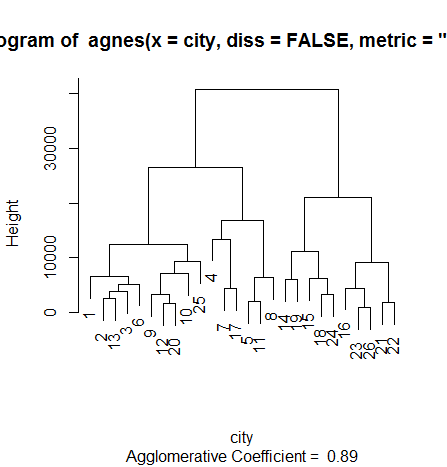
Homework 3

5140809064 张侨伦

1. 第一题： R文件为1RealEstate.R

方法一：聚类分析

1. 聚类分析：根据几个城市的指标聚类分析将26个城市分为4类，使用Hierarchical的聚类分析方法进行分类，树状图如下。



1. 按照上图，取第二次分叉即可得分为四类的结果，结果如下：

第一类：上海，杭州，北京，宁波，南京，青岛，厦门，武汉，沈阳，长沙

第二类：深圳，苏州，广州，天津，大连，天津

第三类：济南，烟台，泉州，扬州，郑州

第四类：重庆，哈尔滨，石家庄，西安，海口

1. 去第一个分叉，就可以分为两类，也就是上面第一类和第二类，第三类和第四类他们之间的距离最近，可以将它们分别合并。

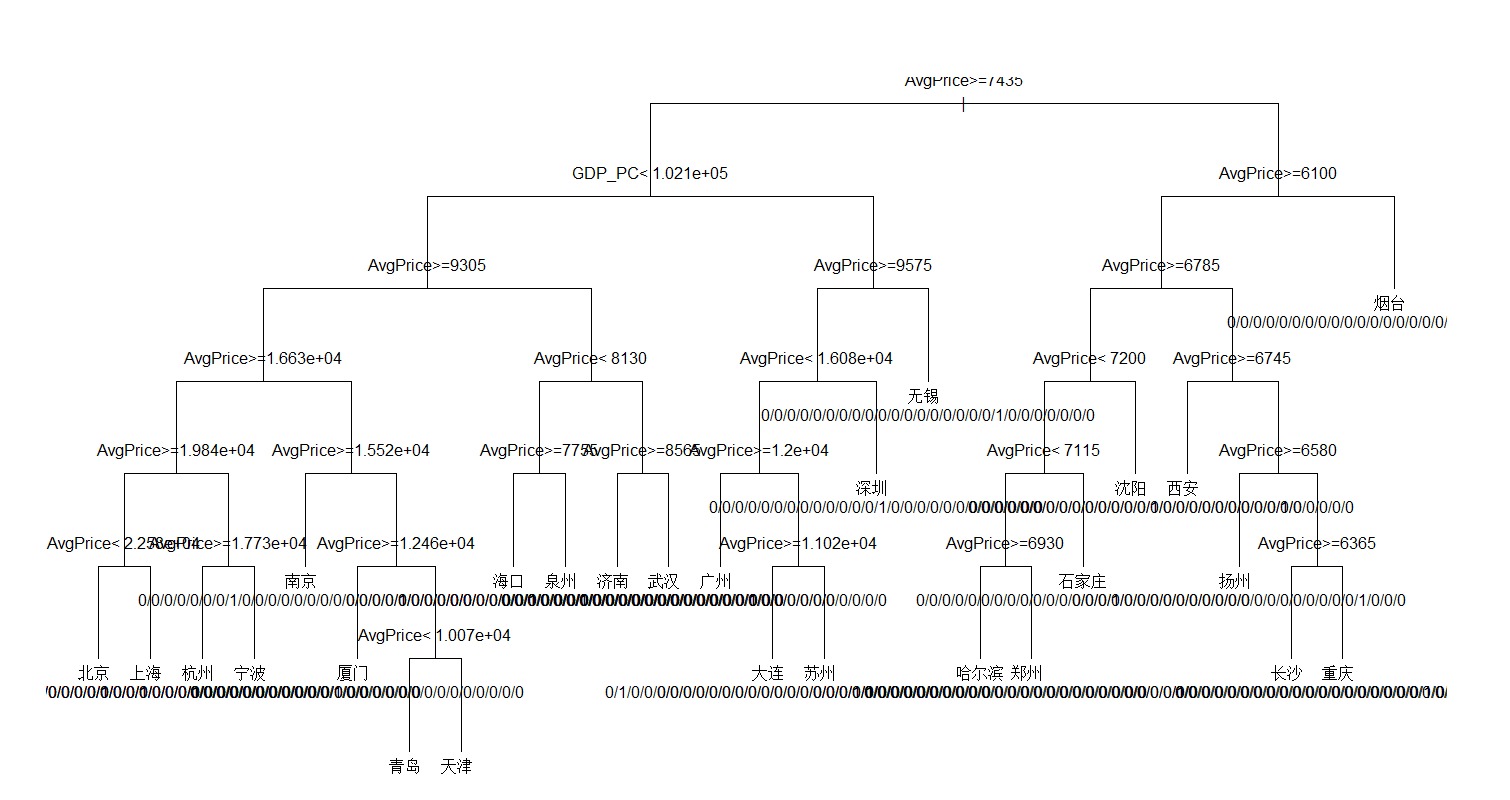
最后的结构就是：

第一类：上海，杭州，北京，宁波，南京，青岛，厦门，武汉，沈阳，长沙，深圳，苏州，广州，天津，大连，天津

第二类：济南，烟台，泉州，扬州，郑州，重庆，哈尔滨，石家庄，西安，海口

方法二：决策树

作图结果如下：



1. 第一类：北京，上海，杭州，宁波，南京，厦门，青岛，天津，海口，泉州，济南，武汉

第二类：广州，大连，苏州，深圳，无锡

第三类：哈尔滨，郑州，石家庄，沈阳，西安，扬州，长沙，重庆

第四类：烟台

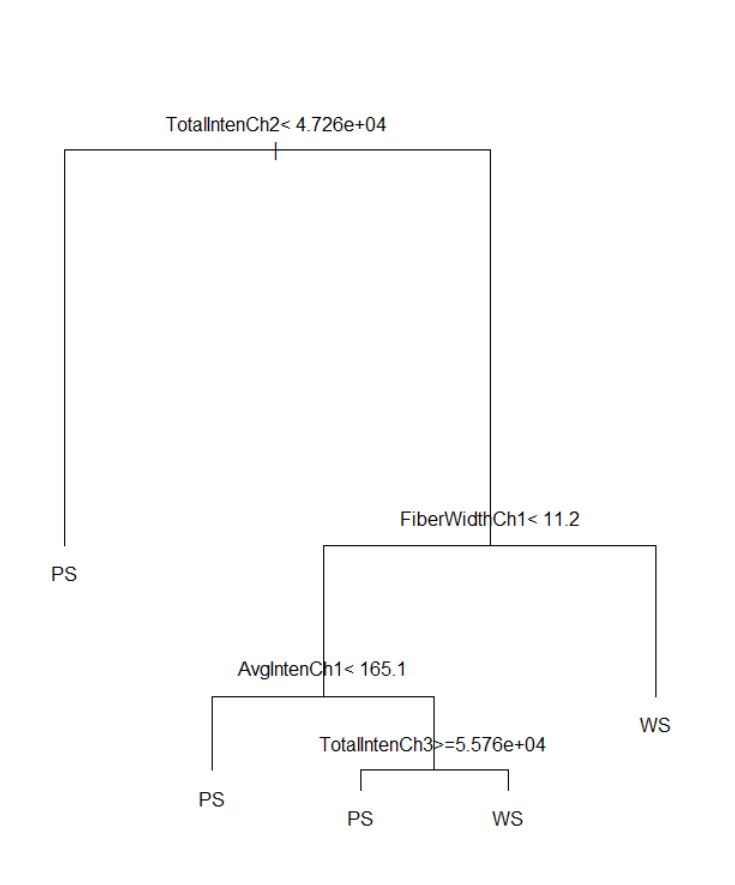
1. 只用将第一次相近的合并在一起就可以，这时候两类就是：

第一类：北京，上海，杭州，宁波，南京，厦门，青岛，天津，海口，泉州，济南，武汉，广州，大连，苏州，深圳，无锡

第二类：哈尔滨，郑州，石家庄，沈阳，西安，扬州，长沙，重庆，烟台

1. 第二题, R文件为2Segmentation.R

决策树：根据数据对市场进行分析，做出图如下：

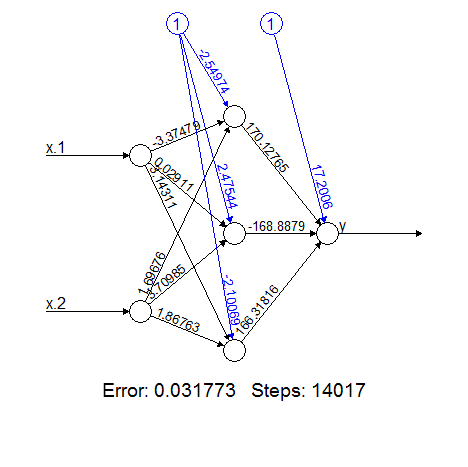


规则为：

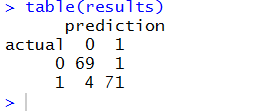
1. 当TotallentenCh2<4.726e+04时，类别为PS
2. 当TotallentenCh2>=4.726e+04且FiberWidthCh1<11.2，且AvgLetenCh1<165.1时，类别为PS
3. 当TotallentenCh2>=4.726e+04且FiberWidthCh1>=11.2时，类别为WS
4. 当TotallentenCh2>4.726e+04且FiberWidthCh1<11.2，且AvgLetenCh1>165.1，且TotallenCh3>=5.576e+04时，类别为PS
5. 当TotallentenCh2>4.726e+04且FiberWidthCh1<11.2，且AvgLetenCh1>165.1，且TotallenCh3<》5.576e+04时，类别为WS
6. 第三题, R文件为3NN.R

人工神经网络

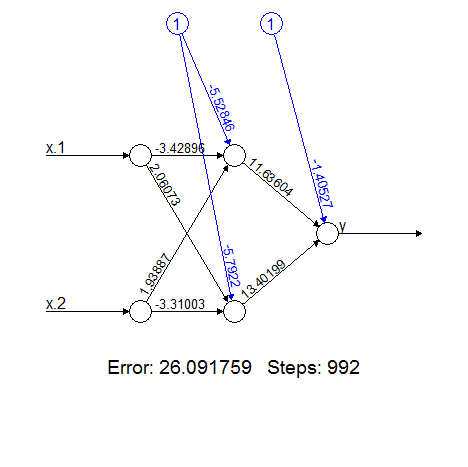
1. 作图结果如下



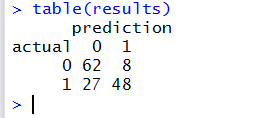
混淆矩阵如下：



1. 作图结果如下

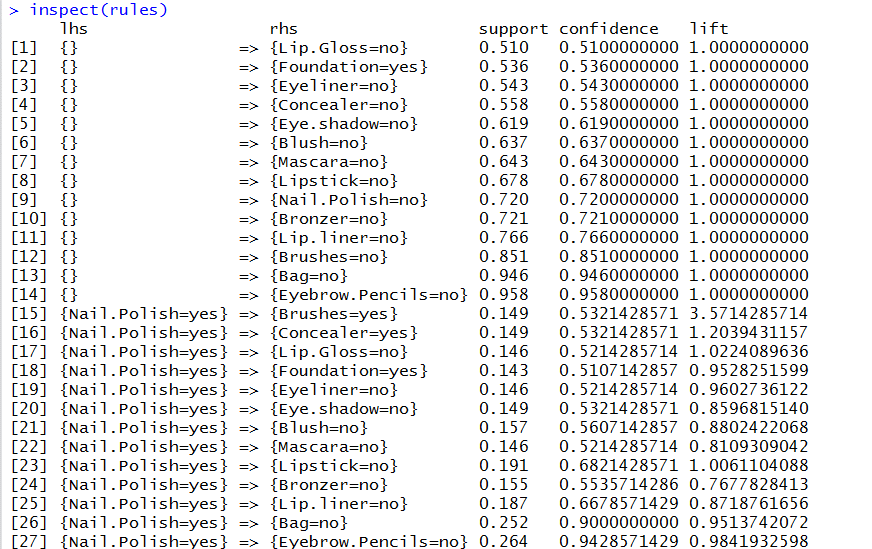


混淆矩阵如下



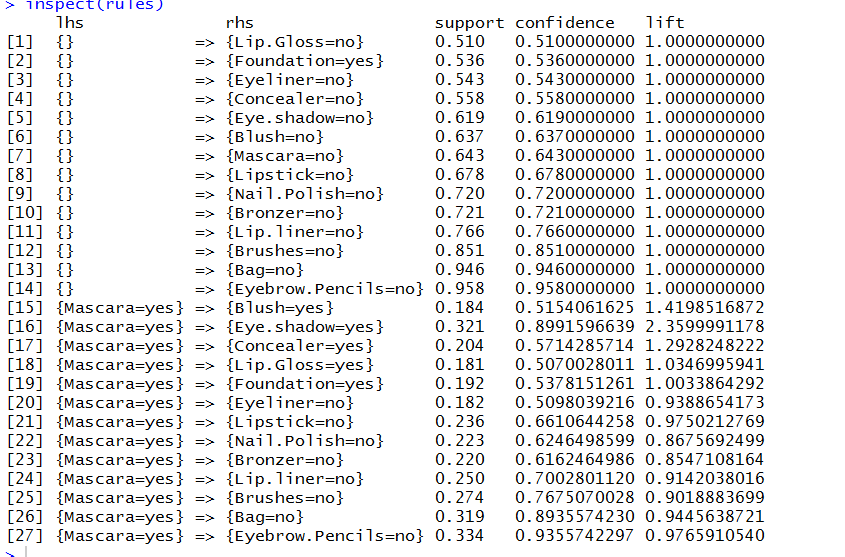
有两个混淆矩阵的结果可知隐藏3个节点比较好

1. 第四题，代码为4Cosmetics.R
2. 可以得到如下图



从中可以看出还会买Brushes，Concealer，Foundation

1. 可以得到如下图



从中可以看出还会买Blush， Eye.shadow, Concealer, Lip.Gloss, Foundation

1. 支持度：同时包括前件和后件的交易数量。

信心水平：包含前件的交易当中同时包含后件的交易的百分比。

比如说如下规则：

就是说Nail.Polish购买了之后同时买了Concealer的交易数量是14.9%，买了Nail.Polish的交易当中同时买了Concealer的交易的百分比约为53%

1. 对问题（a）最有用的规则如下



首先，有用的规则是判断出用户会买另一种商品，那么就只剩下了买Brushes，Concealer，Foundation的几条规则，然后比较这几条规则，这一条的lift比较大，所以这一条比较有用。

对问题（b）最有用的规则如下



首先，有用的规则是判断出用户会买另一种商品，那么就只剩下了买Blush， Eye.shadow, Concealer, Lip.Gloss, Foundation

的几条规则，然后比较这几条规则，这一条的lift，support和confidence都比较突出，整体来说这一条最有用。

1. 第五题
2. 层次聚类方法
3. 因为偶然的数据可能会产生非常明显的类别，但这是不标准的。
4. 1. 了解数据的总体信息（summary statistics）

2. 计算不同分类之间的距离

1. 第六题

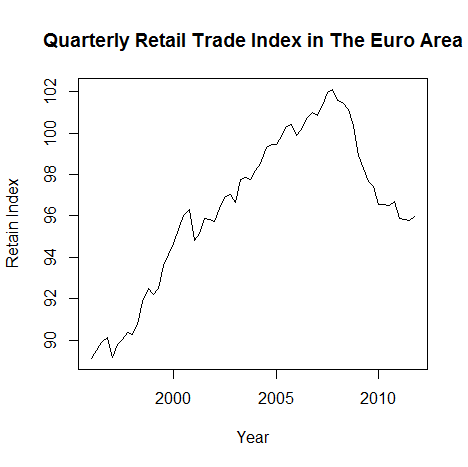
（a）按照郑女士的建议：层次聚类，预期得到去年2000人中被录用和未被录用的分类，然后根据这个将今年的2000人也分到这两类里面，得到一类是可能被录用的，一类是不会被录用的。

（b）非层次聚类：给出指定的类别数量，然后将今年的2000个人分到指定的类别数量里面，这种方法是预先并不知道类别的。

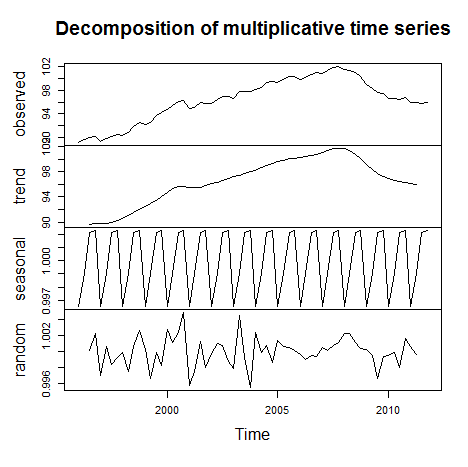
1. 第七题

时间序列画图，代码为7timeseries.R，所得的图如下

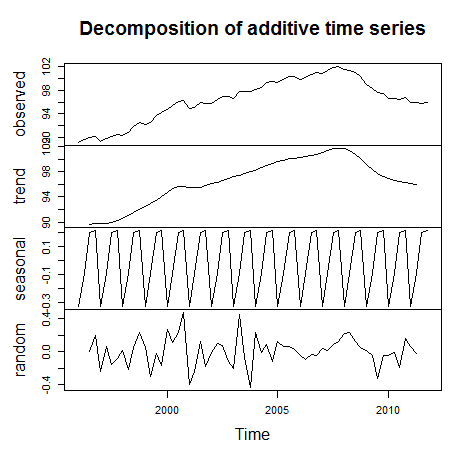
（a）



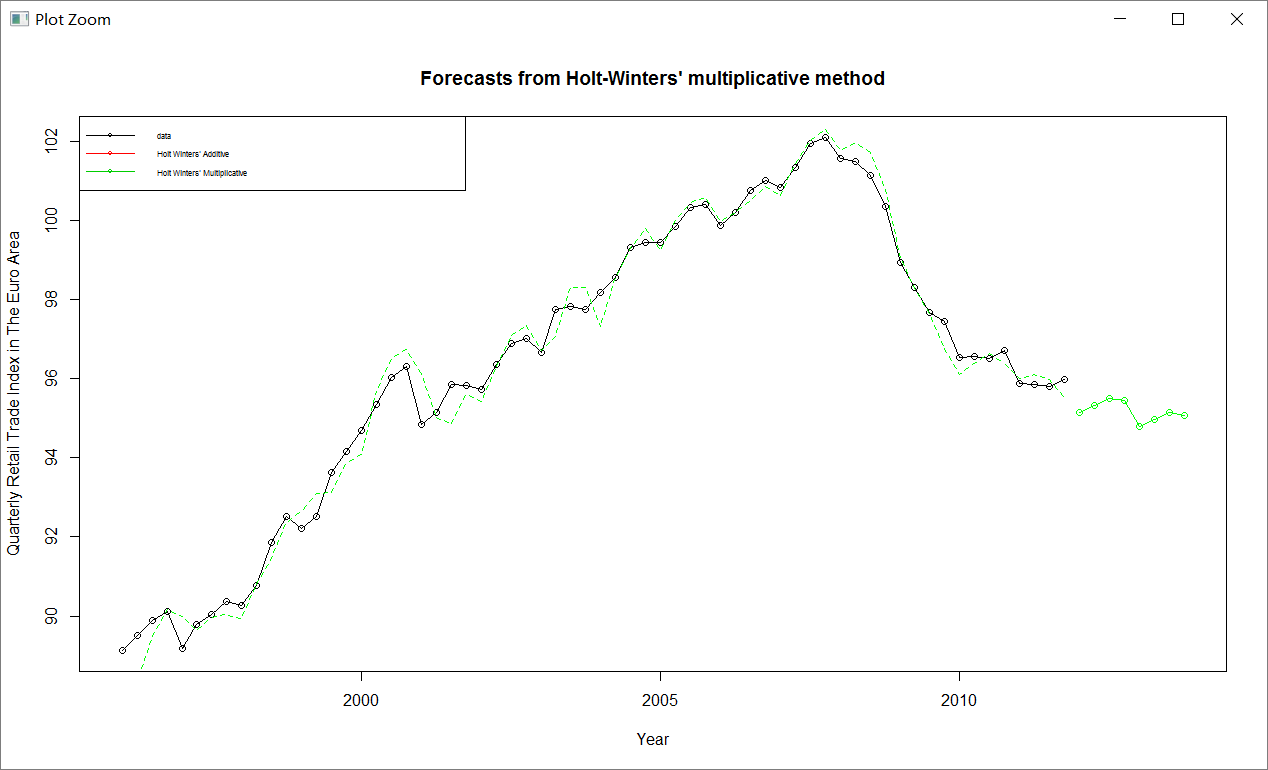
（b）multiplicative form



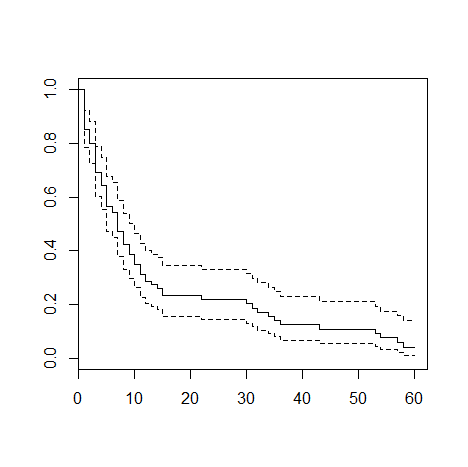
Additive form



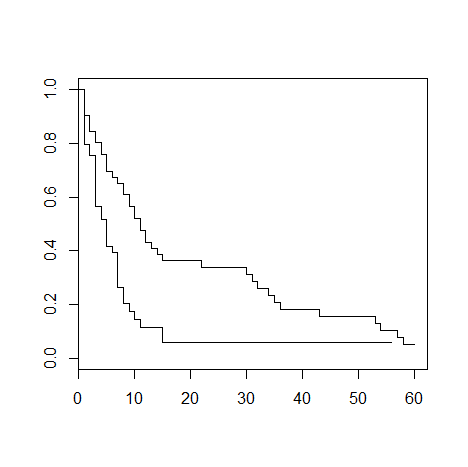
（c）



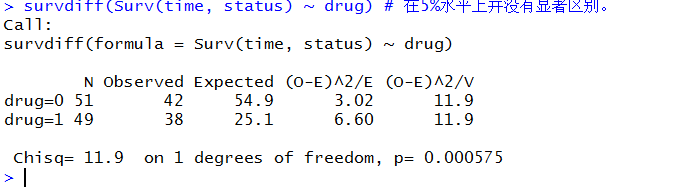
1. 第八题，R语言文件survival.R
2. 作图结果如下



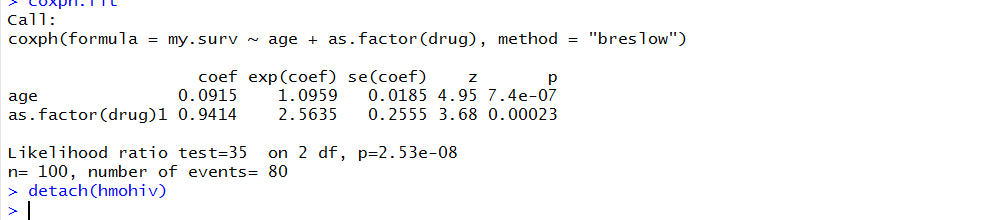
1. 作图结果如下



有以下结果得到p=0.000575，因此在5%显著性水平上有区别



1. 结果如下



由于p=2.53e-08，在5%显著性水平上有显著区别。

由以上结果可以知道年龄和是否吸毒对死亡都有正相关的影响。但是drug的coef比age要大得多，同时吸毒的exp系数更大，综和来说，也就是吸毒对艾滋病患者的影响更大。