1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：P67算法2-有序化的类中提到，决策表中类的顺序非常重要，首先需要寻找训练数据中出现频率最低的类的规则，这样才能保证稀有类能被找到。但是这个算法不是按照类来prepare数据的吗，为什么会存在这样的问题呢？？

讨论后的理解：算法在执行过程中，通过规则学习也会将数据集中学习到的规则覆盖的数据进行remove操作。一开始我纠结的点在于Pos和Neg数据集的划分问题，因为如果每次都将正例置于Pos中，就不会存在稀有类的问题，但是忽略了算法在执行过程中会将相关数据移除。

1. 提出的问题2：P69图3.13第12行，在完成新的候选条件的筛选之后，需要做一步去掉重复值和无意义键值对（条件）的操作，除了书中给出的一个示例之外，“无意义”还应该从哪些方面来考虑呢？

讨论后的理解：书中给出的示例是：同样的条件推出不同的类，说明这个条件还不能很好的作为一个规则，所以就将这个条件看成是无意义。在实际操作中，除了应该考虑理论上的无意义条件，还应该考虑一些应用上无意义的条件，比如极端条件、属性值本身无意义或超过正常范围等。

1. P70算法提到的PruneRule函数，是对规则进行修剪，要删除最大化公式15对应函数的子集，那么什么样的子集会使得PruneRule（）出现最大值呢？换句话说，公式在实际数据中代表的意义是什么？

讨论后的理解：中文版书籍公式有误，根据英文版第二版的解释，p表示当前规则下，验证集中的满足当前规则的正例（PrunePos）的数量；n表示对规则进行剪枝后，验证集中负例满足修剪后规则的数量。所以要找到能够使得函数最大值，实际上可以理解为p为常量，关于n的函数取得最大值。已知，p和n一定是非负数，可以通过绘图或者求导的方式，找到什么时候n能够取得最大值。通过绘图可知，函数x轴右边是单调递减的（p=0时，函数恒等于-1除外），所以当n->0时，函数可以取得最大值，即n取最小值时，v(函数)取得最大值。那么要删除的子集，可以理解为：删除这个子集后，验证集中满足这个规则的负例个数最少。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题4：P 60页的“剪枝之后的规则集合可能不再是互不相交且完全覆盖的”，因为我感觉示例里面P59 3.7(b)的剪枝后的决策树就是互不相交且完全覆盖的，而且我还觉得在这个图里对其他任意的树节点剪枝，得到的结果还是互不相交且完全覆盖的。

自己的理解：在决策树中，当一个节点有多个分支时，如果将中间的某些分支剪掉，那么原本被该分支覆盖的那部分数据将不能被剪枝后的决策树的中所有规则覆盖到。

1. 问题5：64页中为什么说在两条曲线的交点出可以判定分类器的好坏？为什么之后又说这样“有时并不是一个令人满意的答案”，然后又使用面积进行衡量？使用交点和面积两种方式的数学原理和现实原理分别是什么？

自己的理解：ROC曲线实际上记录的是数据集在不同比例的正例负例划分情况下，真阳率与假阳率的比值的连线。我们绘制曲线的目的是找到一个更好的分类器，衡量分类器好坏的标准是总是能够使得真阳率远大于假阳率。因为交点本质上是一个分界点，我们需要做的是在分界点的量变的区域都能找到总是能够使得真阳率远大于假阳率的分类器。然而实际操作中，很难根据假阳率的界值去选择一个分类器，我们需要选择一个在总体上都能满足真阳率>>假阳率的分类器，而AUC(曲线下的面积)代表的是真阳率/假阳率的积分，所以积分越大，就可以理解为其真阳率总是大于或者说远大于假阳率。所以可以通过比较面积来选择分类器。

1. （必填）读书计划

1、本周完成的内容章节：第三章（3.1-3.5）

2、下周计划：第三章（3.6-3.10）

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

2、伪代码的具体实现(选做)