the ways of computing distance beweent vector *

FrankZhou-jun†

2019 年 11 月 15 日

1 引言

最近做实验，在进行分类操作时，如果能有一种方法度量样本间的相似性，则可以较好的进行试验下去。网上查看了一下，方法有：1. 欧氏距离,2. 曼哈顿距离 3. 切比雪夫距离,4. 闵可夫斯基距离,5. 标准化欧氏距离,6. 马氏距离,7. 夹角余弦,8. 汉明距离,9. 杰卡德距离杰卡德相似系数,10. 相关系数相关距离,11. 信息熵.

2 公式

1. Euclidean Distance(欧氏距离)

   计算欧式空间中两点的距离。两个向量为 $a = (x_{11}, x_{12}, ..., x_{1n})$ 与 $b = (x_{21}, x_{22}, ..., x_{2n})$, 则距离为:

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

   矩阵公式为：

   $$\mathbf{d}_{12} = \sqrt{(a-b)(a-b)^T}$$

2. Manhattan Distance(曼哈顿距离)

   这个比较好理解，当你开车要要从一个十字路口，到另一个十字路口时，最短的距显然不是两点间的之间距离，不同的角度看到的到距离是不一样的。

   (a) 二维平面两点 $a(x_1, y_1)$ 与 $b(x_2, y_2)$ 间的曼哈顿距离。

   $$\mathbf{d}_{12} = |x_1 - x_2| + |y_1 - y_2|$$

   (b) n 维平面两点 $a = (x_{11}, x_{12}, ..., x_{1n})$ 与 $b = (x_{21}, x_{22}, ..., x_{2n})$ 间的曼哈顿距离。

   $$\mathbf{d}_{12} = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

3. Chebyshev Distance(切比雪夫距离)

   国际象棋玩过么？国王走一步能够移动到相邻的 8 个方格中的任意一个。那么国王从格子 (x_1, y_1) 走到格子 (x_2, y_2) 最少需要多少步？自己走走试试。你会发现最少步数总是 $max(|x_2 - x_1|, |y_2 - y_1|)$ 步。有一种类似的一种距离度量方法叫切比雪夫距离。

   (a) 二维平面两点 $a(x_1, y_1)$ 与 $b(x_2, y_2)$ 间的切比雪夫距离。

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

*https://zhuanlan.zhihu.com/p/25932717

†研究方向：信号处理，机械故障诊断，深度学习，强化学习，邮箱:zhoujun14@yeah.net

(b) n 维平面两点 $a=(x_{11},x_{12},...,x_{1n})$ 与 $b=(x_{21},x_{22},...,x_{2n})$ 间的曼哈顿距离。

$$\mathbf{d}_{12}=\max_{k}(|x_{1k}-x_{2k}|),k=0,1,2,3..n$$

另一个等价公式：

$$\mathbf{d}_{12}=\lim_{k\to\infty}(\sum_{i=1}^{n}|x_{1i}-x_{2i}|^{k})^{1/k}$$

4. Minkowski Distance(闵可夫斯基距离)

闵可夫斯基距离包含了欧式距离，曼哈顿距离，切比雪夫距离 n 维平面两点 $a=(x_{11},x_{12},...,x_{1n})$ 与 $b=(x_{21},x_{22},...,x_{2n})$ 间的闵可夫斯基为：

$$\mathbf{d}_{12}=(\sum_{i=1}^{n}|x_{1i}-x_{2i}|^{p})^{1/p}$$

其中 p 是一个变参数。

当 p=1 时，就是曼哈顿距离

当 p=2 时，就是欧氏距离

当 p→∞ 时，就是切比雪夫距离

根据变参数的不同，闵氏距离可以表示一类的距离。

闵氏距离的缺点

闵氏距离，包括曼哈顿距离、欧氏距离和切比雪夫距离都存在明显的缺点。

举个例子：二维样本 (身高, 体重)，其中身高范围是 150 190，体重范围是 50 60，有三个样本：a(180,50)，b(190,50)，c(180,60)。那么 a 与 b 之间的闵氏距离（无论是曼哈顿距离、欧氏距离或切比雪夫距离）等于 a 与 c 之间的闵氏距离，但是身高的 10cm 真的等价于体重的 10kg 么？因此用闵氏距离来衡量这些样本间的相似度很有问题。

简单说来，闵氏距离的缺点主要有两个：(1) 将各个分量的量纲 (scale)，也就是“单位”当作相同的看待了。(2) 没有考虑各个分量的分布（期望，方差等) 可能是不同的。

5. Standardized Euclidean distance(标准化欧氏距离)

标准化欧氏距离是针对简单欧氏距离的缺点而作的一种改进方案。标准欧氏距离的思路：既然数据各维分量的分布不一样，好吧！那我先将各个分量都“标准化”到均值、方差相等吧。均值和方差标准化到多少呢？这里先复习点统计学知识吧，假设样本集 X 的均值 (mean) 为 m，标准差 (standard deviation) 为 s，那么 X 的“标准化变量”表示为：

而且标准化变量的数学期望为 0，方差为 1。因此样本集的标准化过程 (standardization) 用公式描述就是：

$$\mathbf{X}^{*}=\frac{X-m}{s}$$

经过简单的推导就可以得到两个 n 维向量 $a=(x_{11},x_{12},...,x_{1n})$ 与 $b=(x_{21},x_{22},...,x_{2n})$ 间的标准化欧氏距离的公式：

$$\mathbf{d}_{12}=\sqrt{\sum_{k=1}^{n}(\frac{x_{1k}-x_{2k}}{s_k})^{2}}$$

6. Mahalanobis Distance(马氏距离)

有 M 个样本向量 X1 Xm，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：

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

而其中向量 X_i 与 X_j 之间的马氏距离定义为：

$$\mathbf{D}(X_i,X_j)=\sqrt{(\mathbf{X}_i-\mathbf{X}_j)^{T}\mathbf{S}^{-1}(\mathbf{X}_i-\mathbf{X}_j)}$$

若协方差矩阵是单位矩阵（各个样本向量之间独立同分布）, 则公式就成了：

$$\mathbf{D}(X_i, X_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^T(\mathbf{X}_i - \mathbf{X}_j)}$$

也就是欧氏距离了。

若协方差矩阵是对角矩阵，公式变成了标准化欧氏距离。

马氏距离的优缺点：量纲无关，排除变量之间的相关性的干扰。

7. Cosine(夹角余弦)

   对于两个 n 维样本点 $a = (x_{11}, x_{12}, ..., x_{1n})$ 与 $b = (x_{21}, x_{22}, ..., x_{2n})$，可以使用类似于夹角余弦的概念来衡量它们间的相似程度。

$$\cos(\theta) = \frac{ab}{|a||b|}$$

8. Hamming distance(汉明距离)

   两个等长字符串 s1 与 s2 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。

9. accard similarity coefficient(杰卡德相似系数)

   (1) 杰卡德相似系数

   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。

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

   杰卡德相似系数是衡量两个集合的相似度一种指标。

   (2) 杰卡德距离

   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：

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

   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。

10. Correlation coefficient (相关系数) and Correlation distance(与相关距离)

    (1) 相关系数的定义

$$\mathbf{p}_{XY} = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathbf{D}(\mathbf{X})}\sqrt{\mathbf{D}(\mathbf{Y})}}$$

    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）。

    (2) 相关距离的定义

$$\mathbf{D}_{XY} = 1 - \mathbf{p}_{XY}$$

11. Information Entropy(信息熵) 11. Information Entropy(信息熵)

    信息熵并不属于一种相似性度量。那为什么放在这篇文章中啊？这个。。。我也不知道。( )

    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：

$$Entropy(X) = \sum_{i=1}^{n} -p_i \log_2 p_i$$