方骏-2020年2月16日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：在连续属性值的问题中，根据阈值修改算法8到11行究竟是怎么修改的？

讨论后的理解：这种修改方法还是比较简单的，和我之前想的是一样的。我们讨论后的一致结果就是进行分类处理。在算法中，开始的时候就直接分情况讨论，分成连续值和离散值的情况，然后针对不同的情况处理就好了，因为本身连续值区别就不大，直接分成几个区间就可以了。

1. 别人提出的问题的理解：
2. 问题2：对于决策树生成算法而言，如何计算sorting of a continuous attribute takes |D|log|D|的时间复杂度

自己的理解：这个就是正常的排序算法。因为对于连续属性值的排序其实就是正常的排序问题，大部分优秀的排序算法，比如快速排序等，时间复杂度都是nlogn。

1. 问题3： 以cross-validation为例，按理来说k次交叉验证后应得到的k个模型分别有不同的accuracy，那取这k个accuracy的平均值有何意义？它并不能代表k个模型中的任意一个。

自己的理解：虽然这k个模型是不同的，因为是不同的训练数据集训练出来的模型，所以模型肯定是会不同的，虽然模型不同，但是用的生成决策树的方法都是一样的，所以本质是一样的。测试出k个accuracy的平均值，意思就是知道该方法到底好不好，如果平均值不够，自然就可以换别的方法了。所以这是测试平均值的意思。

4、 问题4： 比起单纯看模型的accuracy来说，计算recall和precision可以提供哪些额外的信息来进一步评估模型？

自己的理解：对于一些情况，比如网络入侵，因为正常情况下网络入侵的比重是非常小的，所以即使我每次预测预测的都是不入侵，那么accuracy也会很高，这显然是不好的。那么recall和precision这种针对positive标签的评估计算，自然更加准确，其分母是positive的数量，自然会比总数要好很多。

1. 读书计划

1、本周完成的内容章节：3.1-3.3

2、下周计划：第四章

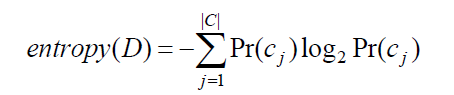
四、读书摘要及理解

1、读书摘要及理解

本次阅读读的内容是有监督学习的基础和决策树以及评价方法。有监督学习是最常用的机器学习方法，通过学习带有标签的数据来训练出新的模型，从而用来判断新数据的标签或分类。训练出的模型可以作为各种形式，像是决策树就是一种分类出的模型形式。训练数据是用于得到训练模型的，测试数据使用来测试模型准确性的。测试数据和训练数据的分布大致相似，都带有标签。评估模型通常采用准确率：

但这往往是不够的。

决策树和正常的树很像，通过一步一步的分支，根据待测数据的属性来选择路径的分支点，从而到达叶子节点，即分到的标签。决策树的一种构成方法是选择属性来作为分支点，再从该分支点下去选择新的属性。属性的选择根据信息熵的定义来选择：



信息熵越大，说明信息复杂程度越大，越不纯。决策树分支的目的就是是分支后的各个支路是纯的，保证所有的样例值对应一条支路。所以选择属性的依据，就是分完后的子树的熵与每分前的子树的熵的差最大，减小的越多越好，这是选择分支的缘由。

评价标准有很多种，但是最常用的往往都是比较简单的。除了之前的accuracy之外，还有一些只针对少数情况的评价标准，在问题中已经讨论过，这里不再赘述。

总之，有监督学习是一种很重要的机器学习方法，在现实应用中更加容易实现并且可以很容易地应用到实际领域中。