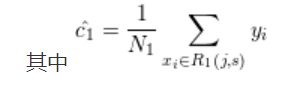
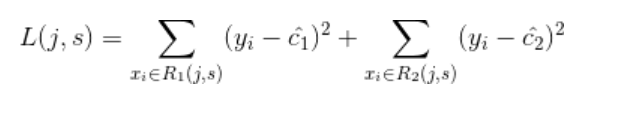
# 医疗花费预测技术报告

17375180 王佳瑞

1. 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。

1. 解决方案

对于回归问题，决策树结点处划分函数

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。

1. 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。

1. 结果展示

在测试集上，排名14，R-square指标未0.8572

740b54775a4fe3fcca6051d9c481466