泰坦尼克数据挖掘

KNN算法以及实现

Titanic.py

基本类

1. Point类

功能:

用于存储属性与类标值,并且提供了常用的类标计算方法.后续任何的数据挖掘算法,计算都要使用该类进行表示记录值.

成员

X: 为一个数组,用于存储属性值;

Y: 为一个对象,用于存储属性值对应的类标;

方法:

* \_\_init\_\_( x, y)

Init方法,用于创建一个对象

x: 输入属性值,要求为tuple类型

y: 属性所对应的类标

* get\_distance(x1, x2, method="E")

类方法,用于计算两个点的类型.

X1: Point类型的一个点

X2: Point类型的一个点

Method: 使用的方法,字符串类型,可缺省,默认使用欧几里得算子

可选参数: 'E': Euclid,

'M': Manhattan,

1. Titanic类

功能

用于读取Titanic.dat数据集,能够将数据集转换为Point类型

成员

Data: 为文件中的有用数据,为上述Point类型.

方法

* \_\_init\_\_()

Init方法,能够初始化数据集

参数:

data\_path: data文件的存放路径,支持相对路径

separtor: data文件中使用的分隔符,可缺省,默认为','

commenting: data文件中的代码注释符,可缺省,默认为'@'

goal\_colum: data文件中,待遇测属性所在的行数,遵循python风格,默认为最后一行"-1"

data\_range: data文件中,需要映射的数据范围

1. KNN类

功能

对Point类型的数据进行KNN,能够预测类标与计算测试集在训练集上的成功率

成员

Data: 为原数据集合

K: 为我们采用K值

Method: 为计算KNN是采用的 K值

Train: 划分的训练集

Test: 划分的测试集

方法

* \_\_init\_\_()

功能:

Init方法,能够初始化KNN算法所需的数据

参数:

Data: 数据集

k, K值

precentage\_of\_training\_data: 训练集比例,为一个0-1的浮点数,默认为0.7

method='E': 采用的距离算法,默认为欧几里得算子

* \_split

划分数据集,采用随机划分

参数:

P: 训练集比例

* check(x):

功能:

根据输入的x预测类标y,返回类标值

参数:

X: 为一个属性值构成的数组

返回值:

返回属性x对应的类标值

* \_get\_all\_distance(x):

功能:

获取某个点到其余点的距离

参数:

x: 数组类型,存储属性值,要求与其余点拥有一样的长度

返回值:

Dis: 为训练集中每一个样本对这个点的距离,为字典类型,通过训练集point对象索引索引

* \_get\_kmin\_point(x, dis):

功能:

获取距离目标点x最小的k个point

参数:

X: 当前的属性值数组

Dis: \_get\_all\_distance的返回值

返回值:

* check\_test():

功能:

获取测试集准确性

参数:

返回值:

一个0-1的浮点数

程序流程:

