数据挖掘 第3周课后作业

3.1 距离计算。

给定两个用元组$(22, 1, 42, 10)$和$(20, 0, 36, 8)$表示的对象：

(1) 计算这两个对象之间的欧几里得距离。

(2) 计算这两个对象之间的曼哈顿距离。

(3) 使用$q = 3$，计算这两个对象之间的闵可夫斯基距离。

(4) 计算这两个对象之间的上确界距离。

(1) $d1 = 6.708203932499369$

(2) $d2 = 11.0$

(3) $d3 = 6.708203932499369$

(4) $d4 = 6.0$

3.2 相似性度量选择。

假设有如下二维数据集：

	A1	A2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

(1) 把该数据看作二维数据点。给定一个新数据点 $x = (1.4, 1.6)$作为查询点，使用欧几里得距离、曼哈顿距离、上确界距离和余弦相似性，基于与查询点的相似性对数据集中的点进行排序。

(2) 规格化该数据集，使得每个数据点的范数等于1。在变换后的数据上使用欧几里得距离对该数据集中的点重新排序。

(1)

数据点	欧几里得距离	曼哈顿距离	上确界距离	余弦相似性
$x1$	0.14142135623730948	0.10000000000000009	0.19999999999999996	0.9999957
$x2$	0.6708203932499369	0.6000000000000001	0.8999999999999999	0.99787613
$x3$	0.28284271247461906	0.20000000000000018	0.40000000000000013	0.99998474
$x4$	0.22360679774997896	0.19999999999999996	0.30000000000000004	0.99951412
$x5$	0.608276253029822	0.6000000000000001	0.7000000000000002	0.9826817

基于相似性对数据排序为 $x1 > x3 > x4 > x2 > x5$

(2) 规格化数据集得到：

数据点	原数据	规格化后数据	规格化后欧几里得距离
x	(1.4, 1.6)	(0.65850461, 0.75257669)	-
$x1$	(1.5, 1.7)	(.66162164, 0.74983786)	0.004149350803200864
$x2$	(2.0, 1.9)	(0.72499943, 0.68874946)	0.09217091457843411
$x3$	(1.6, 1.8)	(0.66436384, 0.74740932)	0.007812321193114019
$x4$	(1.2, 1.5)	(0.62469505, 0.78086881)	0.044085486555962686
$x5$	(1.5, 1.0)	(0.83205029, 0.5547002)	0.2631980507972417

按照欧几里得距离重新排序得到 $x5 > x2 > x4 > x3 > x1$