

王佩的CSDN博客

大数据开发、机器学习、深度学习、神经网络、推荐系统、自然语言处理、图数据库



∠ 写博客





wangpei1949

关注 发私信

访问: 39812次 积分: 751 等级: 8L00 3 排名: 干里之外

原创: 37篇 转载: 0篇 译文: 0篇 评论: 4条

文章搜索

文章分类

机器学习@深度学习 (20)

设计模式 (0)

数据结构与算法 (2)

spark开发 (5)

hbase开发 (0)

hive开发 (0)

python开发 (3)

scala开发 (0)

java开发 (5)

C++开发 (0)

R (0)

Idea (2)

titan图数据库 (0)

neo4j图数据库 (0)

Linux与Shell编程 (1)

阅读排行

Idea安装Python插件并配置P... (3671)Spark MLlib特征处理: One... (3504)Spark MLlib特征处理:均值... (3404)Spark MLlib特征处理: PCA ... (2395)机器学习之sklearn特征工程

Machine Learning --5种距离度量方法

标签:相似度计算 机器学习相似度计算 机器学习距离 ML距离公式 ML相似度计算

2016-10-25 20:32

Q

1953人阅读

评论(0)

收藏

举报

登录 注册

Ⅲ 分类:

机器学习@深度学习(19) ▼

■ 版权声明:本文为博主原创文章,未经博主允许不得转载。

目录(?) [+]

1 前言

在数据挖掘中,我们经常需要计算样本之间的相似度(Similarity),我们通常的 做法是计算样本之间的距离,本文对

距离计算方法做以下总结。

2 距离计算方法

A 欧式距离EuclideanDistance

欧式距离:两点之间的直线距离。

(1)二维平面上两点 $a(x_1,y_1)$, $b(x_2,y_2)$ 之间的欧式距离公式:

$$d_{ab} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

(2) n维空间上两点 $a(x_1,x_2,...,x_n)$, $b(y_1,y_2,...,y_n)$ 的欧式距离公式:

$$d_{ab} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

B 曼哈顿距离(Manhattan Distance)

曼哈顿距离也叫"曼哈顿街区距离"。想象你在曼哈顿街道上,从一个十字路 口开车到另一个十字路口,驾驶距离就

 jieba 结巴分词 常用说明
 (2145)

 Machine Learning --5种距离...
 (1953)

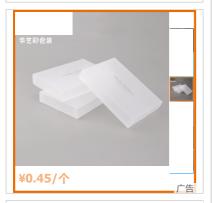
 Spark MLlib特征处理: SVD ...
 (1773)

 padas 生成excel 增加sheet表
 (1733)

 jieba分词支持关键词带空格和...
 (1336)

评论排行

Spark MLlib特征处理: One... (2)Spark Sql 二次分组排序取To... (1)机器学习之sklearn特征工程 (1)spark 提交jar包到集群运行报... (0)RDD笛卡尔操作Cartesian (0)RDD之aggregate操作 Machine Learning --5种距离... Spark MLlib特征处理: Binar... (0)Spark MLlib特征处理: Strin... (0)Shell命令行处理JSON (0)



最新评论

Spark MLlib特征处理: OneHotEncoder... yaoqsm: 你好,我看不懂最后的结果, (3,【0】,【1.0】) 是什么意思啊,我知道one-hot编码的意思,但...

Spark Sql 二次分组排序取TopK licoderli : 代码有点问题,难道CSDN的bu g吗,大概有5处地方被插入了>

机器学习之sklearn特征工程

qq_27892841 : StandardScaler.fit() 与 St andardScaler.fit transfor...



是这个"曼哈顿距离"。

(1)二维平面上两点 $a(x_1,y_1)$, $b(x_2,y_2)$ 之间的曼哈顿距离公式:

$$d_{ab} = |x_1 - x_2| + |y_1 - y_2|$$

(2) n维空间上两点 $a(x_1,x_2,...,x_n)$, $b(y_1,y_2,...,y_n)$ 的曼哈顿距离公式:

$$d_{ab} = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

C夹角余弦

机器学习中可以把两点看成是空间中的两个向量,通过衡量两向量之间的相似性来衡量样本之间的相似性。

(1)二维平面上两向量 $a(x_1,y_1)$, $b(x_2,y_2)$ 之间的夹角余弦公式:

$$\cos\Theta = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{x_1^2 + y_1^2} * \sqrt{x_2^2 + y_2^2}}$$

也可直接通过向量运算:

$$\cos\Theta = \frac{a * b}{|a| * |b|}$$

(2) n维空间上两点 $a(x_1,x_2,...,x_n)$, $b(y_1,y_2,...,y_n)$ 的夹角余弦公式:

$$\cos\Theta = \frac{x_1 * y_1 + x_2 * y_2 + \dots + x_n * y_n}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} * \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}}$$

D 切比雪夫距离 (Chebyshevdistance)

切比雪夫距离:各对应坐标数值差的最大值。国王从格子(x1,y1)走到格子(x2,y2)最少需要多少步?你会发现最少步

数总是max(|x2-x1|,|y2-y1|)步。

(1)二维平面上两点 $a(x_1,y_1)$, $b(x_2,y_2)$ 之间的切比雪夫距离公式:

$$d_{ab} = max(|x_1 - x_2|, |y_1 - y_2|)$$

(2) n维空间上两点a(x₁,x₂......x_n), b(y₁,y₂......y_n)的切比雪夫距离公式:

$$d_{ab} = max(|x_1 - y_1|, |x_2 - y_2|, \dots |x_n - y_n|)$$

E 汉明距离

两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数。

1011101与 1001001 之间的汉明距离是2

2143896与 2233796 之间的汉明距离是3

irie与 rise之间的汉明距离是 3

顶 踩 0

- 上一篇 机器学习之sklearn特征工程
- 下一篇 Spark MLlib特征处理: Binarizer 二值化---原理及实战

相关文章推荐

- 相似性度量方法(欧式距离等各种距离)
- MySQL在微信支付下的高可用运营--莫晓东
- [数据挖掘]数学基础---距离度量方式(马氏距离,...
- 容器技术在58同城的实践--姚远
- 【machine learning】朴素贝叶斯分类方法
- SDCC 2017之容器技术实战线上峰会
- Machine Learning3——LDA算法 (一种经典的...
- SDCC 2017之数据库技术实战线上峰会

- Ensemble method of machine learning 机器学...
- 腾讯云容器服务架构实现介绍--董晓杰
- Machine Learning in Action 学习笔记-(4)基...
- 微博热点事件背后的数据库运维心得--张冬洪
- 菜鸟眼中的machine learning和我的一些机器学...
- 周志华《Machine Learning》学习笔记(2)--性...
- 周志华《Machine Learning》学习笔记(12)--...
- 周志华《Machine Learning》学习笔记 (2) --性...



查看评论

暂无评论

您还没有登录,请[登录]或[注册]

*以上用户言论只代表其个人观点,不代表CSDN网站的观点或立场



公司简介 | 招贤纳士 | 广告服务 | 联系方式 | 版权声明 | 法律顾问 | 问题报告 | 合作伙伴 | 论坛反馈

网站客服 杂志客服 微博客服 webmaster@csdn.net 400-660-0108 | 北京创新乐知信息技术有限公司 版权所有 | 江苏知之为计算机有限公司 |

江苏乐知网络技术有限公司

京 ICP 证 09002463 号 | Copyright © 1999-2017, CSDN.NET, All Rights Reserved



