廣東工業大學

QG 中期考核详细报告书

题　　目　QG-数挖组中期考核报告书

学　　院　计算机学院

专　　业　信息安全

年级班别　19 级（1）班

学　　号　3119005431

学生姓名　欧智昕

2020 年 04 月　24 日

1. kNN 近邻算法

(1)数据集的处理：

将 Iris 数据集读取后，以 2：1 的形式划分训练集和测试集，这样就完成了

(2)算法步骤

(3)思想

kNN 不同于大部分算法，它的训练是“懒惰训练”，就是即训即算，不需要训练时间。

kNN 的核心思想就是——在距离最近的 K 个中选出现频率最高者作为预测

用合适的距离计算公式计算、选取合适的 k、训练集分布的均匀程度、距离的加权方式这些都是影响 kNN 算法优劣的重要因素。

(4)优化之处

对于普通的 kNN 算法，在给前 k 个距离最小所对应的标签分配权值时，是以平均分配的方式，即出现一次就分配 1。但这样的缺点就是，当训练集数据分布不均匀时，有些类别的样本数多，有些少，这样可能会由于数量上的不平衡导致错误标签被筛选到 k 个距离中。当分配权值时，无论该点与测试点距离多远分到的都是 1，最后就很可能就因为数量上的差距导致预测错误，这样吃“大锅饭”就无法正确分辨到底哪些能干、哪些在划水。

解决方法：

对权值的分配进行改进，无论出现次数多少，只要距离远就分配小权值、距离近就分配大权值。

可以使用取倒数、高斯函数等一系列对距离进行加权优化的函数对距离进行

加权优化。

这里我用了高斯函数，提前预设好它的三个参数，使得加权后的最大值不超过 1，同时所有加权得出的数均大于 0，使其对称轴与 y 轴重合，这样离得越近，所得权值越大，越远就越小

优化结果：

经加权优化后，一定程度上减小了由于训练数据分布不均衡导致的预测失误问题，使得“能干的”被赏识，“划水的”被淘汰

(5)算法实现结果评估

```
该模型在50个测试集上的错误率为32.84%
经加权优化后的模型在50个测试集上的错误率为32.84%
sklearn的KNN模型在50个测试集上的错误率为32.84%
```

按照 2：1 的划分，k 设定为 3，最终优化和未优化的模型在测试集上的错误率是 32.84%，与 sklearn 的模型错误率一致。

导致 sklearn 模型与自己搭建的模型错误率一致的原因：

距离计算公式相同，k 选取相同，

优化与未优化的模型错误率一致的原因分析：

1. 训练数据分布比较均匀，经距离加权优化后的模型提升效果不明显
2. 训练数据比较少，无法体现出优化模型的优势

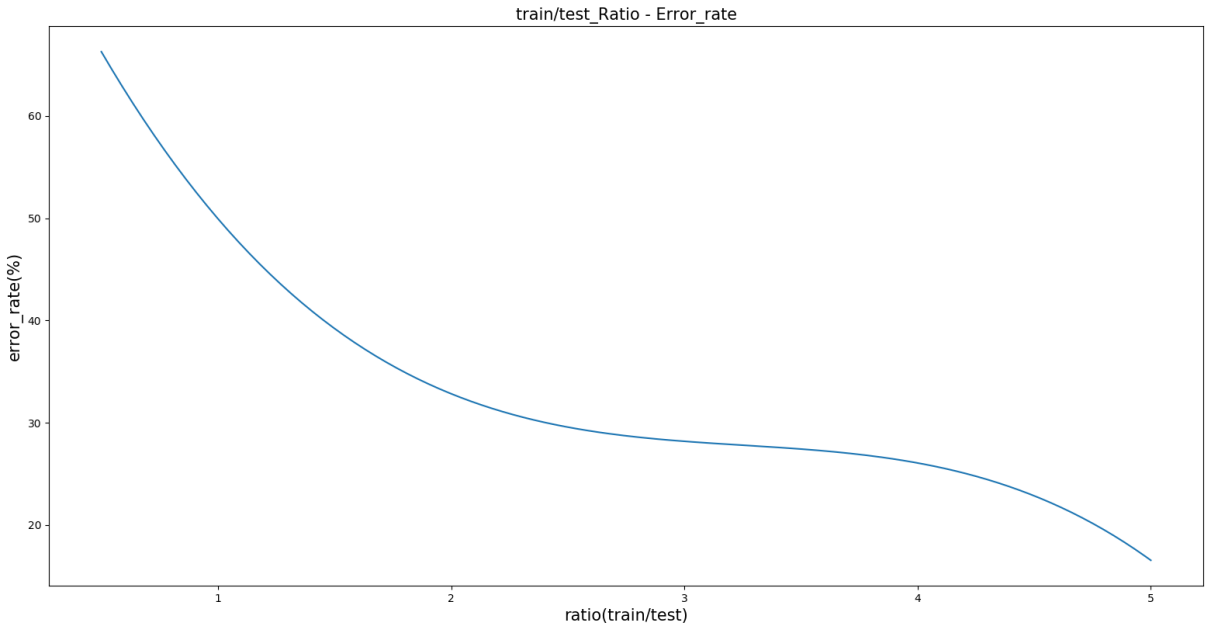
同时，在多次改动 k 值后测得的错误率变化微小，说明该模型对 k 不敏感

当改变训练集与测试集的比例时，发现错误率变化很大

这是训练集：测试集 = 1：2 时的结果

```
该模型在100个测试集上的错误率为66.28%
经加权优化后的模型在100个测试集上的错误率为66.28%
sklearn的KNN模型在100个测试集上的错误率为66.28%
```

下图是训练集与测试集比例与错误率的关系曲线



可以看出该模型的预测效果对训练集与测试集的分割比例还是比较敏感的。因为当训练集较少时，模型可供参考的点很少，这样训练集内点分布可能不是很能体现出类别的不同，导致模型对类别的认知度低，有点像没搞懂知识点就去做作业，最后做出来一塌糊涂。同时测试集较少时，错误率比较低，这说明模型对类别的认知度比较高，把握较精准；当然也有可能是测试集数量较少，无法体现模型真正的预测效果

(6)不足

没有进行归一化的处理，使结果不受数据数值大小的影响。

由于 kNN 算法是计算点与点之间的距离大小，因此当数据的某些特征值在数值上很大或很小时，它会对距离造成影响，可能导致计算出来的距离过大或过小，在预测时造成干扰或误判，这样会大大降预测低准确率。

如果对数据进行归一化，就能使数据的分布规范，不受数值大小的影响，避免“以偏概全”——仅因为数值大而被筛选掉或仅因为数值小而被选中

2. 多元线性回归

(1) 数据集的处理

由于 housing 数据集用不等距的空格分割，因此在用 read_csv 时加上参数 sep='\s+'，同时将数据最后一列的房价抽出来作为真实值，其余作为回归的样本

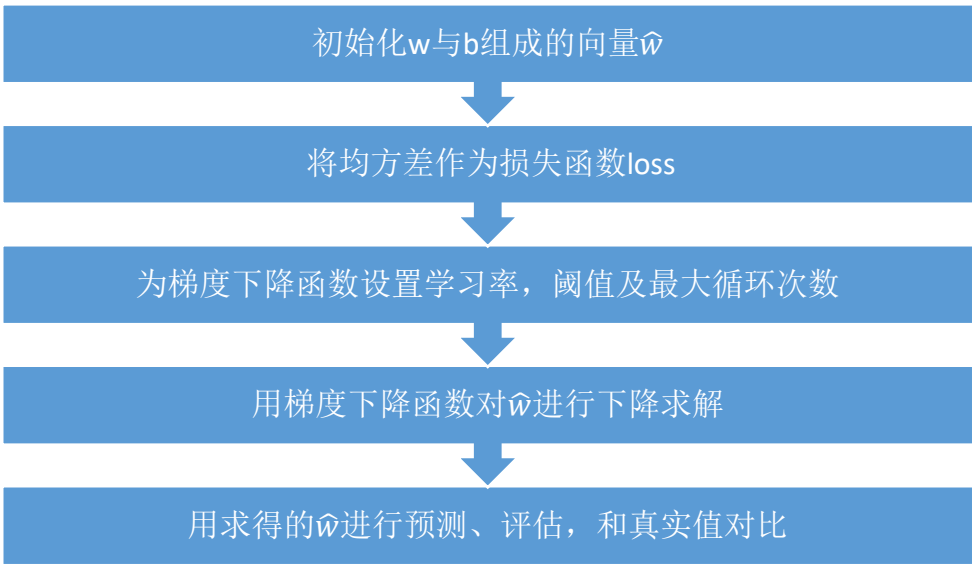
(2) 算法步骤

线性回归主要问题就是求最佳的 w 与 b，使得各点离回归直线的距离最小

1. 公式法（最小二乘法）

   直接利用多元线性回归方程的求参公式可得最优的 w 与 b

2. 梯度下降法



(3) 思想

线性回归，其实在中学里早已有他的身影。

线性回归的核心是，用一条最优直线来拟合样本中的所有离散点，使得该直线与各点的平均距离总和最小。就是尝试在所有离散点之间找到它们之间的线性函数关系。

(4) 算法实现结果评估

下图是梯度下降法模型与 sklearn 模型的预测值和真实值之间的比较

绿色曲线代表真实值，蓝色曲线是梯度下降模型，橙色曲线是 sklearn 库中的模型，横坐标为样本数，纵坐标为房价

下面是两模型的均方差比较

梯度下降模型均方差为24.25 ， sklearn中多元线性回归模型均方差为21.89

可以看出梯度下降模型的效果还是不错的。
从这 250 个样本预测来看，整体的重合部分比较多，但在房价真实值接近 50 的时候预测的效果不是很好。

(5) 不足

数据没有进行归一化，各数据之间数值大小的差距比较大，在梯度下降过程中计算量大，拖慢了梯度下降的速度，若数据量比较多，就会导致梯度下降的效率低，难以找到最低点。

(6) 优化之处

不断调整梯度下降函数中学习率、阈值及最大循环次数，使得梯度下降模型能够在最短时间内找到使损失函数达到最小的$\hat{w}$，使均方差尽可能地减小。还有$\hat{w}$的初始值的选定也可以优化，可以通过多次调整$\hat{w}$初始值再进行梯度下降，再进行评估，以找到最佳的$\hat{w}$。