模式识别第一次作业

2120190442 周宝航

**目录**

[**一、** **问题描述** 1](#_Toc21966617)

[**二、** **基本思路** 2](#_Toc21966618)

[**三、** **算法描述和算法实现** 2](#_Toc21966619)

[**四、** **结果与分析** 5](#_Toc21966620)

1. **问题描述**

K最近邻算法是模式识别的基本算法之一。改算法对测试模式的预测结果为与该模式距离最近的K个模式类别的众数，其中：距离度量为可测的，K为模型的超参数。

使用文件iris.txt中的数据作为实验数据。在该数据文件中，每行为一个模式，前四个数据为模式特征，最后一个数据为模式类别标签。

针对下列问题，给出具有统计意义的答案及解释：

1. 最优K值。
2. 最优预测正确率。
3. **基本思路**
4. 问题1：

首先将数据集按照8:1:1的比例划分为训练集、验证集、测

集。然后，我们设置待选超参数。接着，选取超参数集合中的每个K值得到分类模型，并在训练集上进行训练。最后，我们在验证集上进行测试得到预测准确率。我们选取验证集上最高准确率所对应的K值为最优超参数。

2. 问题2：

在上一问中，我们已经得到了最优K值。现在，我们取该K值得到分类模型。然后，我们将训练集与验证集合并为新训练集，并重新训练分类模型。最后，将训练好的模型在测试集进行测试，得到最优预测正确率。

1. **算法描述和算法实现**
2. **实验设计过程：**

**实验数据集：**

Iris数据集的统计特性如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 模式特征1 | 模式特征2 | 模式特征3 | 模式特征4 |
| 均值 | 5.843333 | 3.054000 | 3.758667 | 1.198667 |
| 方差 | 0.828066 | 0.433594 | 1.764420 | 0.763161 |
| 极小值 | 4.300000 | 2.000000 | 1.000000 | 0.100000 |
| 极大值 | 7.900000 | 4.400000 | 6.900000 | 2.500000 |

其中，数据集共包含150个样本数据，分为3类。各个类别的统计情况如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 1 | 2 | 3 |
| 数量 | 50 | 50 | 50 |

**实验分类器：**

K近邻分类器（超参数包括：K近邻数、距离度量）

在本次实验中，我们采用闵可夫斯基距离作为度量函数：

其中，p的取值会影响距离度量。比如：当时，上述公式即为曼哈顿距离；当时，上述公式即为欧式距离。在后面的结果与分析中，我们将对这两个超参数进行讨论。

**实验评估指标：**

在上面对数据集的统计情况来看，Iris数据集不存在数据类别不

平衡的问题。所以，我们采用准确率作为实验的评估指标。

**实验结果的统计可靠性：**

在实验中，数据集的划分也会对实验结果产生影响。为了保证我

们实验结果的统计可靠性，我们将每组超参数的实验重复1000次，并对这1000次实验的准确率求取均值与方差作为最终评估模型表现的依据。

1. **算法分析及核心代码：**
2. **距离度量函数**



其中参数a为待测试的样本向量，参数b为训练集样本空间矩

阵，参数p为决定闵可夫斯基距离的数值。

1. **K近邻分类器类**



K近邻分类器类的构造参数包括：闵可夫斯基距离的p数值与近邻数k。其calc\_diatance方法负责计算样本间的闵可夫斯基距离。而predict方法则按照K近邻算法进行分类。K近邻算法的伪代码如下所示：

1、计算已知类别数据集中的点与当前待测试样本之间的距离

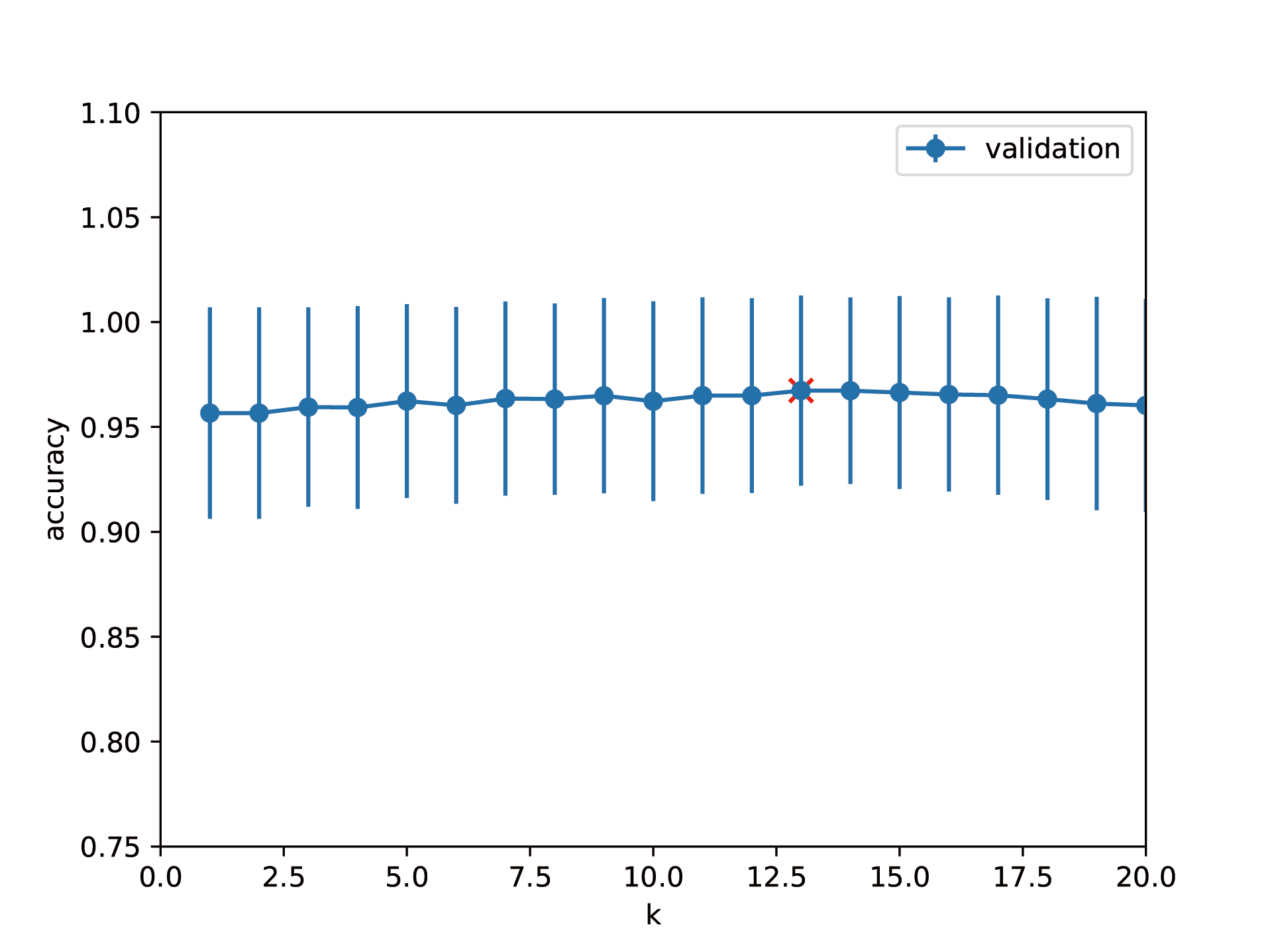
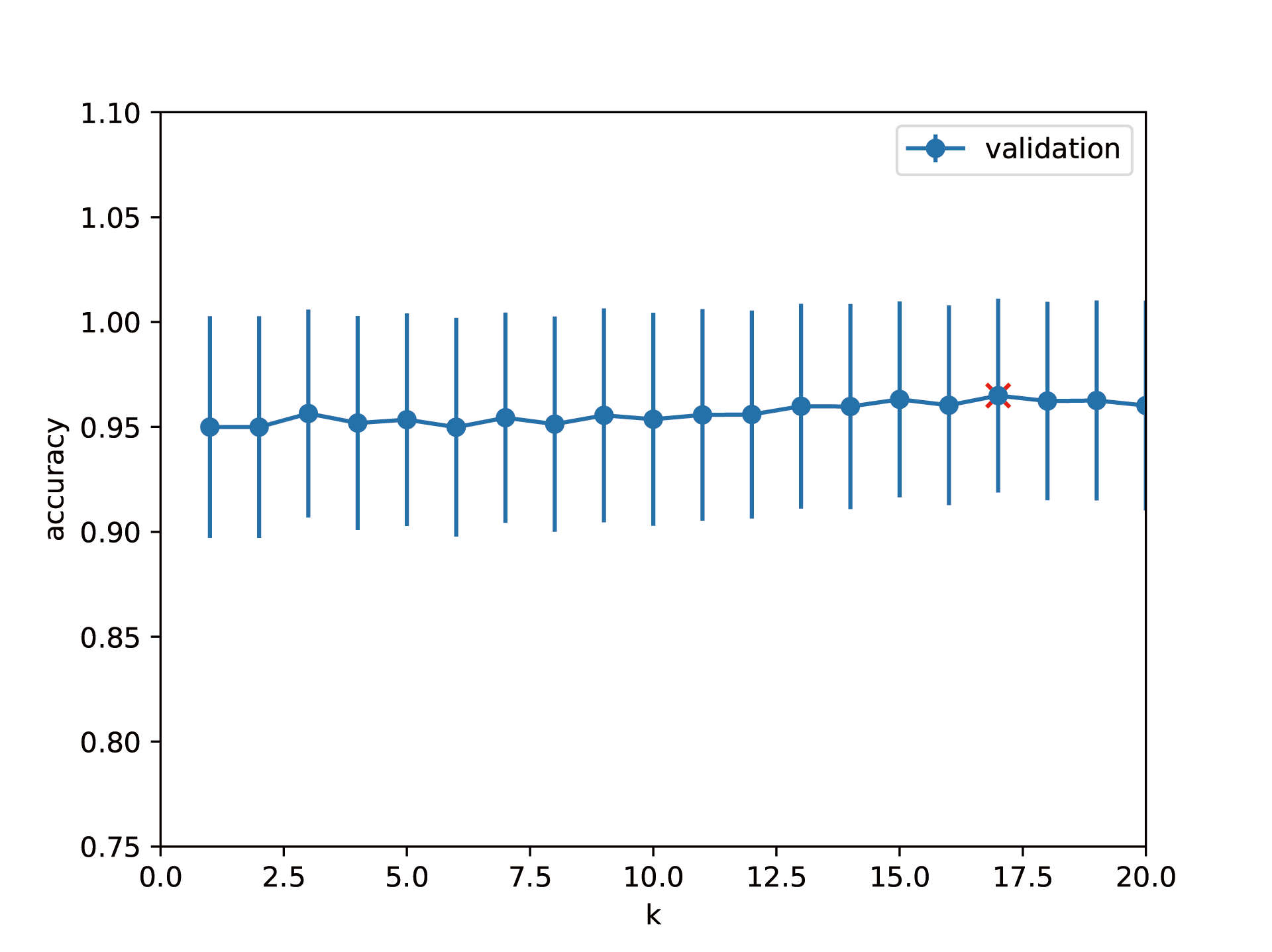
2、按照距离递增次序对距离进行排序

3、选取与当前待测试样本距离最小的k个点

4、统计k个点所属类别的出现频率

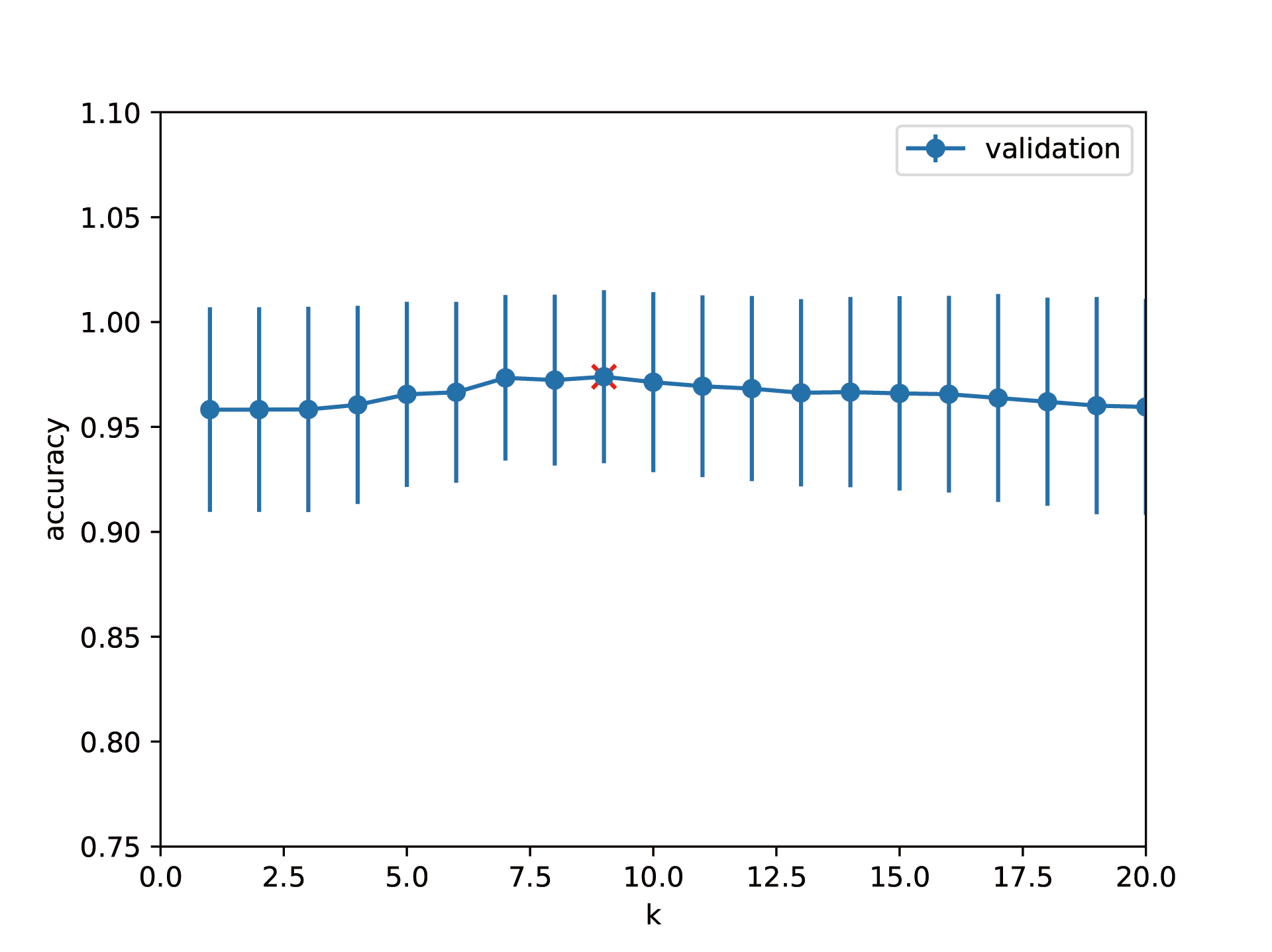
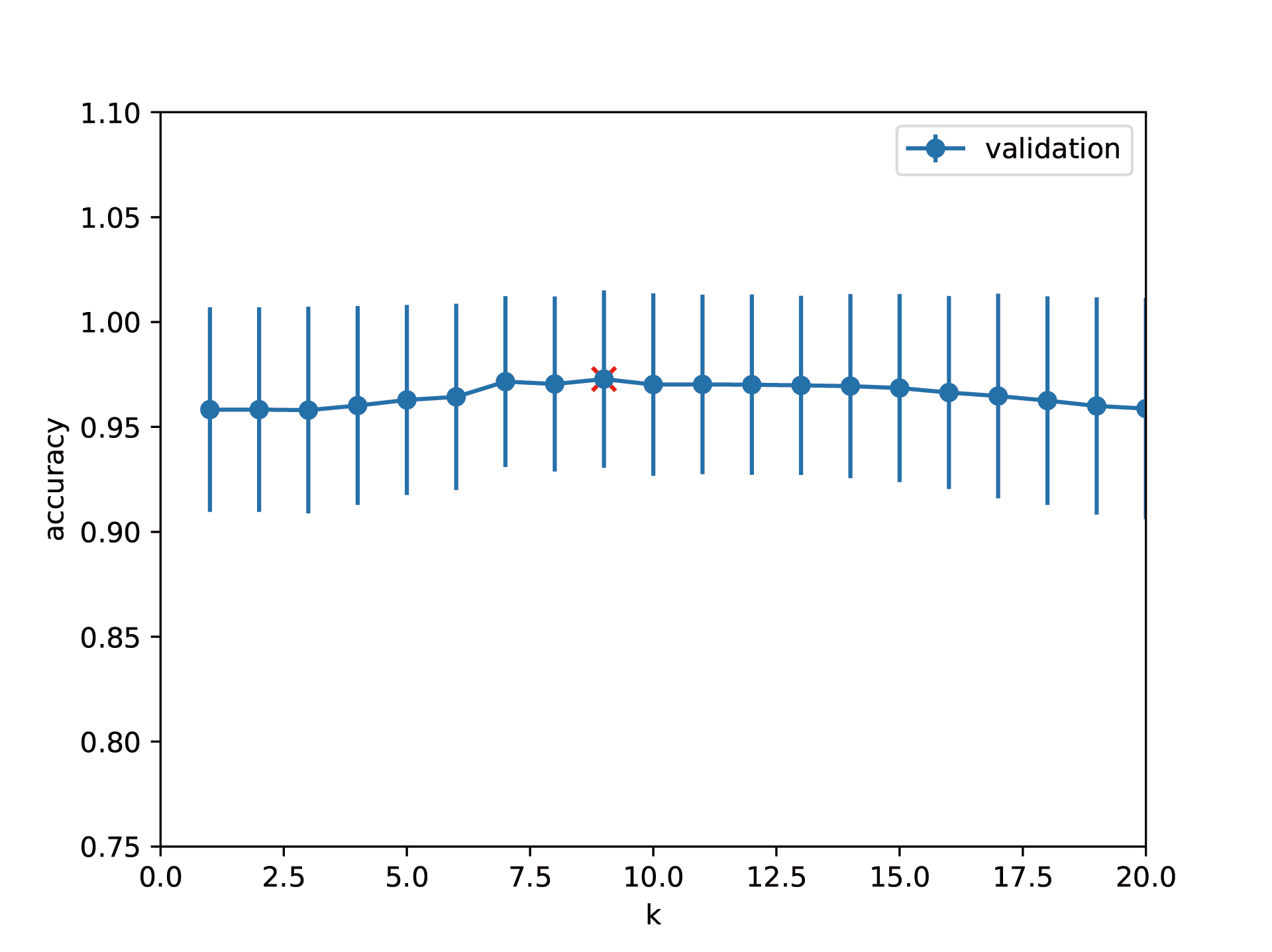
5、返回出现频率最高的类别作为待测试样本的预测分类

1. **结果与分析**



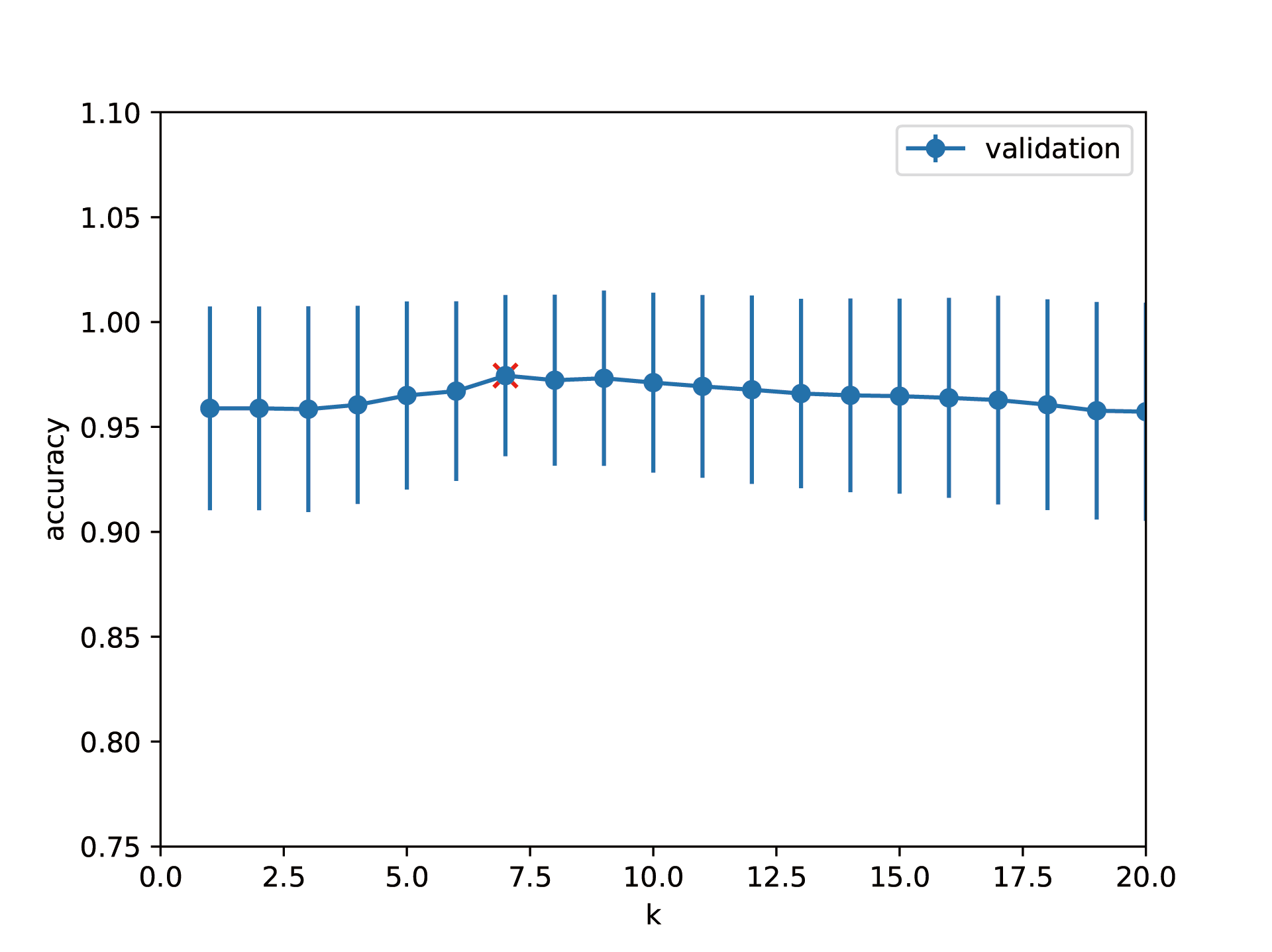
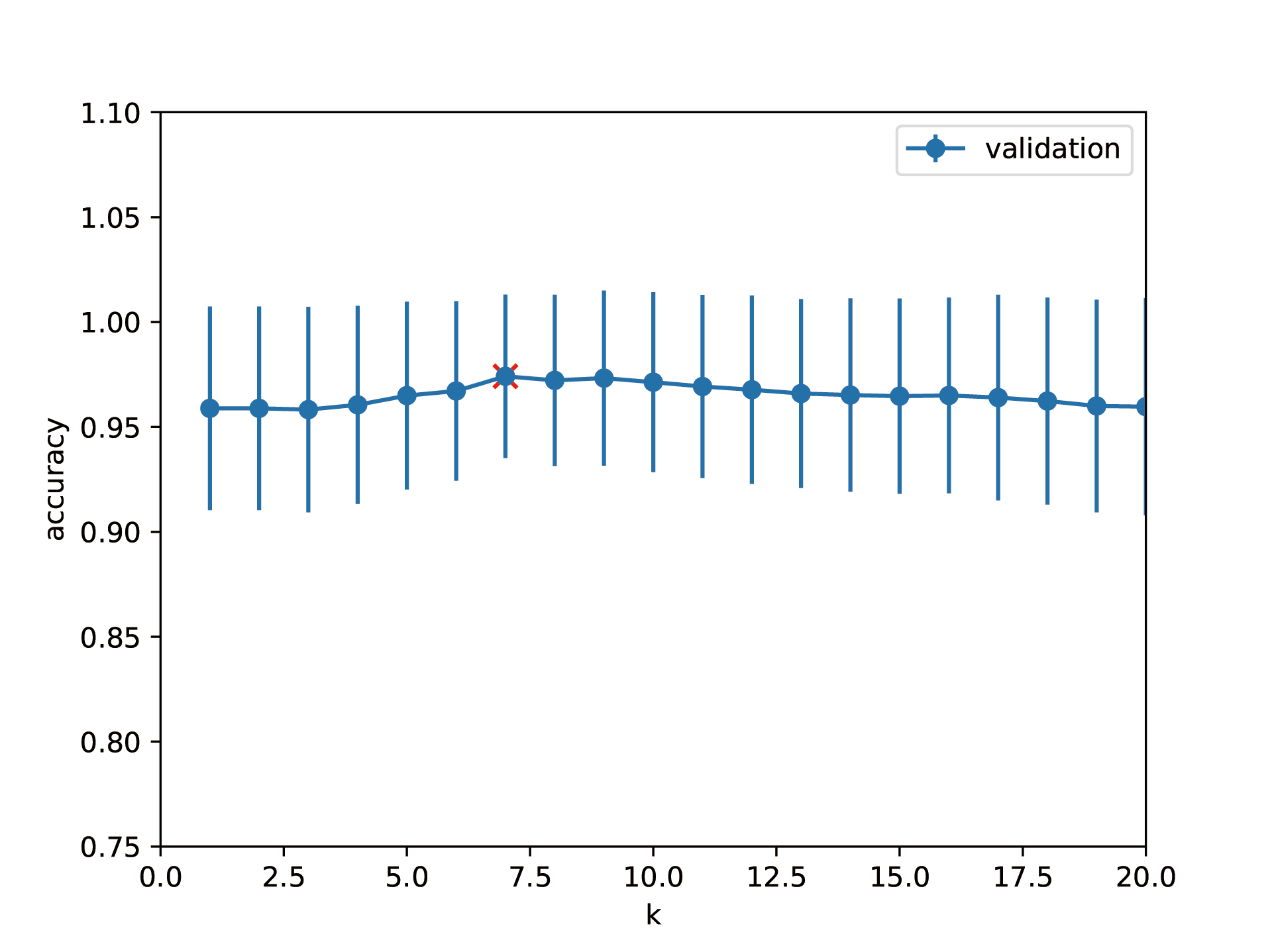
图（2）p=2

图（1）p=1



图（4）p=4

图（3）p=3



图（6）p=6

图（5）p=5

表1 实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 均值(标准差) | | 验证集 | 测试集 | 当前最优K值 |
| P值 | 1 | 96.49%(0.0462) | 96.65%(0.0433) | 17 |
| 2 | 96.73%(0.0453) | 96.75%(0.0425) | 13 |
| 3 | 97.28%(0.0422) | 97.73%(0.0365) | 9 |
| 4 | 97.39%(0.0412) | 97.81%(0.0362) | 9 |
| 5 | 97.41%(0.0390) | 97.67%(0.0366) | 7 |
| 6 | 97.45%(0.0384) | 97.67%(0.0365) | 7 |

图（1）至图（6）展示了在不同p值与K值情况下，分类模型在验证集上的统计性结果。针对每一组p值与K值设定下的分类模型，我们进行了1000次重复实验，保证数据集的划分对测试结果影响降低到最小。每幅图中的点代表1000次实验所得准确率的均值，竖线代表这组准确率的标准差。而红色叉号标注的点即为在验证集上取得最高正确率的K值。同时，我们可以发现随着p值的增大，最优K值随之减小。

根据表1实验结果可以看出：验证集取得最高准确率时对应的P值为：6、K值为：7。由此我们筛选出了模型的最优超参数。然后，我们将验证集与训练集合并为新的训练集来重新训练分类器，并在测试集进行模型的评估。同样，我们重复进行了1000次评估实验，得到了在测试集上的准确率均值与方差。在选取最优超参数时，模型取得的最优预测正确率为：97.67% (0.0365)。