

机器学习中常用的距离度量方式

原创 libo 数据科学与AI 1月11日

机器学习中常用的距离度量方式

一、欧式距离

欧式距离是最常用的计算方式，源自于欧式空间中两点间的距离

- 二维平面两点A(x1,y1)与B(x2,y2)间的距离：

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

数据科学与AI

- 三维空间两点A(x1,y1,z1)与B(x2,y2,z2)间的距离：

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

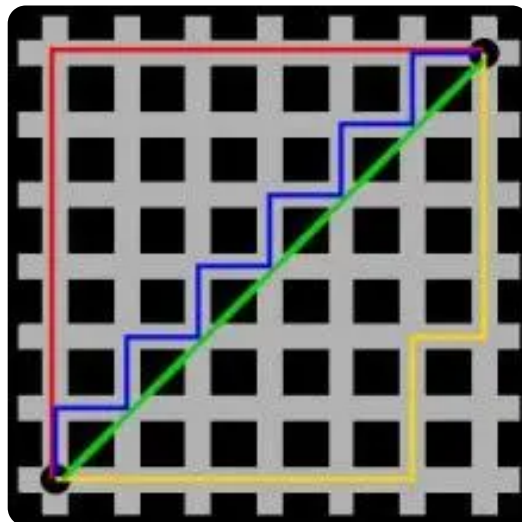
数据科学与AI

- 两个n维向量a(x11,x12,...,x1n)与 b(x21,x22,...,x2n)间的欧氏距离：

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

数据科学与AI


二、曼哈顿距离



如图所示，黑色的是高楼林立，灰色的是街道，从左上角到达右上角的位置，不能如绿色直线到达，只能如红蓝黄三种类似的方式到达。其实三种类似的距离是相等的。

- 二维平面两点A(x1,y1)与B(x2,y2)间的距离：

$$d_{12} = |(x_1 - x_2)| + |y_1 - y_2|$$


 数据科学与AI

- 两个n维向量a(x11,x12,...,x1n)与 b(x21,x22,...,x2n)间的曼哈顿距离:

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

 数据科学与AI

三、切比雪夫距离

| | a | b | c | d | e | f | g | h | |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 5 | 4 | 3 | 2 | 2 | 2 | 2 | 2 | 8 |
| 7 | 5 | 4 | 3 | 2 | 1 | 1 | 1 | 2 | 7 |
| 6 | 5 | 4 | 3 | 2 | 1 |  | 1 | 2 | 6 |
| 5 | 5 | 4 | 3 | 2 | 1 | 1 | 1 | 2 | 5 |
| 4 | 5 | 4 | 3 | 2 | 2 | 2 | 2 | 2 | 4 |
| 3 | 5 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 2 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 2 |
| 1 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 1 |
| | a | b | c | d | e | f | g | h | |

如图国际象棋的棋盘所示，由于国王可以正前后左右走，也可以斜前斜后走，因此距离国王位置f6的最近的8个位置的距离均为1

- 二维平面两点a(x1,y1)与b(x2,y2)间的切比雪夫距离:

$$d_{12} = \max(|x_1 - x_2|, |y_1 - y_2|)$$

 数据科学与AI

- 两个n维向量a(x11,x12,...,x1n)与 b(x21,x22,...,x2n)间的切比雪夫距离:

$$d_{12} = \lim_{k \rightarrow +\infty} \left(\sum_{i=1}^n |x_{1i} - x_{2i}|^k \right)^{\frac{1}{k}}$$

 数据科学与AI

四、闵可夫斯基距离

闵可夫斯基距离不是一种距离，而是一组距离

- 闵可夫斯基距离的定义

两个n维变量a(x11,x12,...,x1n)与 b(x21,x22,...,x2n)间的闵可夫斯基距离定义为:

$$d_{12} = \sqrt[p]{\sum_{k=1}^n |x_{1k} - x_{2k}|^p}$$

 数据科学与AI

其中，p是一个变参数。

当 $p = 1$ 时，就是曼哈顿距离

当 $p = 2$ 时，就是欧式距离

当 $p \rightarrow \infty$ 时，就是切比雪夫距

五、标准欧式距离

- 因为数据各维分量分布不一样，可以进行标准化

标准化过程：

$$X^* = \frac{X - m}{S} \quad m \text{ 为均值, } S \text{ 是标准差}$$

 数据科学与AI

标准化后的值 = (标准化前的值 - 分量的均值) / 分量的标准差

- 两个 n 维向量 $a(x_{11}, x_{12}, \dots, x_{1n})$ 与 $b(x_{21}, x_{22}, \dots, x_{2n})$ 间的标准化欧氏距离的公式：

$$d_{12} = \sqrt{\sum_{k=1}^n \left(\frac{x_{1k} - x_{2k}}{S_k} \right)^2}$$

 数据科学与AI

如果将方差的倒数看成是一个权重，这个公式可以看成一种加权欧式距离

六、马氏距离

- 有 M 个样本向量 $X_1 \sim X_m$ ，协方差矩阵记为 S ，均值记为向量 μ ，则其中**样本向量 X 到 μ** 的马氏距离表示为：

$$D(x) = \sqrt{(X - \mu)^T S^{-1} (X_i - \mu)}$$

 数据科学与AI

而其中向量 X_i 与 X_j 之间的马氏距离定义为：

$$D(x) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

 数据科学与AI

- 若协方差矩阵是单位矩阵（各个样本向量之间独立分布），公式就变成：即欧式距离

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)}$$

 数据科学与AI

七、夹角余弦

几何中夹角余弦可用来衡量两个向量方向的差异，机器学习中借用这一概念来衡量样本向量之间的差异。

- 在二维空间中， $A(x_1, y_1)$ 与向量 $B(x_2, y_2)$ 的夹角余弦公式：

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

 数据科学与AI

- 两个n维样本点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 的夹角余弦

$$\cos(\theta) = \frac{a \cdot b}{|a||b|} = \frac{\sum_{k=1}^n x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^n x_{1k}^2}\sqrt{\sum_{k=1}^n x_{2k}^2}}$$

 数据科学与AI

夹角余弦取值范围为 $[-1, 1]$ 。夹角余弦越大表示两个向量的夹角越小，夹角余弦越小表示两个向量的夹角越大，当两个向量的方向重合时夹角取值最大值1，方向相反时，夹角余弦取最小值-1。

根据欧氏距离和余弦相似度各自的计算方式和衡量特征，分别适用于不同的数据分析模型：欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异；而余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦相似度对绝对数值不敏感）。

八、汉明距离

- 汉明距离的定义

两个等长字符串 s_1 与 s_2 之间的汉明距离定义为将其中一个变为另外一个所需要作的最小替换次数。例如字符串“1111”与“1001”之间的汉明距离为2。

九、杰卡德相似系数

- 杰卡德相似系数(相似度)

两个集合A和B的交集元素在A，B的并集中所占的比例，称为两个集合的杰卡德相似系数，用符号 $J(A, B)$ 表示：

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

杰卡德系数是衡量两个集合的相似度一种指标

 数据科学与AI

- 杰卡德距离(区分度)

与杰卡德相似系数相反的概念是**杰卡德距离**(Jaccard distance)。杰卡德距离可用如下公式表示：


$$J_{\delta}(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

杰卡德距离用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度。 数据科学与AI

十、相似系数与相似距离

- 相关系数

是衡量随机变量X与Y相关程度的一种方法，相关系数的取值范围是[-1,1]。相关系数的绝对值越大，则表明X与Y相关度越高。当X与Y线性相关时，相关系数取值为1（正线性相关）或-1（负线性相关）

$$\rho_{xy} = \frac{Cov(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X - EX)(Y - EY))}{\sqrt{D(X)}\sqrt{D(Y)}}$$
 数据科学与AI

- 相关距离：

$$D_{xy} = 1 - \rho_{xy}$$
 数据科学与AI

十一、信息熵

信息熵是衡量分布的混乱程度或分散程度的一种度量。分布越分散(或者说分布越平均)，信息熵就越大。分布越有序（或者说分布越集中），信息熵就越小。

- 计算给定的样本集X的信息熵的公式：

$$Entropy(X) = \sum_{i=1}^n -p_i \log_2 p_i$$
 数据科学与AI

喜欢此内容的人还喜欢

一个简单回归案例：初识机器学习过程

使用Python玩转数学

多元回归：理解机器学习

使用Python玩转数学

机器学习数学知识结构图