读书报告

71117203-袁佳怡

1. 自己提出的问题的理解：
2. 提出的问题1：在决策树生成算法中，对于属性为连续值时，如何理解sorting of a continuous attribute takes |D|log|D|的时间复杂度？

讨论后的理解：这里的时间复杂度其实是指现在的排序算法能够达到的较好水平，在这里进行指代，对于实例按照属性进行排序，主要还是要看所选择的排序算法才能确定出实际的时间复杂度。

1. 提出的问题2：如何理解that interpolation is needed if such a point cannot be found

讨论后的理解：这里是说可以使用插值的方法，是针对数据缺失人工补足的一种方法，书上并未进行展开论述。

1. 别人提出的问题的理解
2. 问题3：3.2.4节中，处理决策树额外问题中的Handling Skewed Class Distribution的第二种方法Another solution is to rank the new cases according to how likely they may be intrusions.要如何理解？如何对新的样例进行排序？排序后又对调整模型有什么好处？

自己的理解：这里是为了处理决策树偏斜类别分类的问题，方法是根据新来的数据可能是入侵的程度来对它们排序，这样使得用户就可以着重研究那些排名靠前的情况。因为是入侵数据的比率很小，决策树生成出的就是一个叶子结点，所以对于新的数据，如果排序得到它是入侵数据的可能性较高的话就要对其优先进行单独研究，通过单独对较大可能是入侵数据的进行研究，不用决策树直接得到对所有数据的判断，也是为了能够提高决策的正确性。但至于如何判读其可能性又如何排序，书中未涉及相关介绍。

1. 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：accuracy是指决策正确的比例，但对于某些数据而言，这个指标可能是没有意义的，例如书上介绍的只有1%的数据异常的例子，即使accuracy高达99%都是毫无意义的。我们要更多关注的是正例的判断情况，有多少的正例被判断正确，在判断为正例的实例中又有多少是真正的正例，但至于recall和precision侧重哪一个指标还是要看问题的具体要求的。

1. 问题5：Breakeven Point的计算对于分类器评估的意义何在？

自己的理解：其实我们真正理想化的情况是，recall和precision都能够尽量的大，使得我们对于正例的判断更加有效实用，但现实情况是，两者存在着此消彼长，一个的提升要以另一个的下降为牺牲的情况，所以我们尽量使得达到一个平衡点，就是这里的breakeven point，使得能够p和r相等，尽量的大也尽量的均衡，但至于为什么这个情况下就会是最好的，暂无法从数学和实验的角度给出解释。

1. 问题6：他说任何的threshold value在vi和vi+1之间的有同样的效果，具体解释一下？

自己的理解：对于连续属性，我们应该选择能将信息增益最大话的阈值，为此我们需要尝试所有可能的阈值取值，因为处于vi和vi+1间的所有阈值在划分训练数据集的效果上是一样的，都是把数据分成两个集合为了把连续属性按照两个区间划分开，vi和vi+1是两个离散值，用于将连续的进行二分，具体切在哪一边都没有区别。

1. 读书计划

1、本周完成的内容章节：如2.6-3.3

2、下周计划：第四章、第五章