一揽子距离问题

zhoulu NJU Apr 12
zhoulujxys@gmail.com

Norm and distance

- A **norm** is a function that assigns a strictly positive *length* or *size* to each vector in a vector space.
- **Distance** is a numerical description of how far apart objects are. In mathematics, in particular geometry, a distance function on a given set M is a function d: $M \times M \rightarrow \mathbf{R}$, where **R** denotes the set of real numbers.
- 范数的本质是距离，是具有“长度”概念的函数。存在的意义是为了实现比较。

More about distance funt

- $D(x,y) \geq 0$, and $d(x,y) = 0$ [if and only if] $x = y$.（非负性）

- $d(x,y) = d(y,x)$. (The distance between x and y is the same in either direction.) （对称性）

- $d(x,z) \leq d(x,y) + d(y,z)$. （满足三角不等式）

$$\text{1-norm distance} = \sum_{i=1}^{n} |x_i - y_i|$$

$$\text{2-norm distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

$$p\text{-norm distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

$$\begin{aligned}\text{infinity norm distance} &= \lim_{p\to\infty}\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \\ &= \max\left(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\right).\end{aligned}$$

- 证明：夹逼准则

一系列的距离定义

- （以二维空间为例，n维空间类似）
- 欧几里得距离：两点之间的直线距离
- 曼哈顿距离：两个点在标准坐标系上的绝对轴距之
  $|x_1 - x_2| + | y_1 - y_2|$
- 切比雪夫距离：定义为其各座标数值差的最大值
  Max($|x_1 - x_2| + | y_1 - y_2|$)

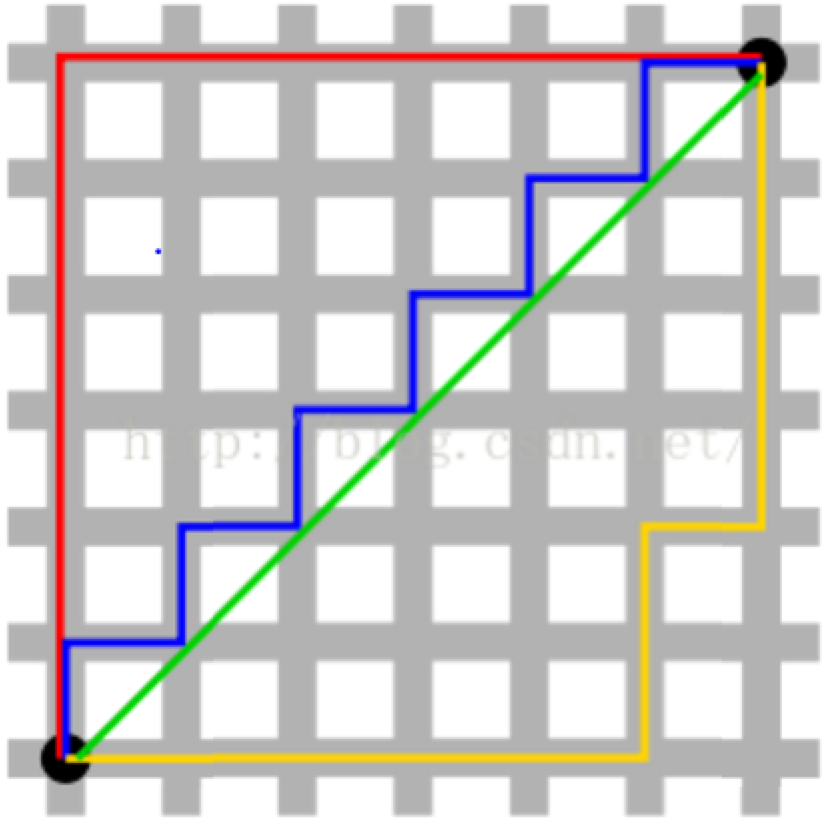
闵可夫斯基距离： $$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}.$$

看看上述距离定义是否满足 General metric

	非负性	对称性	三角不等式
欧式距离	满足	满足	满足（证明）
曼哈顿距离	满足	满足	满足（证明）
切比雪夫距离	满足	满足	满足（证明）

欧式距离： 取α，β 表示n维向量。
|α+ β|=√（α+β，α+β)=√(|α|²+2(α，β)+|β|²)<=√(|α|²+2|α|·|β|+|β|²）（柯西-施瓦茨不等式 $|\langle x, y\rangle| \le \|x\| \cdot \|y\|.$ ）=|α|+|β|
其中（α+β，α+β）=（α，α）+2（α，β）+（β，β）
（α，β)表示内积。
曼哈顿距离： 取等 iff 第三个点在初始的两个点所围成的区域中。
切比雪夫距离：取α，β 表示n维向量,取 Z = α+β
（绝对值不等式）

$$\max_i |z_i| = \max_i |x_i + y_i| \le \max_i(|x_i| + |y_i|) \le \max_i |x_i| + \max_i |y_i|.$$



通俗理解： 欧式距离：斜边

曼哈顿距离：两直角边和

切比雪夫距离：两直角边中的较大者

问题讨论：量化距离

- 假设你是服装厂的老总，你手上有一批数据：包括社会上随机人群的身高和体重。
- 你需要根据上述的大量数据设定服装厂生产的服装 size 的产能分配。

- 即问题需要根据身高和体重来度量人的体型的相似程度 。

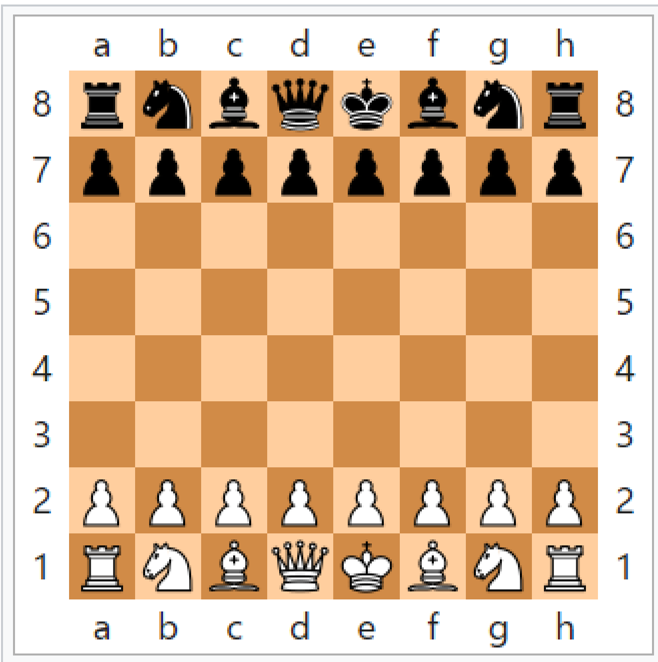
- 举个例子：张三身高180cm，体重100kg，李四身高175cm,体重90kg。我们应该怎么样来定义上述两人的体型的距离。

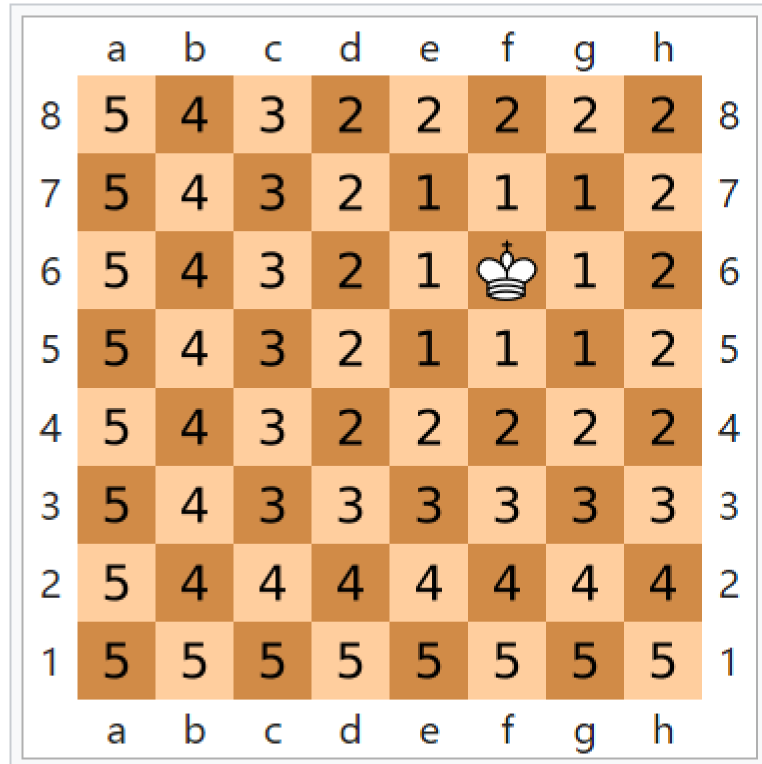
上述不同距离的性质

- 欧式距离：将样品的不同属性（即各指标或各变量）之间的差别等同看待。同时从距离角度来看，它认为两点之间，**始终是可以通过直线距离到达的（更适用于欧氏空间）**。
- 从原点到目标点，往往直走不能到达的。曼哈顿距离加入了一些这方面的考虑。但是在其距离计算中，认为各个维度对于距离d的贡献权重是一样的。
- 切比雪夫中说的距离则是两者的结合体（即可直线走，也可沿着格子走）。二个点之间的距离定义是其**各坐标数值差绝对值的最大值**。此距离中，加入了优化的成分，通过最值来定义距离。

- 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。
- 所以人类发明了**标准化欧氏距离：针对**不同量纲权重差异可能带来的影响。（感兴趣的同学就直行查阅下吧）

上述距离的应用





- 王：横、直、斜走均可，但每次只能走一格