**Homework2 KNN高斯分布 实验报告**

**实验要求：**

有2类二维空间点，A类和B类。

A类点以（0，0）为中心、（1，0；0，1）为协方差矩阵的二维高斯分布；

B类点以（1，2）为中心、（1，0；0，2）为协方差矩阵的二维高斯分布；

随机生成300个A类点，200个B类点，并用k-最近邻的方法进行分类。

**实验思路：**

1. 生成数据集：首先生成这500个点，以列表的形式表示位置坐标，再把所有的坐标放入一个大的列表中。再生成一个label列表，包含300个’A’和200个’B’。
2. 为了更加清晰直观，把生成的300个A类点和200个B类点分别用两种颜色的符号绘制在一个二维坐标系内。
3. 划分数据集。采用十折交叉验证的方法，以9:1划分数据集，分别作为训练集和测试集，每次对每个集中的每个点都对应正确的label。
4. 对于每个待测试的k，用十折交叉验证对十个不同的测试集中的各个点进行测试，若测试结果与点本身的标签相符，则记为正确一次。共测试十轮，统计正确率
5. 比较不同k值下的正确率，取出正确率最高的k值，作为best k。
6. 增加一些附加功能，如绘制正确率与k的折线统计图、在原先的ab点分布图中绘制出决策面来划分AB两个区域

**实验原理：**

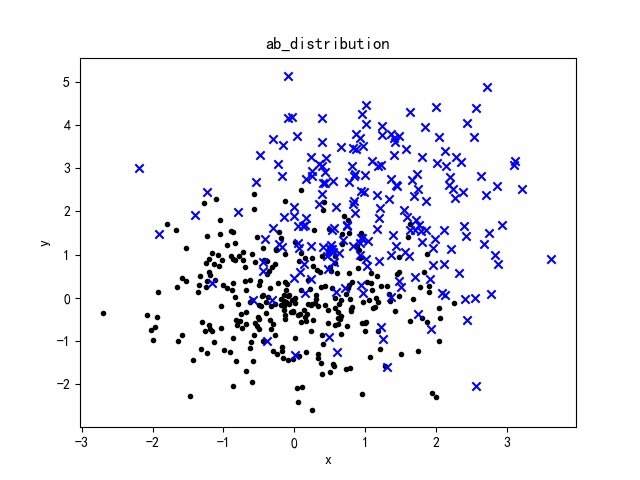
KNN(K-Nearest Neighbor)：k近邻法是一种分类决策方法，规则往往是多数表决，即由输入实例的k个邻近的训练实例中的多数类，决定输入实例的类。

K折交叉验证：就是把数据集以(k-1):1的比例分成训练集和测试集，而分出来的每1/k份轮流作为测试集，剩余的作为训练集。因此共能测试k次，且每一次的测试集都不会与其他次的测试集有重复部分。这样扩大了实验数据样本，也保证了科学性和实验结果的可靠性。

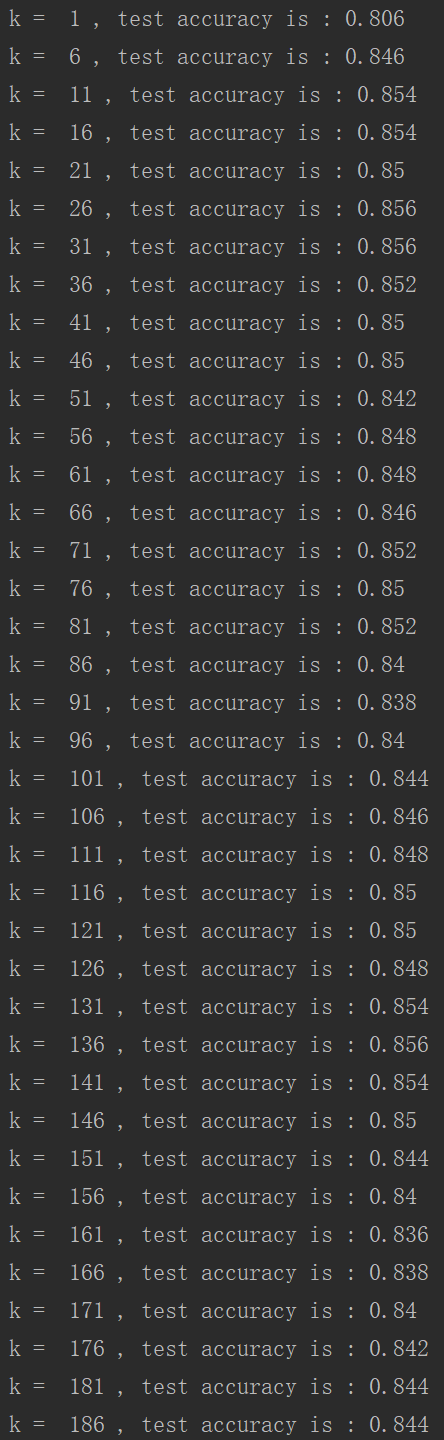
在本实验中，训练集共450个点，测试集共50个点，测试十轮，共能收集到500个测试结果，取其中分类结果与点本身标签正确匹配的点的总个数，除以500，即为这个k下的正确率。

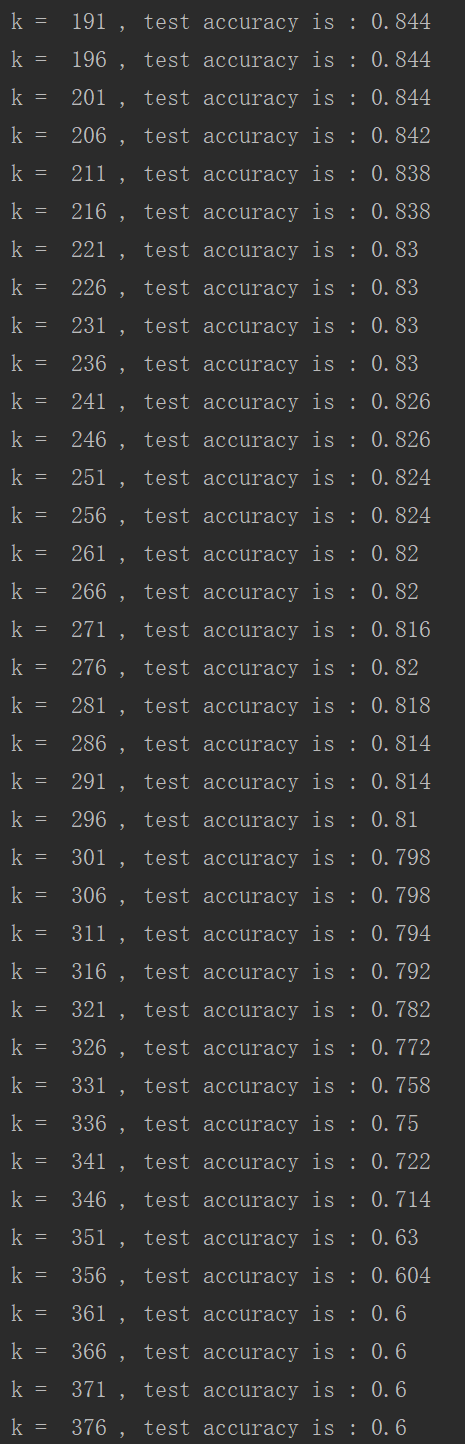
**实验结果：**

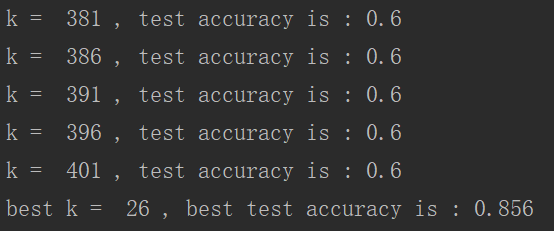
这是生成的所有500个点，黑色代表A点，中心为(0,0)，蓝色代表B点，中心为(1,2)。



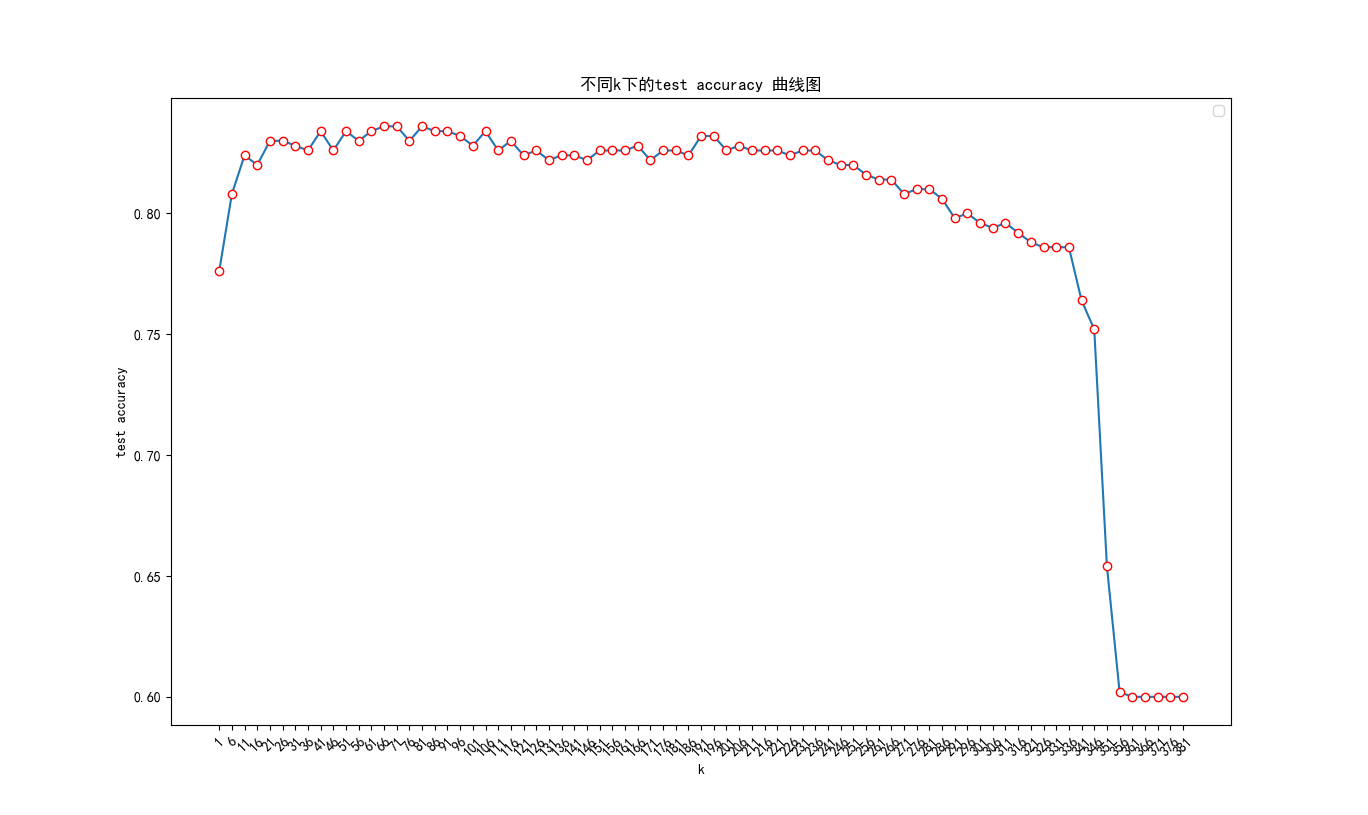
选取了其中一次的实验结果：







该次试验中，最佳的k值是26，对应的最高准确率是85.6%。而进行多次测试发现，最佳的k值在20左右到100左右之间不定，k值在这个范围内时准确率都较高，且比较接近。最高的准确率都在85%左右。因此k的最佳取值在20到100之间。

这是根据结果绘制出的k-accuray曲线：

这是利用python的plt.pcolormesh函数，经过knn决策划分的AB区域分类图：

