondeterministic model(Fair k)。

unfair full model：使用所有属性（LSAT，GPA，sex，race）对因变量 FYA 进行 Logistic 回归拟合并预测申请者入学后一年后的平均成绩（FYA）

unfair unaware model：仅使用非受保护属性（LSAT，GPA）对因变量 FYA 进行 Logistic 回归拟合并预测申请者入学后一年后的平均成绩（FYA）

fair deterministic model (Fair add)：将 GPA、LSAT 和 FYA 建模为具有与种族和性别无关的附加误差项的连续变量

$$GPA = b_G + w_G^R R + w_G^S S + \epsilon_G$$

$$LSAT = b_L + w_L^R R + w_L^S S + \epsilon_L$$

$$FYA = c_F + a\epsilon_G + b\epsilon_L$$

步骤 1：使用训练集线性拟合前两个模型，每个模型分别使用种族和性别来单独预测 GPA 和 LSAT，然后计算两个模型的残差 ϵ_G 和 ϵ_L（训练集）

步骤 2：同步骤 1，使用测试集线性拟合前两个模型，获取残差 ϵ_G 和 ϵ_L（测试集）

步骤 3：使用残差 ϵ_G 和 ϵ_L（训练集）来拟合 FYA，使用残差 ϵ_G 和 ϵ_L（测试集）进行预测。

2.1 nondeterministic model(Fair k)

假定一个潜变量：学生知识（K）影响 GPA，LSAT，FYA

$$GPA \sim N(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G)$$

$$FYA \sim N(w_F^K K + w_F^R R + w_F^S S, 1)$$

$$LSAT \sim Poisson(exp(b_L + w_L^K K + w_L^K R + w_L^R S))$$

步骤 1（获取训练集 K）：假设 $sigma$ 先验服从半正态分布，其余待测变量先验均服从正态分布，用训练数据对上述三个模型进行贝叶斯后验推断（MCMC 方法进行拟合，迭代 2000 次），推断出 K 值及其他待测变量的后验分布。

步骤 2（获取测试集 K）：假设 K 值服从正态分布，用测试数据对 GPA 和 FYA 模型进行拟合，方法同步骤 1，推断出 K 值的后验分布（其他待测变量直接采用步骤 1 中对应待测变量后验分布的均值）。

步骤 3：使用训练集 K 来线性拟合 FYA，使用测试集 K 进行预测。

3 实验设定

设备：python 3.6

训练集：测试集 = 8 : 2

模型评估：RMSE（均方根误差），越小越好

main 文件：CF.py

预处理：包括导入数据集，划分自变量和因变量，对分类数据进行 OnehotEncoder 处理，划分训练集和测试集

每个算法对应 CF.py 文件中一个函数，其中 nondeterministic（）函数（对应 fair,nondeterministic 模型）引入两个额外的函数，分别对应于 law_school_train.py 中的 nondeterministic_tr() 函数 (用来获取训练集 K) 和 law_school_only_u.py 中的 nondeterministic_te() 函数 (用来获取测试集 K)

预期 RMSE 值：unfair full model < unfair unaware model < fair nondeterministic model(Fair k) < fair deterministic model（Fair add）