**k-NN**

1. **实验要求：**
   1. 1) Develop a k-NN classifier with Euclidean distance and simple voting
   2. 2) Perform 5-fold cross validation, find out which k performs the best (in terms of accuracy)
   3. 3) Use PCA to reduce the dimensionality, then perform 2) again. Does PCA improve the accuracy? Which dimensionality is better?
2. Plus (at most 3/10)
   1. Try other distance metrics or distance-based voting (+0.5~1)
   2. Try other dimensionality reduction methods (+1~2)
   3. Perform feature selection (+1)

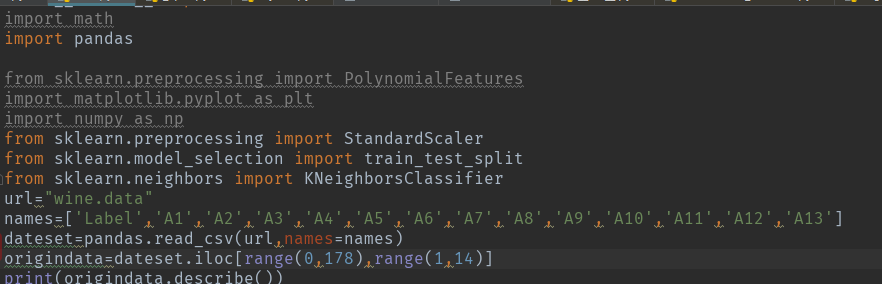
How to set the k value, if not using cross validation? Verify your idea (+3)

**二. 实验步骤：**

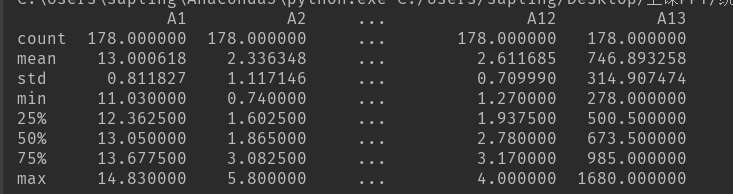
**要求1:采用欧氏距离来并编写一个k-nn分类器。**

首先我们先把数据加载进去，并且可视化数据,这里我们是采用了read\_csv函数先把数据读出来，然后在使用descirbe函数去统计每个特征集的各个值。

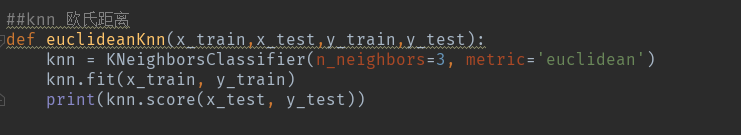
具体代码如下：



**实验结果:**



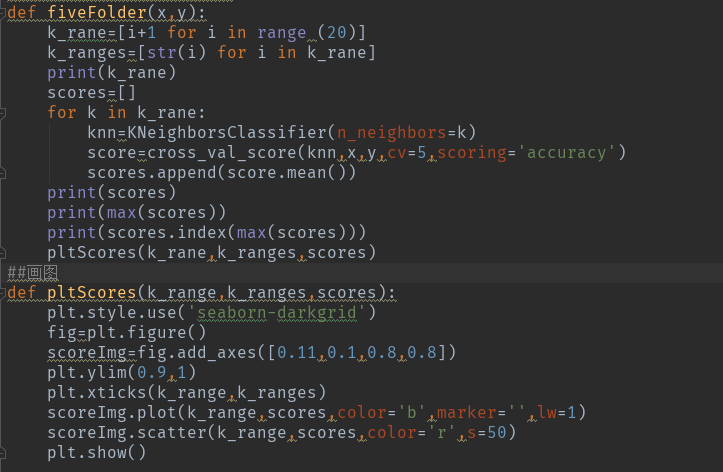
采用欧式距离做简单的knn选择。具体代码如下：



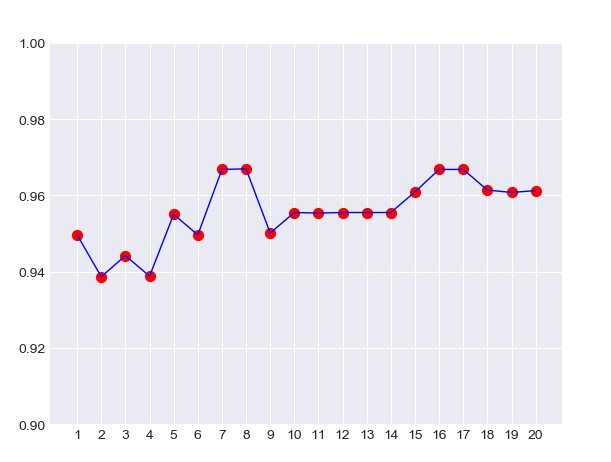
实验结果： 分类准确率为1.0

**要求2:** **执行5折交叉验证，找出在准确性方面表现最好的k**

具体实现代码：



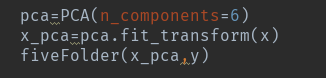
实验结果：从实验结果可以看出k值在8的的时候最大，分类的准确率最高，所以我们使用k值为8，准确率为：0.966940469。



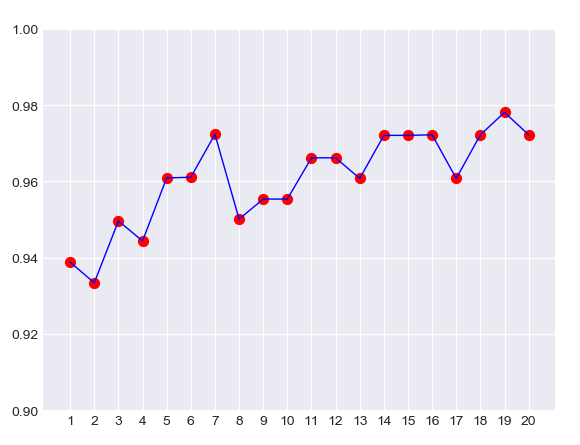
**要求3:** **采用PCA降维**

首先采用pca降维，然后重复要求二。得到最大的值。

代码如下:



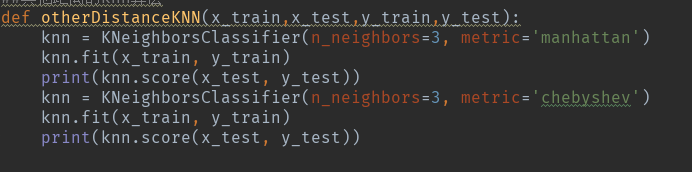
实验结果如下：从实验结果可以看出来在采用pca进行降维了以后我们的准确率发生了变化，现在在k取19的时候最大，准确率为0.9780780708.



**要求4:** **采用不同距离的KNN**

这里我们另外选取两种距离方式，一种曼哈顿距离，一种切尔雪夫距离；

具体代码：

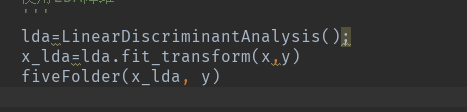


实验结果：从实验结果中我们可以看出来，曼哈顿距离的准确率为1.0.切尔雪夫的距离为0.93333。

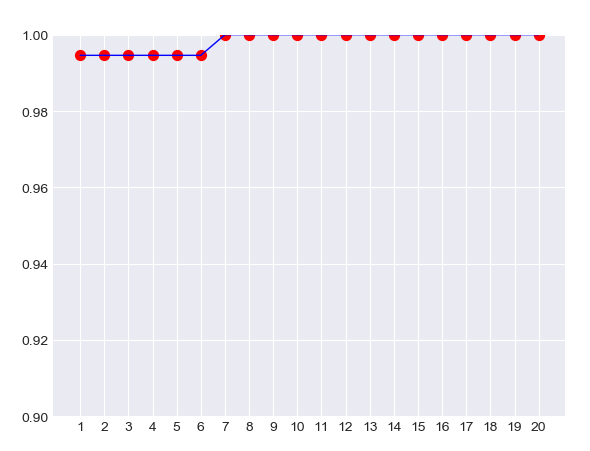


**要求5:** **采用LDA进行降维：**

具体代码如下:



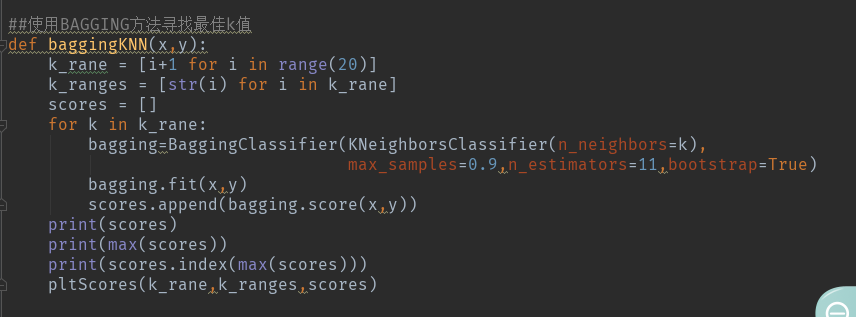
**实验结果：从实验结果可以看出来，使用LDA降维比PCA的效果好许多。**



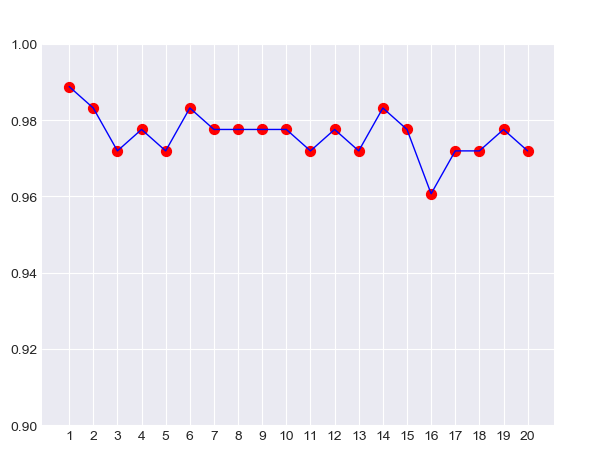
**要求5:** **使用BAGGING方法寻找最佳k值**

除了K折交叉验证法，我们还可以利用委员会算法寻找最佳K值。Bagging 是其代表，其基本思想是i.对一个单独数据集，利用bootstrap（放回抽样）产生M个自助数据集。ii.对每一个K值，我们利用这M个自助数据集训练M个独立模型iii.对每一个实例，委员会预测为M个独立模型的平均值**。**

**具体代码如下：**



**实验结果:** 我们发现K=1时，取到最高准确率。这是因为K=1时，模型方差很大，但是加和求平均后偏置很小，从而产生了预测效果的提升。利用bagging算法，我们选取K=6,此时准确率为0.983

**.** 

.参考文献

[1]李航 统计学习方法