医疗花费预测技术报告

17375180 王佳瑞

一、问题分析

根据数据中的输出（charges 列）易知这个问题是回归问题。对于输入列，有连续型数值（age,bmi,children),也有标签类数据（sex,smoker,region)。考虑这个问题可以用决策树进行回归问题，再用 bagging 或者随机森林的思路集成学习优化效果。

二、解决方案

对于回归问题，决策树结点处划分函数

$$L(j,s)=\sum_{x_i\in R_1(j,s)}(y_i-\hat{c_1})^2+\sum_{x_i\in R_2(j,s)}(y_i-\hat{c_2})^2$$

其中 $\hat{c_1}=\frac{1}{N_1}\sum_{x_i\in R_1(j,s)}y_i$

一般对于连续性特征，决策树结点选择一个数值作为交界，以这个特征的数值将结点分为两个子树的根节点。在这个问题中，对于连续型数值标签，我就是用这个做法进行尝试。

对于标签类数据，观察到除了地区特征，其余的特征都是有两个标签，因此我分别对地区类标签尝试分为四类，分为两类（（0，1）/（2，3）；（0，2）/（1，3）；（0，3）/（1，2））。这两种做法对应文件分别是 DecisionTree.py; module.py.

在设计决策树时，采用预剪枝方案，设置 tolN 参数限制任一子结点集合大小小于 tolN 时不做分割，设置 tolS 参数限制按照最优特征切分前后误差小于 tolS 则不切分。

三、实验对比

在实验中，采用 bagging 方式，以及采用结点选择 k 个特征的方式进行实验(见 dataloader)。并在上述两种处理 region 的思路上分别尝试不加入 bagging，不特征随机选择；加入 bagging，不随机选择特征；加入 bagging 并且随机选择特征。最终结论是 bagging 阶段数据又放回的分为 10 份，做出 10 个决策树最后去平均值，在决策树结点处选择所有标签计算，将地区分为 4 类，这个实验效果最好。

四、结果展示

在测试集上，排名 14，R-square 指标未 0.8572

14	17375180_王佳瑞	4	12/29/20	0.8572 (14)