机器学习的5种距离度量方法

凡人机器学习 2018-05-08

点击蓝字关注这个神奇的公众号~

在机器学习领域中有非常多的问题需要求距离，常见的是向量距离的计算。比如判断A、B、C三种商品之间的相似性，可以先按照商品特征构建A、B、C的各自的向量，然后求向量间的距离，距离近就表示彼此相似度高。今天讲下常见的几种距离计算方法。

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：

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

B 曼哈顿距离(ManhattanDistance)

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：

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

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：

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

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(x_1,x_2.....x_n)，b(y_1,y_2.....y_n)的切比雪夫距离公式：

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

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