KNN读书笔记\_周师严

1. 算法概述（算法名称及原理）

K最近邻(kNN，k-NearestNeighbor)分类算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

1. 算法设计（流程图及主要分段代码，附详细代码注释）

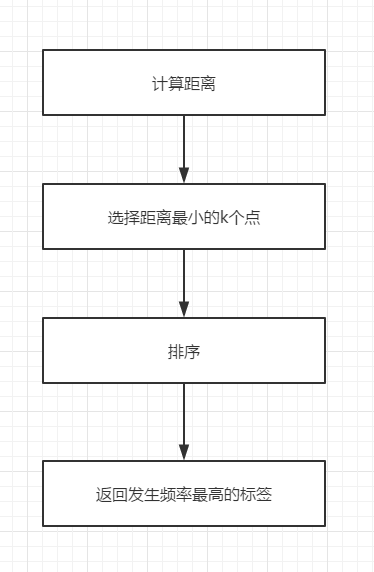


图1

**KNN算法：**

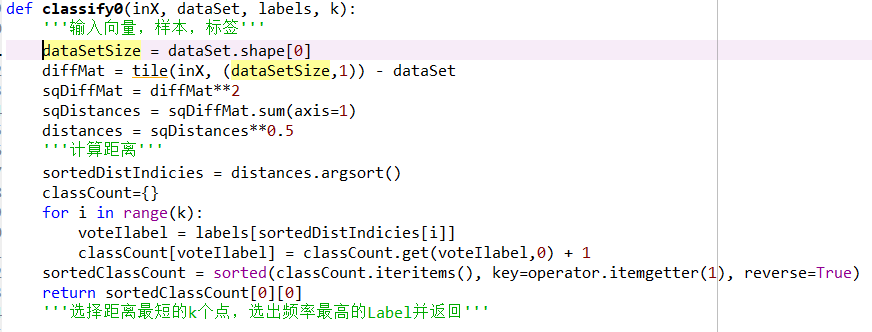


图2

**分类器：**

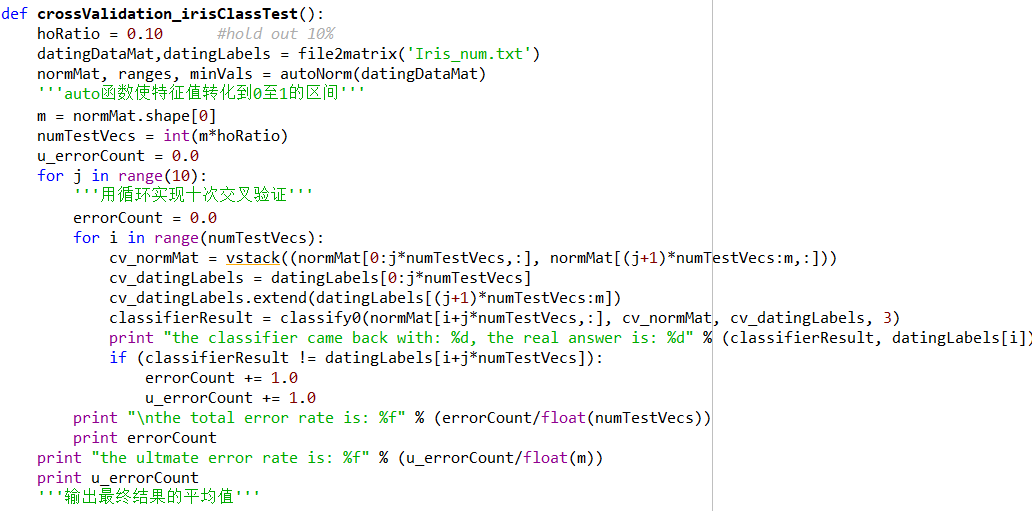


图3

1. 选用数据（数据集描述，包括来源，行数，列数，格式等）

**来源：**UCI——**Iris Data Set**

数据集包含3个类别，每个类别50个实例，其中每个类别指的是一种iris（鸢尾属）植物。

**Predicted attribute:**

class of iris plant

**Attribute Information:**

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

5. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

**为方便处理已经将class 转换为数字1，2，3**

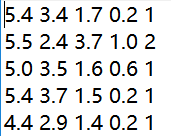


图4

1. 评价方法（说明训练集和测试集分配方法及评价指标）

使用十次交叉验证法，选错误率最小。

1. 实验结果截图

k取3时的结果：

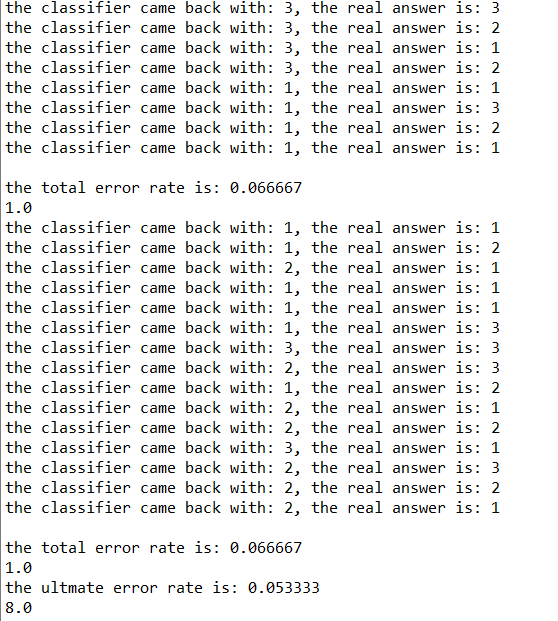


图5

1. 实验结果分析及比较

通过改变k的数值，找到最好的结果

图6

由上图可看出，错误率随着k值增加而减少，当超过6时又开始增加。

**当k取6时，错误率最小为3.3%，即准确率96.6%，这是一个相当不错的结果。**

1. 遇到的问题及解决方法，实践心得

**问题：**编写代码用for 循环实现交叉验证法时，出现问题。

普通的列表list和numpy中的数组array是不一样的，np.array 需要使用 c = vstack(a,b) 合并list，而不能使用extend

**心得：**Knn中k的取值不能太大也不能太小

CART读书笔记\_周师严

1. 算法概述（算法名称及原理）

CART是一种著名的决策树学习算法，分类和回归任务都可用，使用属性的基尼系数（Gini index）来选择划分属性。

CART 假设决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支

是取值为“是”的分支，右分支是取值为“否”的分支。这样的决策树等价于递

归地二分每个特征，将输入空间即特征空间划分为有限个单元，并在这些单元上

确定预测的概率分布，也就是在输入给定的条件下输出的条件概率分布。

1. 算法设计（流程图及主要分段代码，附详细代码注释）

**CART 分类树⽣成算法：**

1. 创建根节点N
2. 为N分配类别
3. if T都属于同一类别or T中只剩下 一个样本则返回N为叶节点，否则为其分配属性
4. for each T\_attributelist中属性执行该属性上的一个划分，计算此划分的GINI系数
5. N的测试属性test\_attribute=T\_attributelist中最小GINI系数的属性
6. 划分T得到T1 ，T2子集
7. 对于T1重复（1）-（6）
8. 对于T2重复（1）-（6）

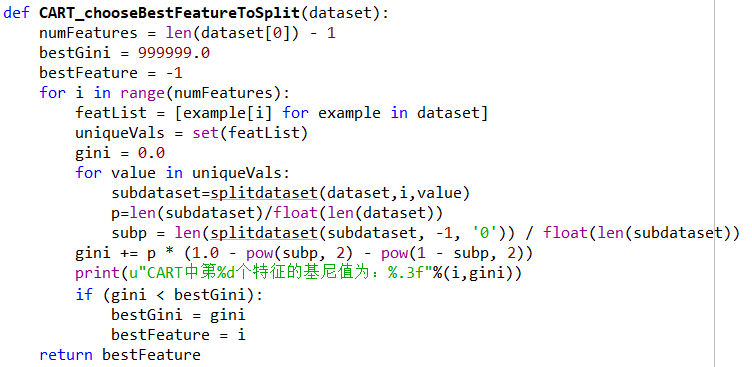


图1

1. 选用数据（数据集描述，包括来源，行数，列数，格式等）

**来源：**UCI——**Iris Data Set**

数据集包含3个类别，每个类别50个实例，其中每个类别指的是一种iris（鸢尾属）植物。

**Predicted attribute:**

class of iris plant

**Attribute Information:**

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

5. class:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

**为方便处理已经将class 转换为数字1，2，3**

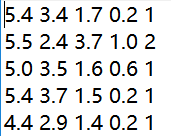


图2

1. 评价方法（说明训练集和测试集分配方法及评价指标）

使用十次交叉验证法，结果选预测均值。

1. 实验结果截图

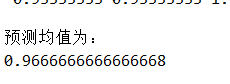


图3

1. 实验结果分析及比较

最终为准确率96.6%，这是一个相当不错的结果。

1. 遇到的问题及解决方法，实践心得

**问题：**CART分为模型树和回归树，一开始混淆了概念导致不少误解。