各种距离

谢逸

2019 年 4 月 29 日

南京大学匡亚明学院
171240536@smail.nju.edu.cn

问题

请查阅资料，介绍曼哈顿距离、欧几里得距离、契比雪夫距离分别是什么意思，他们的典型应用是什么。还有哪些创意，来定义二进制位串之间的距离？

回顾

问题：什么是距离？

距离的定义

设 X 是非空集合，对于 X 中任意的两个元素 x 与 y，按某一法则都对应唯一的实数 $d(x,y)$，而且满足下述三条公理：
(1)(正定性)$d(x,y) \geq 0$, 且 $d(x,y) = 0$，当且仅当 $x = y$;
(2)(对称性)$d(x,y) = d(y,x)$;
(3)(三角不等式) 对于任意的 $x, y, z \in X$，恒有 $d(x,y) \leq d(x,z) + d(z,y)$.
则称 $d(x,y)$ 为 x 与 y 的 **距离**，并称 X 是以 d 为距离的 **距离空间**，记作 (X,d)。通常，在距离已被定义的情况下，(X,d) 可以简单地将 X 中的元素称为 X 中的 **点**.

三种距离

欧几里得空间中三种距离介绍:
(1) 欧几里得距离;
(2) 曼哈顿距离;
(3) 切比雪夫距离

欧几里得空间

定义

设V是实数域R上的线性空间（或称为向量空间），若V上定义着正定对称双线性型g（g称为内积），则V称为（对于g的）内积空间或欧几里德空间（有时仅当V是有限维时，才称为欧几里德空间）。[3] 具体来说，g是V上的二元实值函数，满足如下关系：

（1）g(x,y)=g(y,x)；

（2）g(x+y,z)=g(x,z)+g(y,z)；

（3）g(kx,y)=kg(x,y)；

（4）g(x,x)>=0，而且g(x,x)=0当且仅当x=0时成立。

这里x,y,z是V中任意向量，k是任意实数。

欧几里得距离

在欧几里得空间中，点 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$ 之间的欧几里得距离定义为

$$d = (\sum_{i=1}^{n} |x_i - y_i|^2)^{\frac{1}{2}}$$

又称 l2 距离.
就是我们所理解的" 直线距离" !

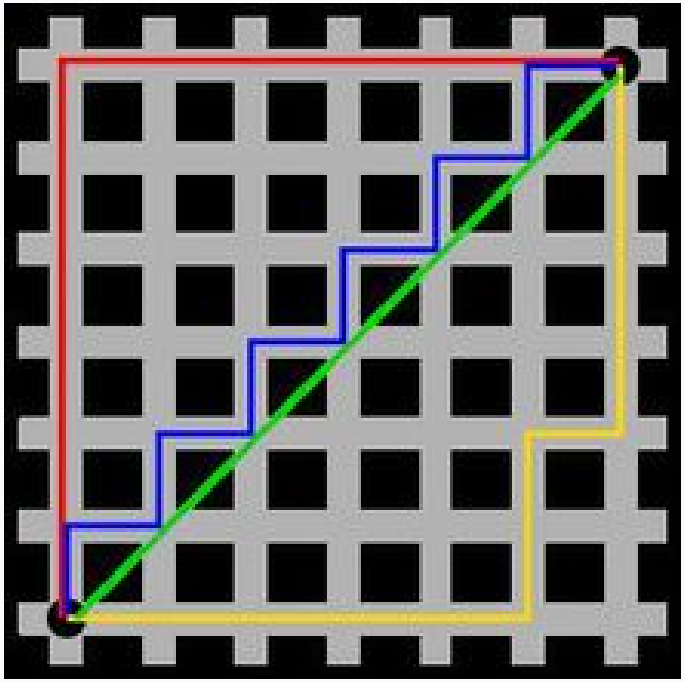
曼哈顿距离

在欧几里得空间中, 点 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$ 之间的曼哈顿距离定义为

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

又称为出租车度量, l1 距离.

曼哈顿距离



曼哈顿距离

物理意义:
曼哈顿的大部分街道都是网格状布局, 出租车想要从十字路口到达另一个十字路口只能沿着网格状街道行驶, 所行驶的最短距离就是曼哈顿距离, 也被称为出租车度量.

曼哈顿距离

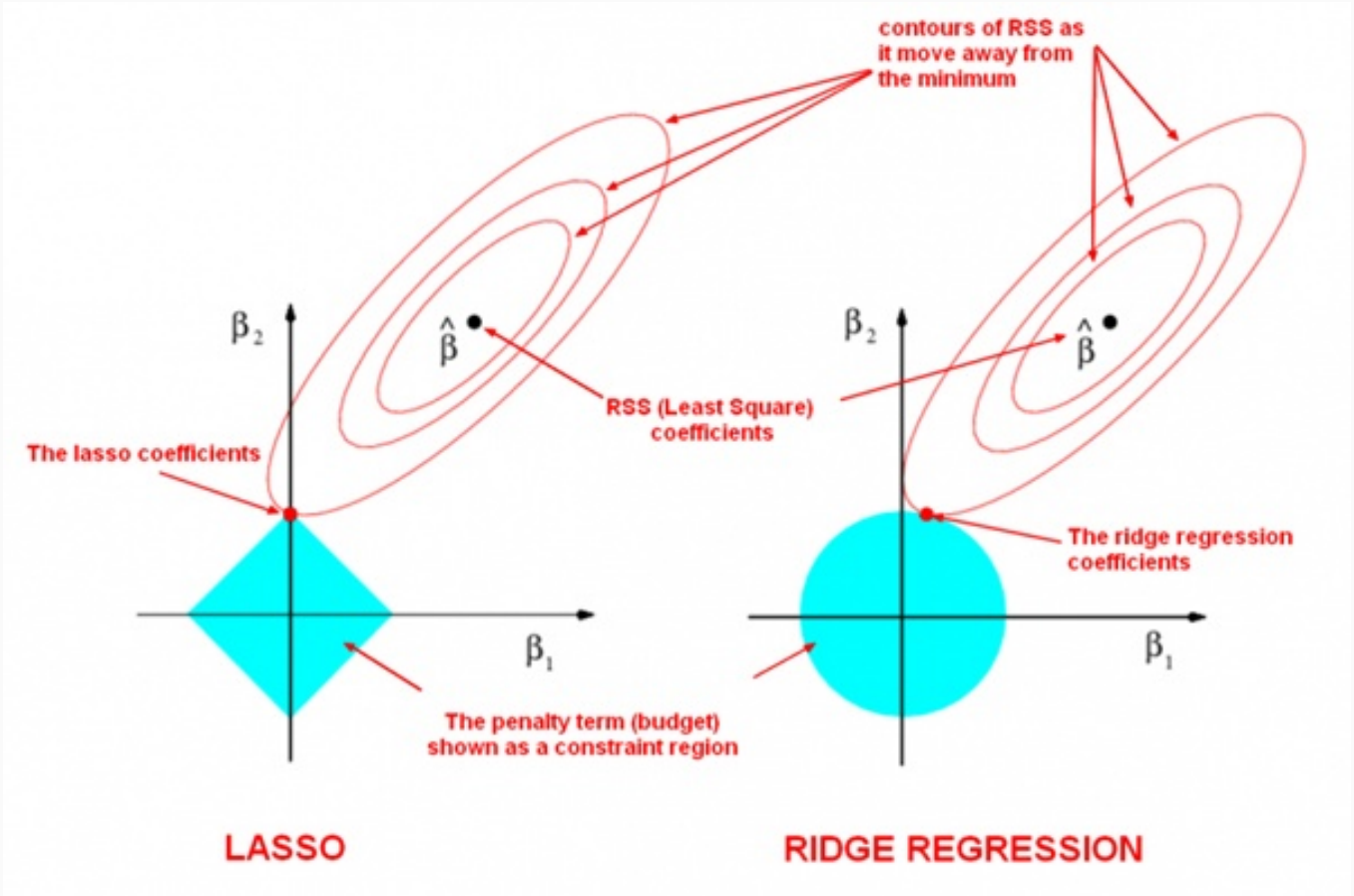
问题：二维情况下，按照曼哈顿距离，”圆”是什么样的？三维中的”球”呢？

曼哈顿距离应用

(1) 网格状公路中的路径问题.
(2) 国际象棋中某些只能横着或竖着走 (或只能斜着走) 的棋子.
(3) 将 n 位 01 编码看做 n 位坐标, 则汉明距离就是曼哈顿距离.
(4) 在 sklearn.linear.model.SGDClassifier 中, 参数 $l1\_ratio$ 决定了弹性网络惩罚项中距离的种类. 当 $l1\_ratio = 1$ 时, 使用 l1 距离 (即曼哈顿距离); 当 $l1\_ratio = 0$ 时, 使用 l2 距离 (即欧几里得距离); 否则按照 $l1\_ratio : (1 - l1\_ratio)$ 的比例混合使用曼哈顿距离和欧几里得距离. l1 距离和 l2 距离分别对应两种回归方式: lasso 和 ridge.

曼哈顿距离应用

lasso 在优化过程的目标函数中使用如下的 l1 penalty, 从而把一些线性回归项的系数“逼成”零, 在维持简洁性上更具优势;
ridge 是用 l2 penalty，旨在把系数变得小一些，但非完全成零, 在防止过拟合上力压群芳。



曼哈顿距离的优点

如果点坐标为整数，那么曼哈顿距离的计算可以避免浮点数运算，并且不管累计运算多少次，都不会有误差. 因此在早期的计算机图形学中广泛应用.
但是，坐标轴变化后，曼哈顿距离可能会改变!

切比雪夫距离

在欧几里得空间中, 点 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$ 之间的切比雪夫距离定义为

$$d = \max_{i=1}^{n} |x_i - y_i|$$

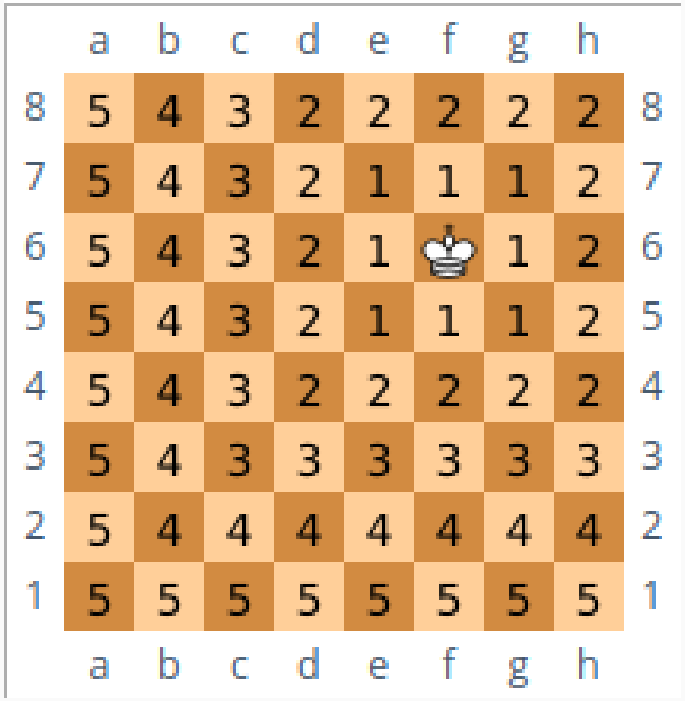
又称 l_∞ 距离 (度量).

切比雪夫距离

问题：二维情况下，按照切比雪夫距离，”圆”是什么样的？三维中的”球”呢？

切比雪夫距离

被称为棋盘距离，因为在国际象棋游戏中，国王从棋盘上的一个位置到另一个位置所需的最小移动次数等于正方形中心之间的切比雪夫距离.



切比雪夫距离应用

(1) 切比雪夫距离有时用于仓库物流，因为它有效地测量了高架起重机移动物体所需的时间（因为起重机可以同时在 x 和 y 轴上移动）。
(2) 在平面中操作的许多工具，例如绘图或钻孔机，光绘机等，通常在 x 和 y 方向上由两个马达控制，类似于桥式起重机。

闵可夫斯基距离

问题: 为什么这三个距离分别被称为 l_2 距离, l_1 距离和 l_∞ 距离?

闵可夫斯基距离

三大距离的统一: 引入闵可夫斯基距离的概念.
在欧几里得空间中, 点 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$ 之间的闵可夫斯基距离定义为

$$d = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

闵可夫斯基距离

$p = 1$, 闵可夫斯基距离就是曼哈顿距离;
$p = 2$, 闵可夫斯基距离就是欧几里得距离;
$p = \infty$, 闵可夫斯基距离就是切比雪夫距离;(为什么?)

闵可夫斯基距离

令 d_C 表示切比雪夫距离，d 表示闵可夫斯基距离，即

$$d_C = \max_{i=1}^{n} |x_i - y_i|$$

，

$$d = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

因为

$$d_C = ((\max_{i=1}^{n} |x_i - y_i|)^p)^{\frac{1}{p}} \leq d \leq (n(\max_{i=1}^{n} |x_i - y_i|)^p)^{\frac{1}{p}} = n^{\frac{1}{p}} d_C$$

$$\lim_{p \to \infty} d_C = d_C$$

$$\lim_{p \to \infty} n^{\frac{1}{p}} d_C = d_C$$

夹逼法则:

$$\lim_{p \to \infty} d = d_C$$

二进制位串/字符串间的距离

(1) 汉明距离 (你们都懂了, 略)
(2) Lee 距离
(3) Levenshtein 距离
(4) Jaro–Winkler 距离 (因为不满足三角不等式所以略去, 感兴趣的自己 wiki)

Lee 距离

在字符串的汉明距离计算中，只关心了每个字符是否相同．如果要需要定义不同字符间距离不同呢？

Lee 距离

Lee 距离是两个字符串之间的距离. 字母表为 q 进制字母表 $\{0, 1, \cdots, q-1\}$, $(q \geq 2)$ 字符串 $x_1 x_2 ... x_n$ 和 $y_1 y_2 ... y_n$ 之间的 Lee 距离定义为

$$\sum_{i=1}^{m} \min(|x_i - y_i|, q - |x_i - y_i|)$$

思想: 假设字母表是一个环, 以字母表为 a-z 为例, a 和 c 距离为 2, a 和 z 距离为 1.

Levenshtein 距离

如果字符串的长度不一样呢?

Levenshtein 距离

两个单词之间的 Levenshtein 距离是将一个单词更改为另一个单词所需的单字符编辑（插入，删除或替换）的最小数量，也称为编辑距离.

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

这是一个 dp 过程. 终止条件为字符串之一长度为 0, min 中的三个分别对应于删除 a(插入 b), 插入 a(删除 b), 修改 (a 或 b)(一样就不修改). 常用于 DNA 间距离测定.

思考题

思考题: Levenshtein 距离是否满足三角不等式? 为什么?

谢谢!