教程 2 : K-NN

提交日期：2020.7.18

提交人：詹紫琦

目录

[1. 题目 3](#_Toc30573)

[1.1 K近邻算法（KNN） 3](#_Toc15809)

[1.2 运行环境 3](#_Toc3805)

[2. 算法阐述或实验步骤说明 3](#_Toc1205)

[2.1变量说明 3](#_Toc10073)

[2.2 KNN回归算法阐述及步骤说明 4](#_Toc13274)

[2.3 KNN分类算法阐述及步骤说明 5](#_Toc28223)

[3. 实验结果与截图 5](#_Toc10185)

[3.1 KNN回归实验结果 5](#_Toc26540)

[4. 总结 12](#_Toc11884)

[4.1 总结 12](#_Toc8992)

[5. 参考文献 12](#_Toc5528)

# 题目

## 1.1 K近邻算法（KNN）

KNN[1]是一种分类算法，其核心思想就是对于距离的把控，通过散点分布之间不同点的距离来判断该点的类别或者是数值。借由KNN这一简单机器学习算法的特性，可以将KNN应用于分类或者是回归问题中。本文中就介绍了KNN算法应用于分类以及回归问题中的不同情形中的不同定义。KNN是通过测量不同特征值之间的距离进行分类，选择距离最近的K个点，来综合评判该预测点的值或者是类别的一种简单算法，基本思路是，如果一个样本在特征空间中的K个最相似（即特征空间中最邻近）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。

## 1.2 运行环境

系统：Ubuntu20.04，python3.7，Anaconda集成工具Jupyter编写。

# 算法阐述或实验步骤说明

## 2.1变量说明

：K近邻算法的关键参数，表示算法中需要选择的与样本特征空间内距离最近的其他样本点的个数。

：该算法中所应用的数据集，总共有m项。

O\_distance：该算法中所应用的距离测算方法，使用的是欧氏距离。

X\_mean ：取K个最近邻点的均值

K\_MSE：不同均值下的均方误差MSE

Predict\_x：预测的样本点的值

Predict\_classes：预测样本点的分类

## 2.2 KNN回归算法阐述及步骤说明

在本次实验中分别进行了KNN回归和分类两种算法的应用，首先进行数据集选取，KNN回归数据集的选取的初步分析如图2-1：其中包含两种特征值，这里分别用feature\_one和feature\_two替代

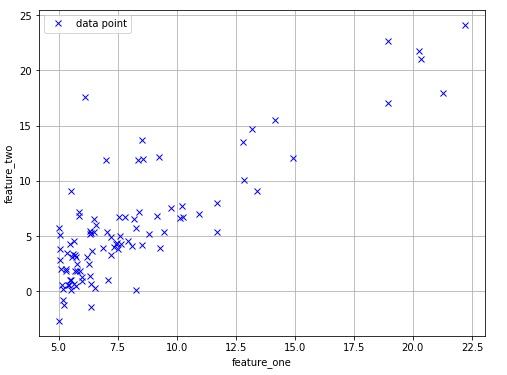


图2-1 KNN回归数据集点的分布

进行了数据集的读取后，定义欧式距离函数，这里我们测算样本间的距离采用的是欧式距离，如公式2-1

 （2-1）

定义完欧式距离函数后，进行KNN回归功能函数编写，首先依次遍历样本中每一个点，针对该点而言，计算其他点与该点的欧式距离，找出最小距离的K个点，这里我们回归预测采用的是均值预测，如公式2-2

 （2-2）

即对最近邻K个样本点，取均值作为预测值。直到遍历完所有样本点。然后进行误差分析，这里采用MSE均方误差来评判优劣，如公式2-3

 （2-3）

最后编写MSE函数，获取回归模型的误差值，进行误差分析。

## 2.3 KNN分类算法阐述及步骤说明

与KNN回归算法不同的是，KNN分类算法采用的是UCI数据集中的鸢尾花数据集，首先对数据集进行预处理，将花的三种种类字符串类型转换为0，1，2三个数字替代三种种类。初步分析数据，发现数据特征值总共有四种。然后编写欧几里得距离函数，同公式2-1，然后进行KNN分类功能函数的编写，与回归算法相似，依次遍历样本中的每一个点，计算出该点与其他点的欧氏距离，选取最近邻的K个样本点，找出这K的样本点的类别号的众数，即最多的种类是哪一种，在这里编写获取众数功能函数，最终得到预测分类。与回归算法的评判标准不同，分类算法使用的是分类正确率计算，分别计算设定不同K值下的分类正确率。进行误差分析。

# 实验结果与截图

## 3.1 KNN回归实验结果

初始化K值分别为3，6，9进行初步回归预测，设置K值为3，得到的预测结果与真实数据的对比图如图3-1：我们可以初步发现，预测的点的值相较于样本点的范围进行了一个收缩。然后设置K值为6，得到的预测结果与真实数据的对比图如图3-2。接着设置K值为9，得到的预测结果与真实数据的对比图如图3-3。通过K值的不断增大，发现每一个样本点其中包含K个样本点的均值的方差开始变大，所以误差会变得更大，这里我们取K为1，2，3......，50总共50个K值，进行分析MSE均方误差的值，得到的不同K值下使用KNN回归算法对样本进行预测的不同MSE误差值，同时包含对每一个特征值的预测值与真实值的MSE，以及两种特征MSE之和的变化，如图3-4。从误差图中可以看出，针对第一类特征值，预测误差的大小并不是与K值单调相关，而对于总误差而言，K值越大，总误差越大。

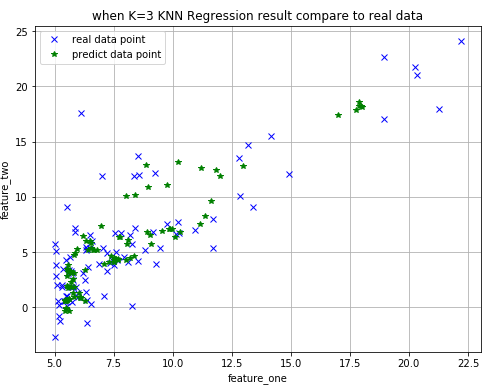


图3-1 当K值为3时回归预测结果与真实值的对比

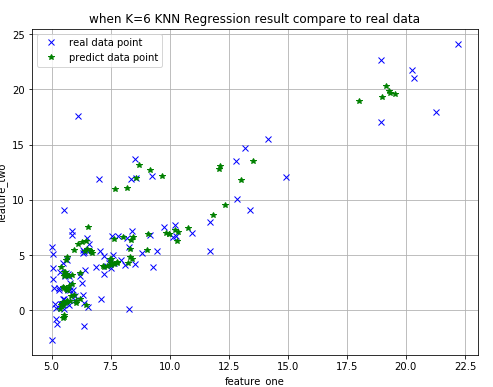


图3-2 当K值为6时的回归预测结果与真实值的对比

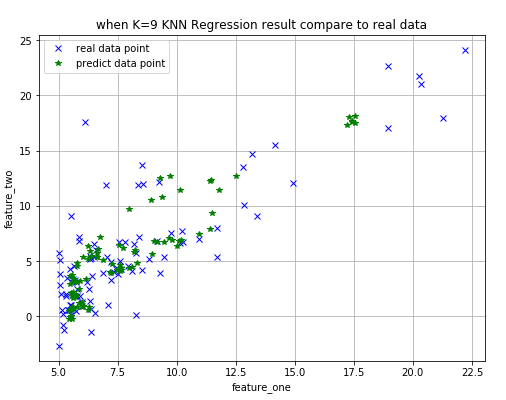
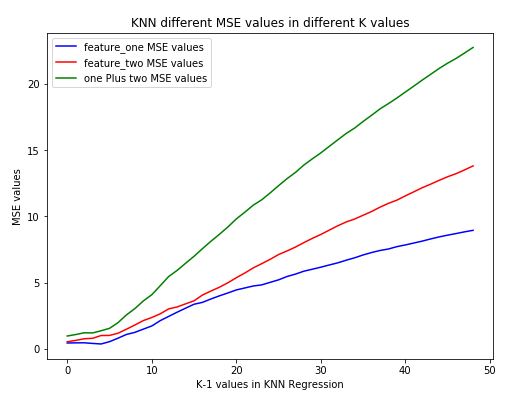


图3-3 当K值为9时的回归预测结果与真实值的对比

图3-4 不同K值下的MSE大小

## 3.2 KNN分类实验结果

与回归算法评判标准不同，这里采用的是分类正确率来进行分类算法优劣评判，首先观察鸢尾花数据的类别分布，其中用0，1，2分别表示不同的鸢尾花类别，初始数据如图3-5：

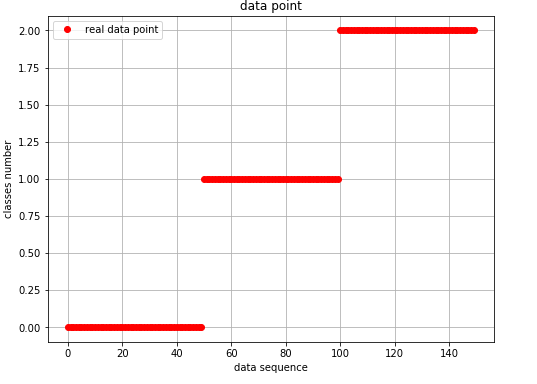


图3-5 鸢尾花数据初始数据类别

分别设置K=3，6，9初步观察分类结果与真实数据对比，取K为3时，分类结果如图3-6，取K为6时，分类结果如图3-7，取K为9时，分类结果如图3-8。绿色点与红色点重合程度越高代表分类效果越好。

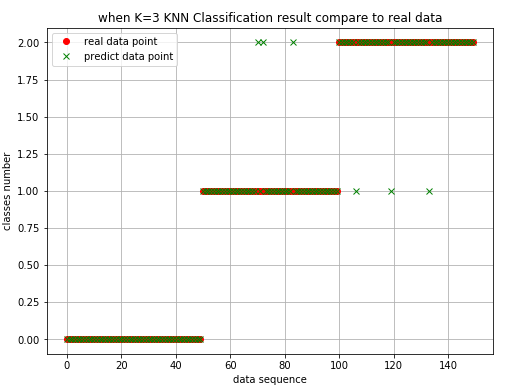


图3-6 K取3时的分类结果

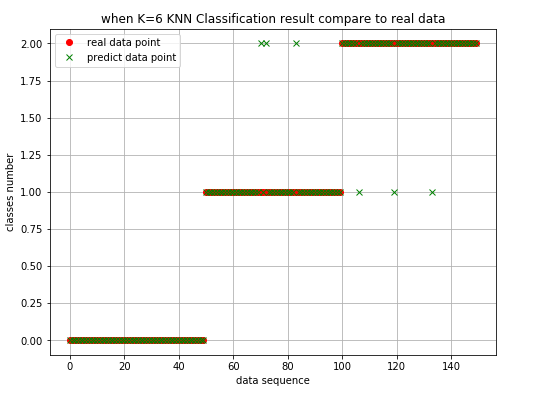


图3-7 K取6时的分类结果

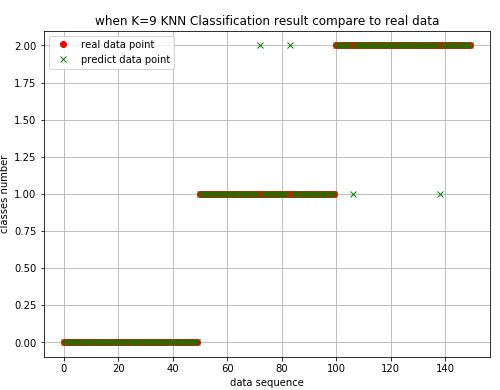


图3-8 K取9时的分类结果

这里同样取K值为1到50的序列，得到的不同K值下的分类准确率如图3-9所示，得到的最高分类准确率为98.00%，此时的K值为15，19，21。

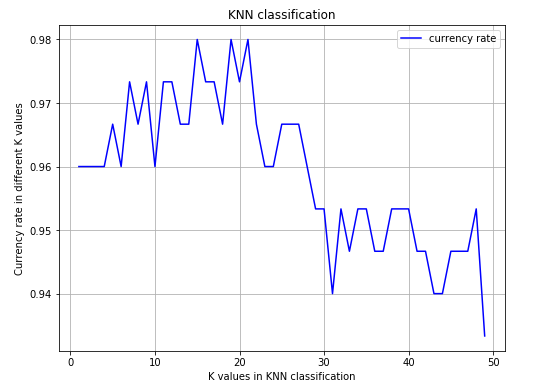


图3-9 不同K值下KNN分类算法准确率

# 总结

## 4.1 总结

通过了解到KNN这一包含分类和回归的算法，了解到了特征空间中各特征点的距离的定义，在本次实验中采用的都是欧氏距离定义法，而KNN这一算法中的距离定义还包含着多种类型的距离，例如汉明距离等。在KNN回归中，我们发现K值不宜取得过大，当K值过大时，待预测样本点就需要综合更多相对较远距离的点，使得预测样本点的值误差变大，而KNN分类中，分类准确率不随K的值而线性变化，最后的分类准确率能达到98%，说明KNN这一简单的算法思想却能达到好的分类效果，所以KNN一般都用于分类中。相比于其他复杂的分类算法，KNN实现简单，运行速度快，也不存在消耗大量资源的情况，而且还能得到比较好的分类效果。

# 参考文献

[1] 周志华等.机器学习（西瓜书）