**哈尔滨工业大学**

**数据挖掘理论与算法实验报告**

**(2016年度秋季学期)**

**课程编码** S1300019C

**授课教师** 邹兆年

**学生姓名** 杨志烨

**学 号** 16S103168

**学 院** 计算机科学与技术

**决策树实验**

**数据集说明：**

决策树实验使用了三种数据集，分别是Iris、bank、winequality。

**(1).Iris**

样例：

5.1,3.5,1.4,0.2,Iris-setosa

说明：

前4个属性为花的特征，最后一个属性为花的种类。

**(2).wine**

样例：

1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065

说明：

第一个属性代表酒的起源种类，后13个属性代表酒的化学成分

**(3).winequality**

样例：

7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6

说明：

前11个书名为酒的特征，最后一个属性为酒的品级。

**实验环境说明：**

系统：Mac OS 10.11.5

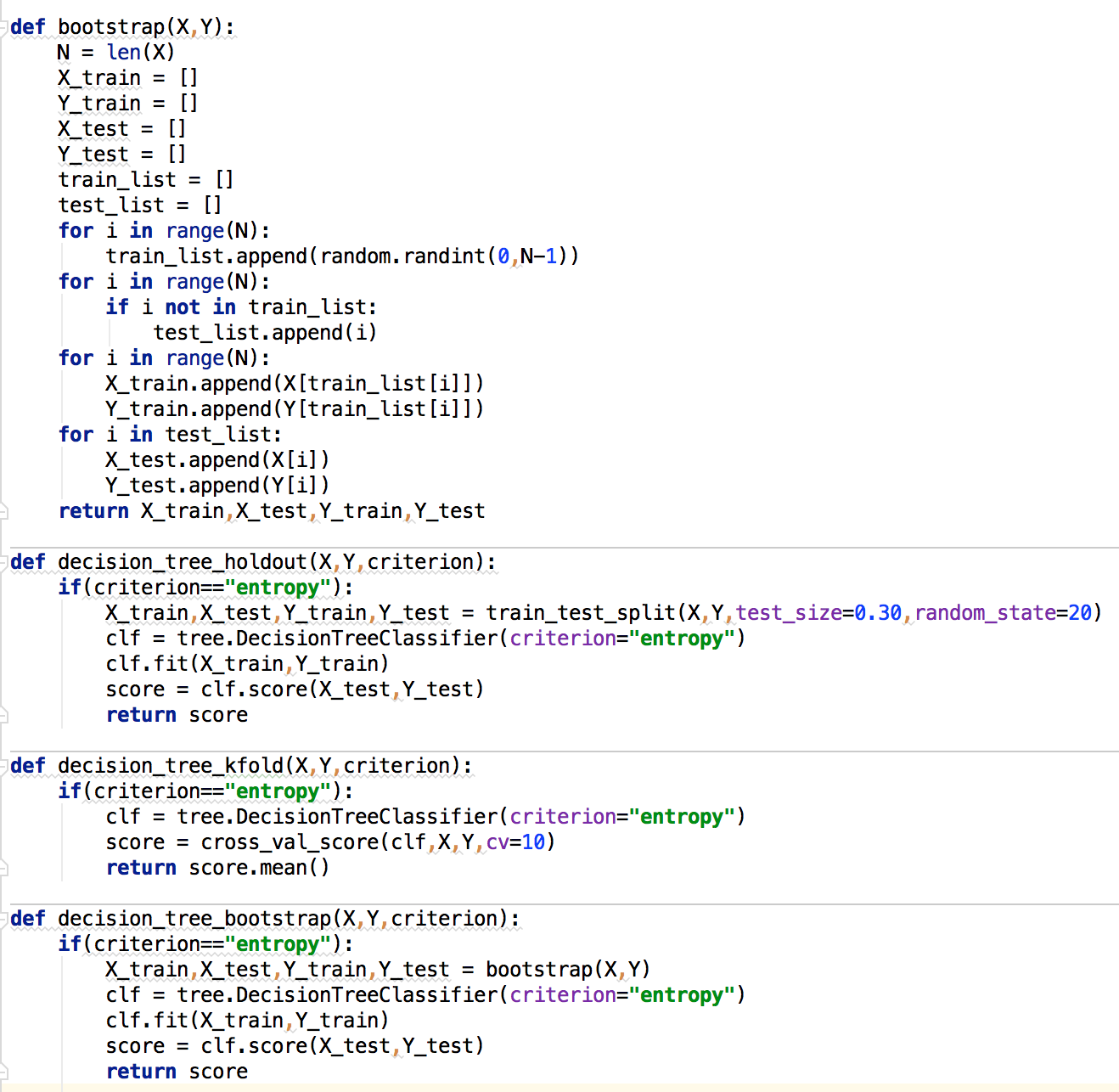
语言：python2.7

IDE：pyCharm

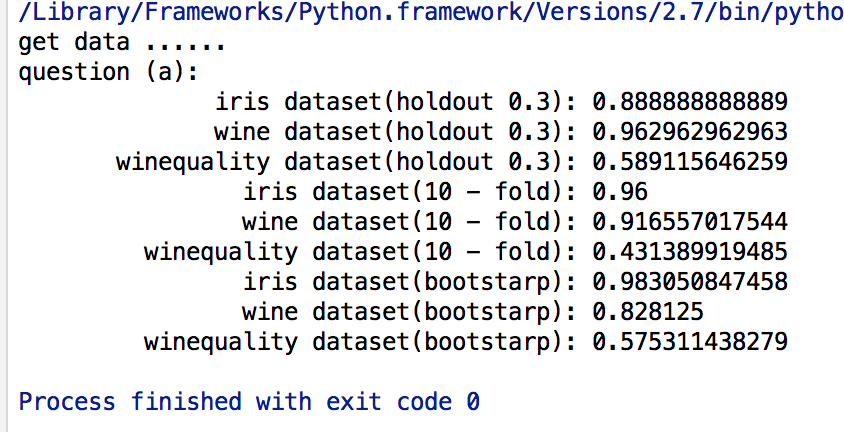
注：代码太多就不粘了，需要的话联系QQ：445922703

**问题(a):使用信息增益作为决策树的分类条件，并使用holdout、10-fold、Bootstrap方法评估决策树。**

答：sklearn包中封装了信息增益分类决策树和k-fold方法，但holdout和bootstrap评估方法需要自己实验一下，关键代码如下：

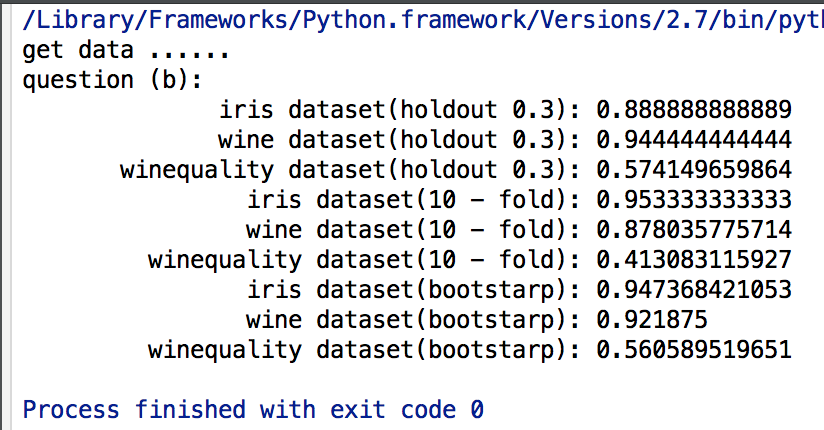


实验结果如下：



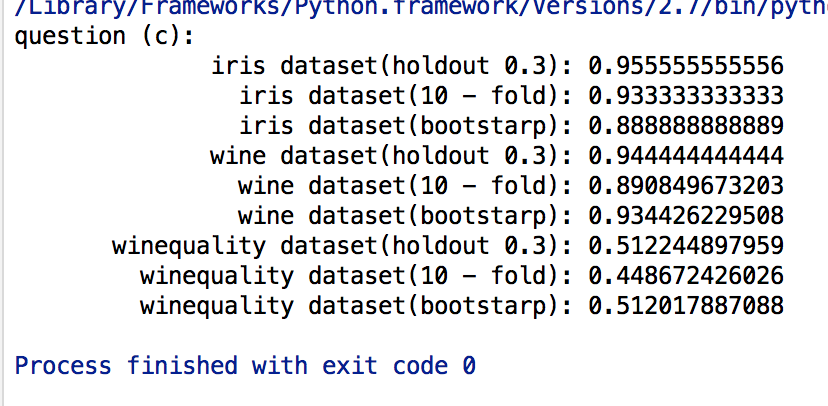
**问题(b):使用Gini作为决策树的分类条件，并使用holdout、10-fold、Bootstrap方法评估决策树。**

答：同问题(a)，只需把参数criterion由”entropy”改为”gini”，实验结果如下：



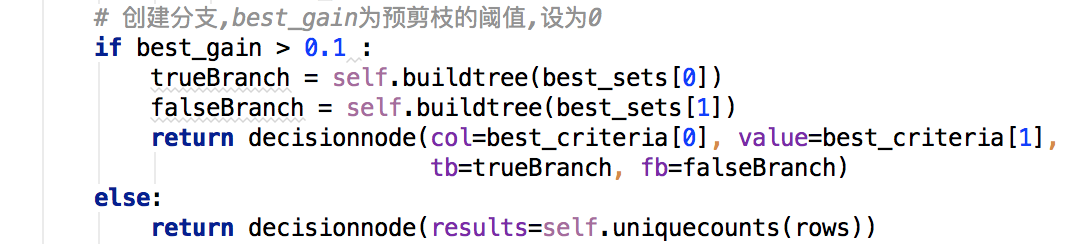
**问题(c):使用误分类率作为决策树的分类条件，并使用holdout、10-fold、Bootstrap方法评估决策树。**

答：sklearn中没有误分类率作为分类判断，在github上找了python实现的决策树，读懂代码之后，加上了误分率方法，结果如下：



**问题(d):实现pre剪枝和post剪枝**

答：关键代码如下：



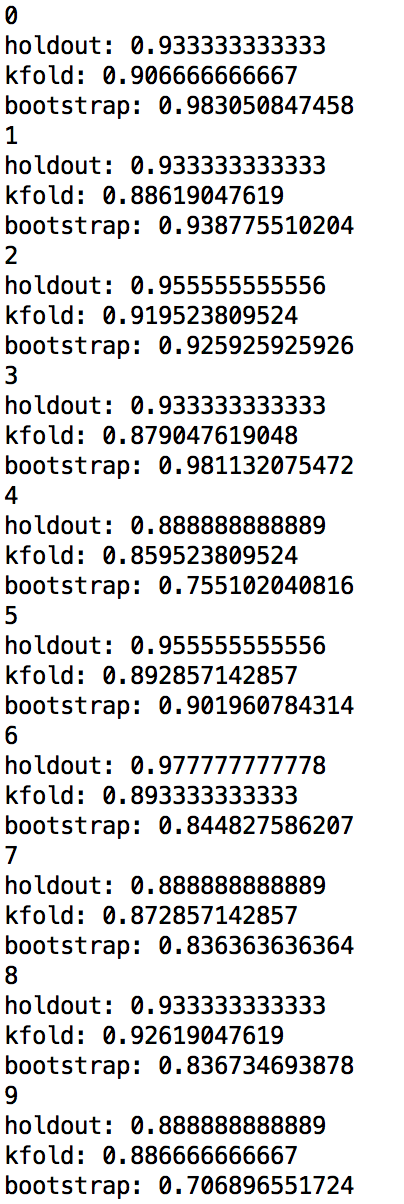
本实验使用最小gain作为阈值进行剪枝，如果gain小于0.1，则剪枝。

**问题(e):决策树处理缺省值**

答：把标签当成特征，缺失值当成标签训练决策树，可以处理缺省值。但是这种情况对多个属性缺失不太适用，之前遇到的情况我都是删掉这条数据，如果有缺失值的数据比较多，那么可以用先验概率或者平均值填充上这个空位。

**问题(f):AdaBoost算法**

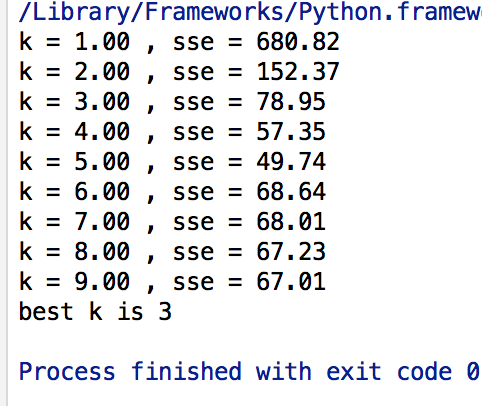
答：Iris AdaBoost结果如下：



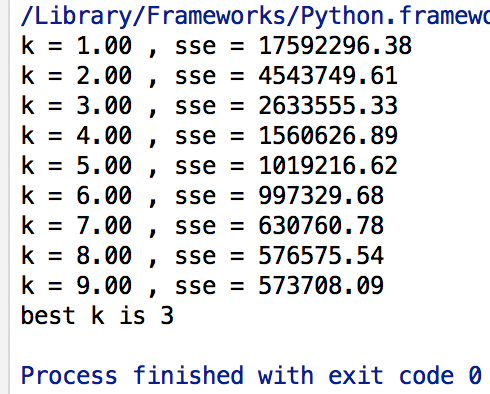
**聚类实验**

**问题(a):解释k对SSE的影响，找到最优的k。**

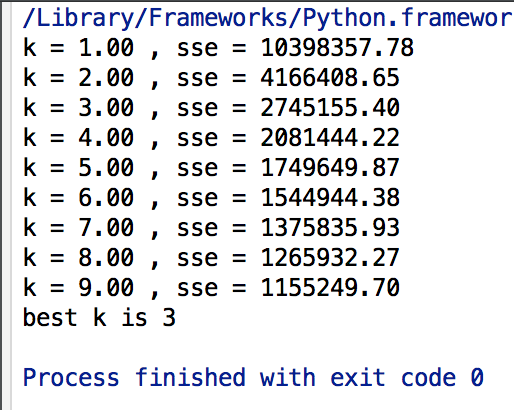
答：对数据集Iris：

****

对数据集wine：



对数据集winequality：



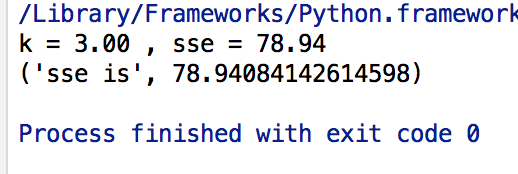
不同的k的取值，当kmeans停止迭代时会对SSE造成影响，当k=1时，SSE很大，当k=样本数量M时，SSE等于0，本实验使用肘部法则，即k值增大的过程中，畸变程度的改善效果下降幅度最大的位置对应的k值就是肘部。

本实验中，对三个数据集进行测试，最优k值都为3.由于数据是有标签的，可以看出前两个数据集k值选取正确，最后一个数据集k值选取错误(观察数据k选4更合理)。

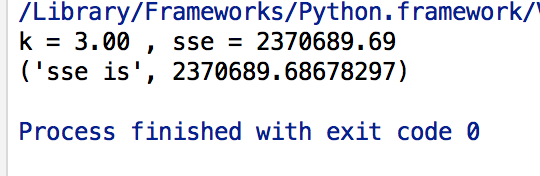
**问题(b):对于确定的k值，解释初始化质心对聚类质量的影响。**

当k=3时，随机初始化质心位置10次，计算每次初始化之后的SSE，选择对应SSE最小的那组质心初始化位置。

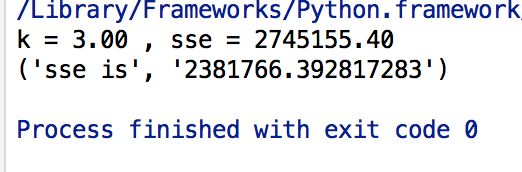
对Iris：



对wine：



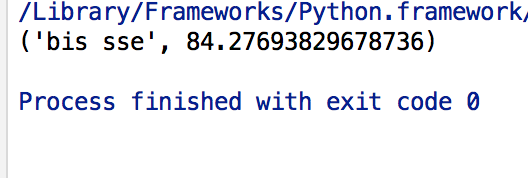
对winequality：



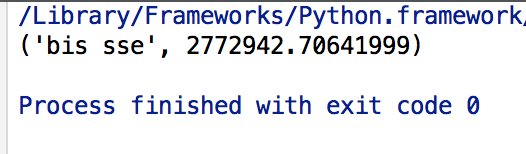
可以看出，当k值一定时，不同的质心初始化对，对聚类效果有影响，当选择初始化SSE小的质心作为初始质心时，通常最后的SSE会变小，即聚类效果更好。

**问题(c):实现并比较二分k-means算法和普通k-means算法**

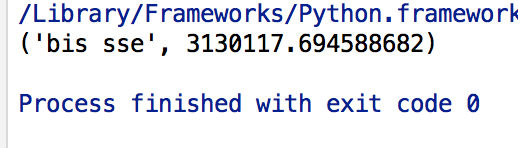
对Iris：



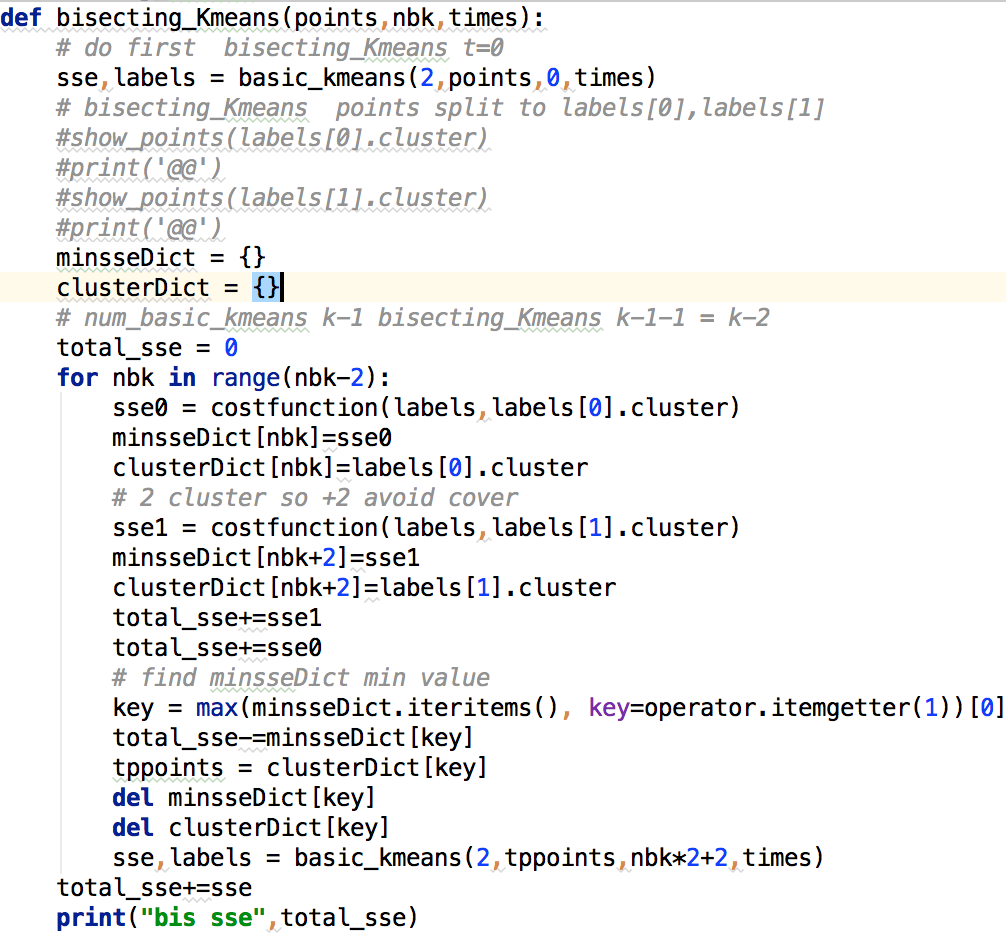
对wine:



对winequality:



主要代码



通过SSE结果可以看出，二分kmeans对Iris数据集表现好，对wine和winequality数据集表现不好，查阅资料总结如下：

1. 相比于普通kmeans，二分kmeans算法的执行速度更快，因为每次不是计算所有点，时间复杂度由n下降至logn。
2. 二分kmeans不受初始化问题的影响
3. 但二分kmeans也是保证也每一步的误差最小，但同样无法收敛于全局最优。