**Q1**

1.2）1/（5+1）=1/6

1.3）1/3\*1\*1/4=1/12

2/3\*2/3\*3/4=1/3

P(MAX\_SEV\_IR|WEATHER\_R=2 and TRA\_CON\_R=0)=

1.7）由于训练集中为1的个数比为0的个数多，因此认为验证集结果均为0

1.8）经计算 1.7的做法正确率为0.505

1.6的做法正确率为0.532

因此使用朴素贝叶斯分类器的预测错误率低。

有用，但由于正确率只大一点，因此作用不是很明显

**Q2**

2.1）kMeans聚类 距离指标：Euclidean Distance 计算在m维空间里两个点的真实距离或者向量的自然长度

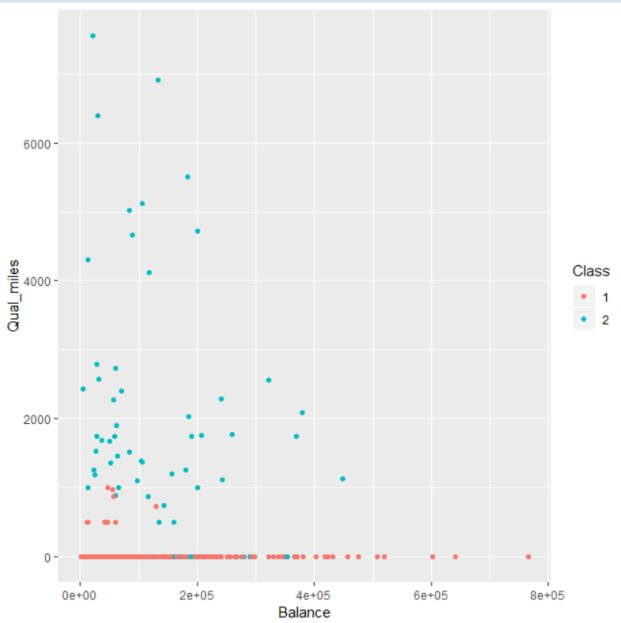
2.2）第21行：所有数据记录通过 class 进行分组汇总之后的 Balance 数值构成的向量在各组的“长度”（length），个数越多则多个 Balance 数值组成的向量越长，实际上这统计了数据记录在各组的计数

第22行：这条语句则统计了所有数据记录通过 class 进行分组汇总之后，每个变量在各组的均值（mean）

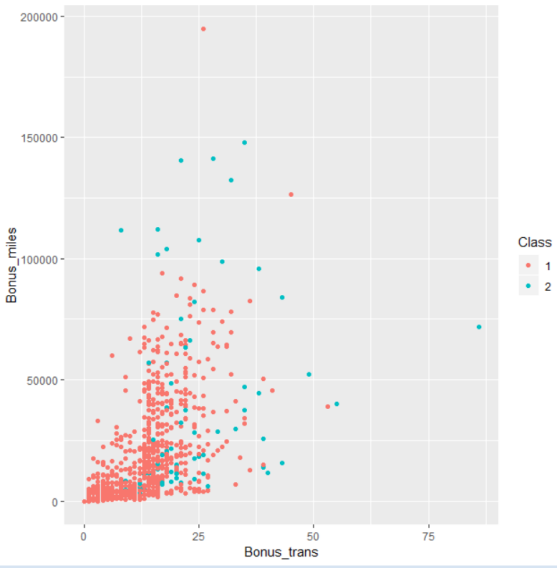
2.3）在EastWest中的数据中，将第1,4,8,9,15,17,19个簇聚为一类，剩下的聚为一类，k变为2，也就是两个类。这样做的目的是将Qual\_miles均值较小的分为一类，剩下的分为另外一类。

2.4）根据‘可用于兑换行程的奖励数’分类，class1里的样本点远多于class2，也就是说class1的客户更多。可是从均值来看，class2的客户拥有的平均可用于兑换行程奖励的里程数几乎是class1的两倍。说明class2的客户虽然人少，可拥有的可用于兑换行程奖励的里程数较class1的客户要多得多。

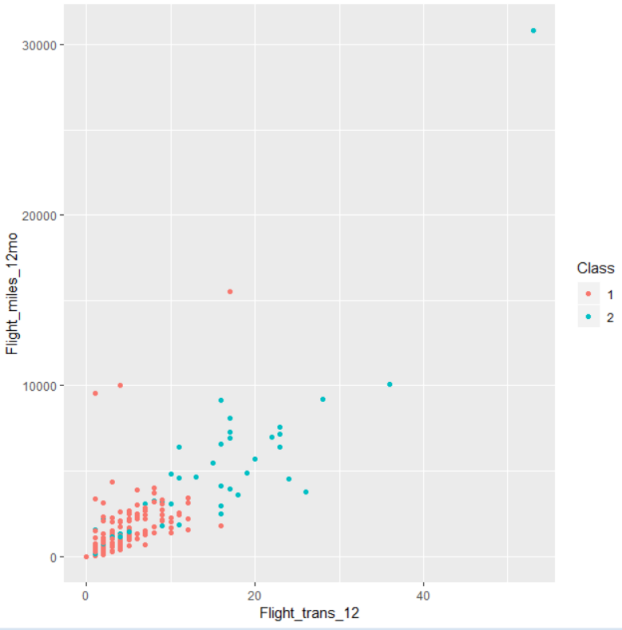
2.5）



以Balance和Qual\_miles为横纵坐标，可分析出class1的特征是几乎没有用于兑换高级服务的里程数，有可兑换行程服务的里程数，class2的特征是既有可以用于兑换高级服务的里程数也有可兑换行程服务的里程数；



以bonus\_miles和bonus\_trans为横纵坐标，类别一和类别二无明显区别，主要集中于有少量‘过去12月内通过非航空交易订单获得的里程数’和‘过去12月内获得非航空奖励的交易次数’。

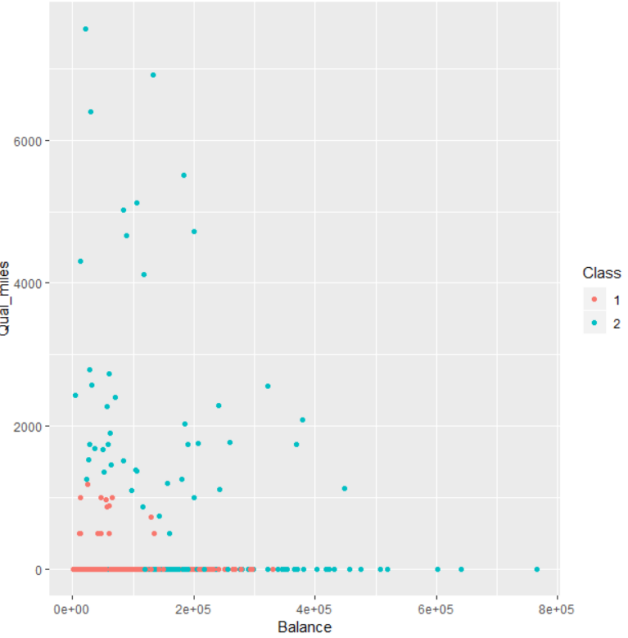


以Flight\_trans\_12和Flight\_miles\_12mo为横纵坐标，class1的客户‘过去12月内通过航空交易订单获得的里程数’及‘过去12月内获得航空奖励的交易次数’都很少，而相较之下，class2的客户拥有的更多。

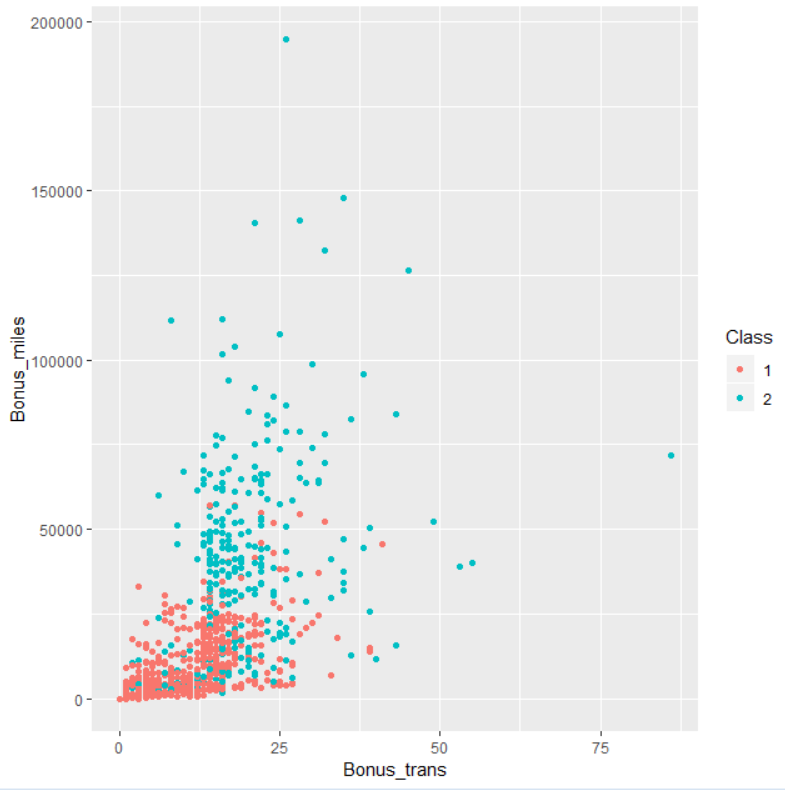
从图中分析航空公司应该喜欢class2的客户，因为他们既有可用于兑换高级服务的里程数也有可兑换行程服务的里程数，还有一定数量的通过航空交易订单获得的里程数，首先说明了他们在该航空公司有一定量的机票支出，其次说明他们可能还会带来后续的经济效益。

2.6） no.Award 与 with.Award 中样本点最多的都是class1，说明class1的客户较多。在class1中又是no.Award的客户数量最多, class2中又是with.Award的客户数量最多。说明第一类客人的特征之一是没有兑换最后一次旅程，第二类客人的特征之一是兑换了最后一次旅程。

2.7）ward指标更好一些。原因如下：



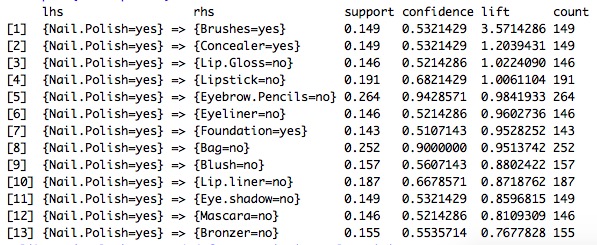
第一幅图与用average指标画的图几乎没什么差别



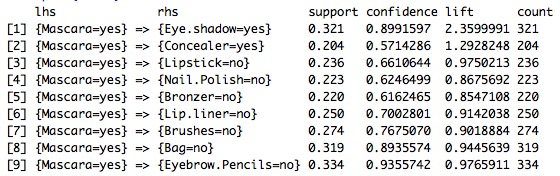
从第二幅图来看，以Bonus\_trans和 Bonus\_miles为x，y轴坐标，class1和class2的区分度更高，因此可以看出class1的客户拥有的平均‘过去12月内通过非航空交易订单获得的里程数’和‘过去12月内获得非航空奖励的交易次数’都很少，而相较于class1，class2的客户拥有的以上两项都较多一些。

**Q3**

3.1）

如果某个女士购买了 Nail.Polish，那么在支持度 support 至少 0.1,信心水平 confidence 至少 0.5 的情况下该女士还会购买Brushes和Concealer。

3.2）



如果某个女士购买了 Mascara,那么在支持度至少 0.2,信心水平至少 0.5 的情况下该女士还会买Eye shadow和Concealer。

3.3)

lhs rhs support confidence lift count

[1] {Mascara=yes} => {Eye.shadow=yes} 0.321 0.8991597 2.3599991 321

上述规则是3.2) 中得出的第一条规则。

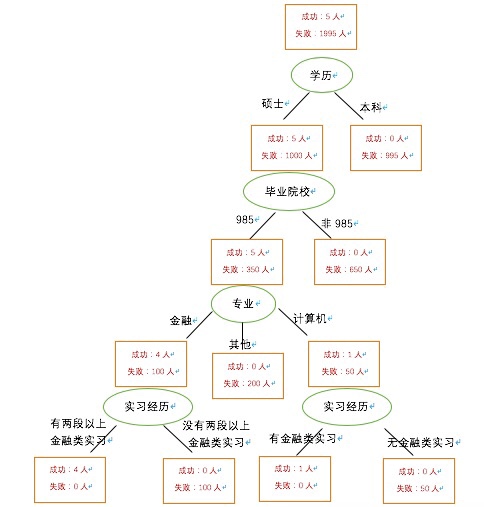
其中的support代表支持度，支持度是指在所有项集中{X, Y}出现的可能性，即项集中同时含有X和Y的概率（或者说，前件和后件同时出现的概率）。计算公式为：支持度=同时包括前件和后件的交易数量/总交易数量。在此项规则中，支持度为0.321，表示在所有交易中，人们同时购买Mascara和Eye shadow的概率为0.321。

confidence代表信心水平（置信度），信心水平表示在先决条件X发生的条件下，关联结果Y发生的概率（包含前件的交易当中同时包含后件的交易的百分比）。计算公式为：信心水平=同时包含前件和后件的交易数量/包含前件的交易数量。在此项规则中，信心水平为0. 8991597，表示在所有含有购买Mascara的交易中，同时包含购买Masacara和Eye shadow的交易数量占0.8991597。

3.4）关联规则的评估标准是置信度（confidence）和提升度（list）。3.1）中，首先刨除提升度小于1的规则，剩下的为前四条规则（因为提升度>1且越高表明正相关性越高，提升度<1且越低表明负相关性越高，提升度=1表明没有相关性）。最有用的规则是规则[1]，因为其提升度最高，置信度排规则[4]之后，而规则[4]提升度远小于规则[1]，并且研究人们买Nail Polish的同时不买Lipstick的概率没有太大意义，所以规则[1]最有用。3.2）中，在提升度大于1的规则中，规则[1]的置信度最高，所以规则[1]最有用。

**Q4**

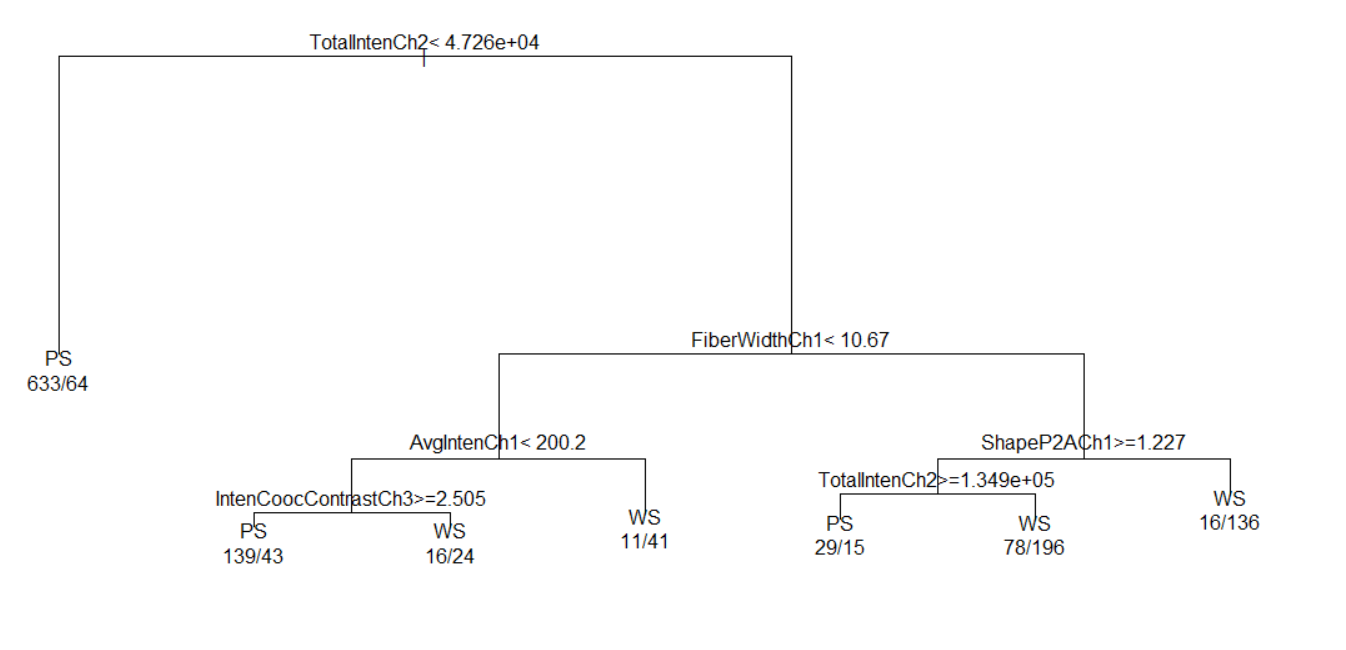
4.1）决策树。 用决策树分析去年的2000份申请材料，可以得到成功申请到这五个岗位的人都满足哪些申请条件。例如，学历是本科还是硕士，毕业学校是否为985，专业是金融类还是非金融类，有无金融实习等等。假如每个岗位只招收1人，根据上述这些条件，我们便可以构造一棵决策树，比如下图：



利用决策树算法和去年的2000份申请材料信息可以得出一个决策树模型（上述图片只是一个简化版本的决策树）。通过这个模型，可以清晰的看到去年申请到岗位的应聘者满足哪些申请条件。

4.2）聚类分析法。可以通过聚类分析把所有的申请者根据各自的申请情况例如学历、毕业院校、专业、工作经历等等来划分成不同的组别，这样就可以得到今年申请者们大概可以分为哪几类人。可以采用层次聚类，得出分好类的树状图，这样郑女士可以通过得出的树状图对申请者们的情况有个大体的认知。

**Q5**



规则：是否TotalIntenCh1>4.726\*10^4

是否 FiberWidthCh1<10.67

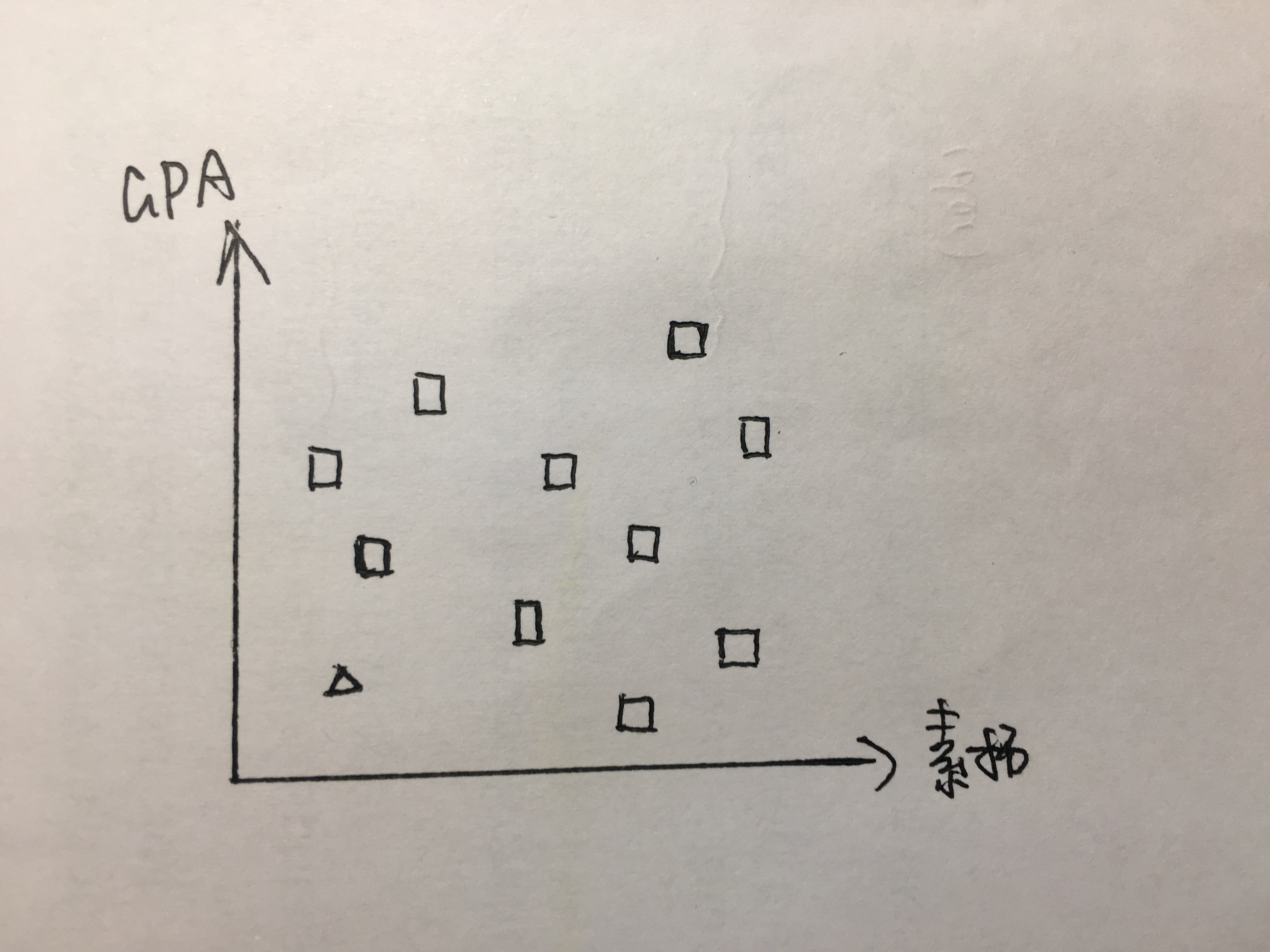
是否 AvgIntenCh1<200.2

是否 ShapeP2ACh1>=1.227

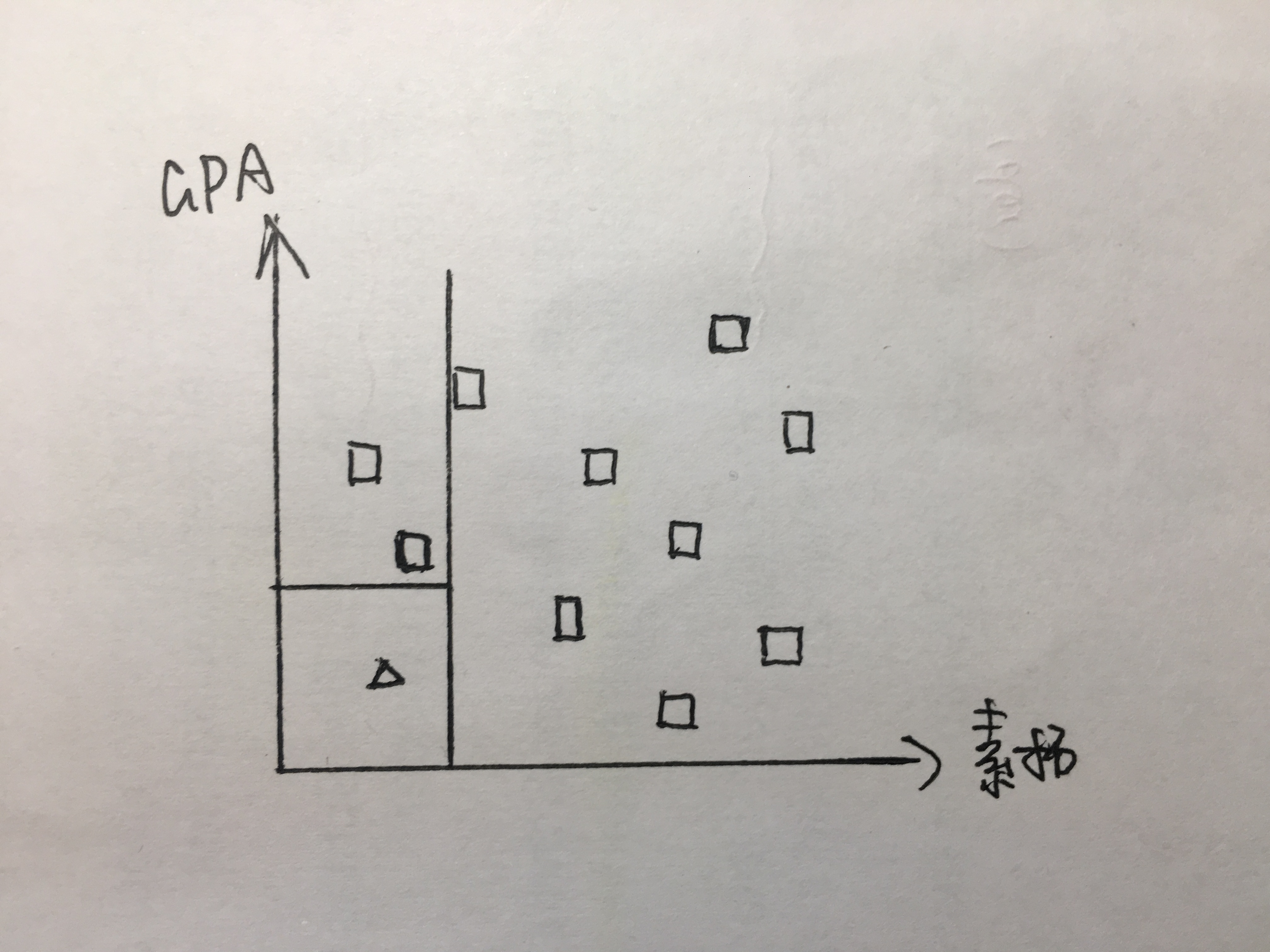
**Q6**

例子：给予退学警告的条件是GPA<1.7且素拓<10。

三角形代表“给予退学警告”，方形代表“不给予退学警告”，由于满足“给予退学警告”条件的人很少，因此“给予退学警告”的样本数很少。



此种情况下，适用于决策树，因为仅需要两次分割，即可得到准确判断。



此种情况不适用于KNN，是因为其样本数据及其不均衡，当K>=3是不可能判断出正确的“给予退学警告”的情况，而当K<=2，也只有较小概率能判断正确。