Web data mining部分：

第三章讨论部分：

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：

结合西瓜书第二章

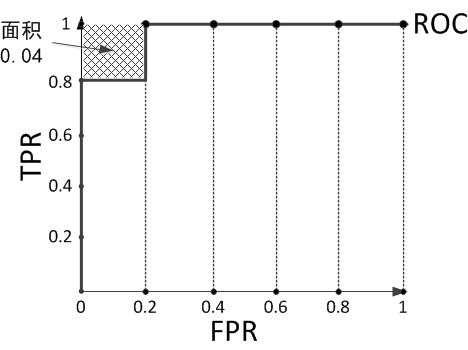
为什么有AUC=1-Lrank。没有理解

讨论后的理解：

这个问题看了b站上的举例和csdn上的数学推理得出两种解释。

解释一：举例说明实际意义：

我们不妨举个例子，预测结果序列为[1, 1, 1, 1, 0, 1, 0, 0, 0, 0]。



计算上图的AUC为0.96与计算正例与排在负例前面的概率0.8 × 1 + 0.2 × 0.8 = 0.96相等，而左上角阴影部分的面积则是负例排在正例前面的概率0.2 × 0.2 = 0.04。刚好是Lrank的值。

解释二：数学理解：

roc曲线上图积分的意义实际上相当于，相当于给定最小时间单位下分段递增匀速运动时间速度曲线的路程积分问题。TPR相当于速度，每段“时间”的起始TPR越高，这段“时间”内正例排在负例前面的概率也就越大，而这个概率就是路程；FPR则相当于时间，这个可以这么理解：每次旅行总会结束，而FPR也总会到达100%（x值为1），FPR每向x轴正方向移动一步，相当于离“旅行结束”又近了一步。这样，一个积分概率，一个积分路程，大概就能近似理解了。

1. 提出的问题2：

第68页倒数第二段，尝试所有的键值对不现实，因为可能的数据呈指数增长，不是应该是排列组合问题嘛，增长函数不应该是指数？

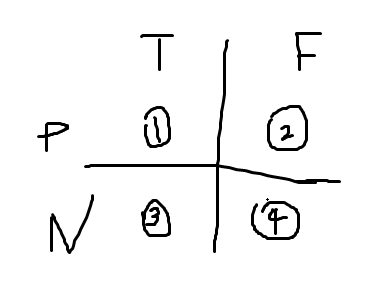
讨论后的理解：

书上的解释应该只是说明可能的数据呈巨大增长，增长的函数应该是A（m，n），并不能算指数增长。

1. 提出的问题3：

结合西瓜书第二章阅读，西瓜书p34页倒数第三段为什么，前一个标记为（x，y），若当前为真正例，则为（x，y+1/m+），若为假正例，则为（x+1/m-，y）？

讨论后的理解：



y=3/3+4，X=1/1+2.

在已知当前的样本为真正例的时候，新的x不变，y就等于y0+1/m+，在已知当前样本为假正例的时候，新的y不变，x就等于x0+1/m-。

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：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？

自己的理解：

因为规则学习的本质是，通过一定的数据学习到相应的规则，然后删除本来的学习集的数据，那么假定稀有类的数据的规则体部分被大数据的规则体所包括。

那么如果大数据的数据先被规则学习的话，学习完进行删除，很可能删除了所有稀有类的数据，导致稀有类数据的规则无法被学习，所以要先对稀有类进行规则学习。

1. 问题5：我没有理解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：

这个问题我们进行了激烈的讨论，首先明确规则剪枝是既可以自顶向下也可以自底向上。

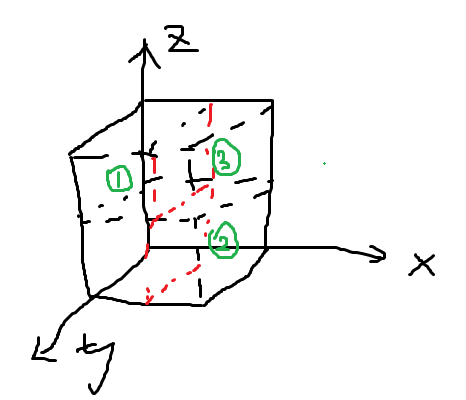
首先回答互不相交的问题：

这里可以有两种想法：（1）举例解释，对于一个连续的二维（x，y）数据集合，假定规则好瓜-》0<x<0.5,0<y<0.5

后来我们发现x属性对于瓜好坏基本上没有影响，那么规则可以剪枝为好瓜-》0<y<0.5。

同理坏瓜-》0.5<x<1,0.5<y<1,可以剪枝为0.5<x<1。那么必定存在集合（0.5<x<1，0<y<0.5）在规则下既是好瓜又是坏瓜，这显然是不合理的，需要对他重新进行评估。

1. 利用图形解释，基于向量空间。



如图所示的1号区域为好瓜，如果发现x对于分类影响不大可以扩充到（1和3）为好瓜。如图所示的2号区域为坏瓜，如果发现z对于分类影响不大可以扩充到（2和3）为坏瓜。那么显然3号区域重复覆盖了。

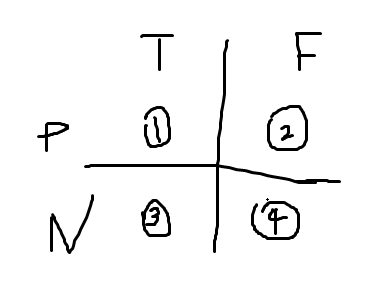
然后回答完全覆盖的问题：

打个比方说，西瓜可以通过x属性划分为条纹，黑色和绿色，我们发现在x属性为条纹的时候基本上就是好瓜，x属性为黑色的时候就是中瓜，如果x属性为绿色，那么无法判断是好，中，坏瓜，所以可以将x属性=绿色这个键值对减去，那么最后得到的结果必然是有漏规则的，所以规则剪枝会存在互不相交的问题。

1. 问题6：在ROC曲线中，为什么主对角线（代表随机猜测，即每个类分成正例为一个固定比例）上测试数据的假阳率FPR总是和真阳率TPR相等？

自己的理解：

还是先看图，fpr和tpr的定义。



Tpr=1/1+2

Fpr=3/3+4

由图显然可得1/3=2/4

证明Tpr=Fpr即为证明1/1+2=3/3+4.

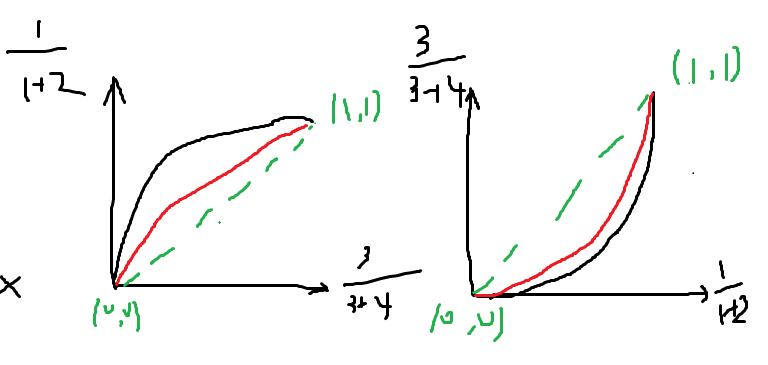
显然上式取倒数加1再取倒数可得证。

7、问题7：64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：

对于这个问题，我们不妨将roc曲线关于y=x轴对称。

可得原曲线如左图到右图所示。



对于右图，我们可得1/1+2为正确的值在所有预测值的比例，这是好的，

而3/3+4是不好的，那么一个好的分类器，必然要求在x相同的情况下，y越小越好

那么如图所示的黑线是显然优于红线的，那么翻转回去，roc形成的面积越大，分类器显然是越好的。

1. （必填）读书计划
2. 本周完成的内容章节：

（1）3.1到3.5完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

（2）为配合第三章的阅读，同时参考并且阅读了西瓜书的第二章，第四章和第十五章，并且放到一起综合学习

2、下周计划：3.6-3.10的阅读，并且联系第二章的关联规则再次理解3.5的内容

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）

通过第三章的阅读，发现涵盖了很多东西，额外阅读了西瓜书的对应章节，分类器的评估，决策树，规则学习，分别对于着《西瓜书》的第二章，第四章和第十五章。

作为第三章的补充，将分类器的评估，决策树，规则学习中额外第三章没涵盖到的知识点，额外补充如下：

分类器的评估：

1.roc曲线上图积分的意义实际上相当于，相当于给定最小时间单位下分段递增匀速运动时间速度曲线的路程积分问题。TPR相当于速度，每段“时间”的起始TPR越高，这段“时间”内正例排在负例前面的概率也就越大，而这个概率就是路程；FPR则相当于时间，这个可以这么理解：每次旅行总会结束，而FPR也总会到达100%（x值为1），FPR每向x轴正方向移动一步，相当于离“旅行结束”又近了一步。这样，一个积分概率，一个积分路程，大概就能近似理解了。

决策树：

1预剪枝使得决策树的很多分支都没有“展开”，这不仅降低了过拟合的风险，还显著减少了决策树的训练时间开销和测试开销。但另一方面，有些分支的当前划分虽然不能提升泛化性能，甚至可能导致泛化性能下降，但是后续划分可能提高他们，预剪枝禁止这些分支展开，给预剪枝带来了欠拟合的风险。

1. 后剪枝可能带来规则的冲突和不完全覆盖，上面已经综合论证过了。
2. 但是后剪枝开销比预剪枝大得多，却能减少欠拟合的风险

规则学习：

规则的生成本质上是一种贪心搜索的过程，需要一定的机制来缓解过拟合的风险，最常见的做法是剪枝，剪枝也分为预剪枝和后剪枝，预剪枝的常用算法是CN2算法，LBS越大，说明采用规则集进行预测与直接使用训练集正反例比例进行猜测的差别越大；LBS越小，说明规则集的效果越可能仅是偶然现象。在数据量大的情况下，通常LBS很大就可以停止伸张

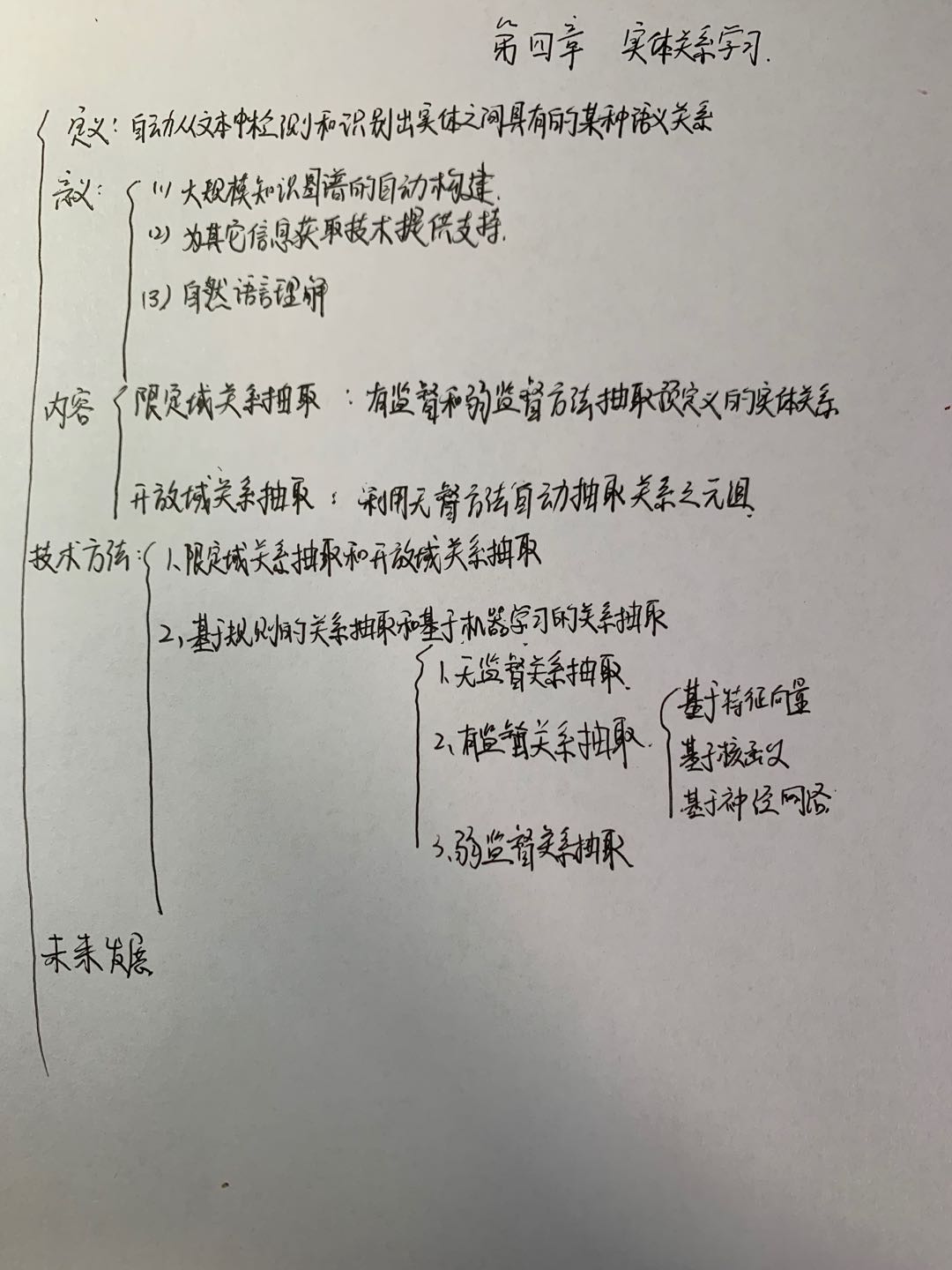
知识图谱部分：

读书报告内容：

分章节知识整理知识框架如下：

第四章：

知识框架：



遇到问题：

有监督关系抽取部分中的基于核函数不能理解，打算下周仔细阅读《西瓜书》支持向量机部分，再回头理解算法原理。

读书计划：

**知识图谱自己完成了第四章的阅读。**

学习方式：

第一遍完成阅读，有不理解的地方圈画作标记，圈画自己认为重点做标记。

第二遍再次阅读，完成知识框架的构建。

第三遍回顾解决问题。

下周计划：

1继续完成至少一章的阅读

2 完成师兄安排的后续type inference的相关任务。