2013 教育部-IBM 产学合作专业综合改革项目
厦门大学《数据挖掘原理及实践》课程习题

第 2 章 数据的属性和表现形式

1. 进一步考察余弦度量和相关性度量:

(1) 对于余弦度量, 可能的值域是什么?

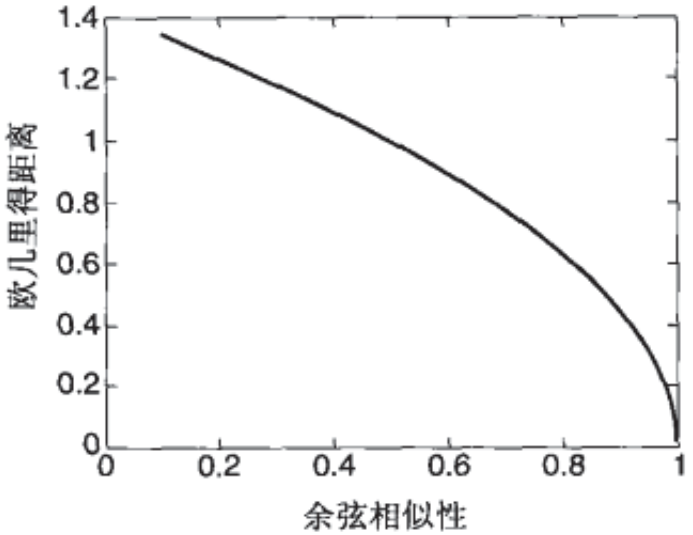
(2) 如果两个对象的余弦度量为 1, 它们相等么? 解释原因。

(3) 如果余弦度量与相关性度量有关系的话, 有何关系? (提示: 在余弦和相关性相同或不同情况下, 考虑诸如均值. 标准差等统计量)

2. 请举出 3 个数据可视化的例子。

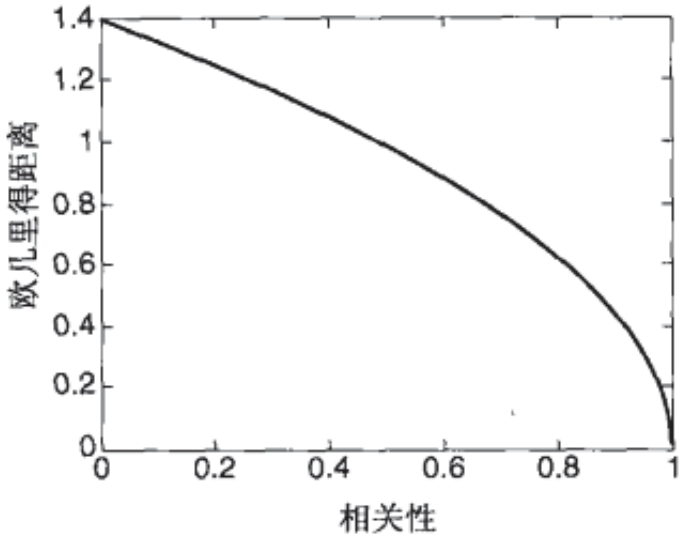
3. 数据属性主要有哪些类型, 它们的特点是什么?

4. 下图显示了 100000 个随机生成的点的相关性度量和欧几里德距离之间的关系, 其中这些点已经规范化, L_2 长度为 1。当向量的 L_2 长度为 1 时, 关于欧几里德距离与余弦相似性之间的关系, 能得到什么一般的观测结论?



欧几里得距离与余弦度量之间的关系

5. 在下图显示了 100000 个随机生成的点的相关性度量与欧几里德距离之间的关系，这些点已经标准化，具有均值为 0，标准差为 1。当向量已经标准化，具有均值 0 和标准差 1 时，关于欧几里德距离与相关性之间的关系，能得到什么的一般观测结论?



欧几里得距离与相关性之间的关系

6. 给定一个区间[0,1]取值的相似性度量，描述两种将该相似度变换成区间[0,∞]中的相异度方法。

7. 请简述数据的中心趋势度量有哪些指标参数。

8. 通常，邻近度定义在一对对象之间。

(1) 阐述两种定义一组对象之间邻近度的方法。

(2) 如何定义欧几里德空间中两个点集之间的距离?

(3) 如何定义两个数据对象集之间的邻近度? (除邻近度定义在任意一对对象之间外，对数据对象不做任何假定。)

9. 给定欧几里德空间中的一个点集 S，以及 S 中每个点到 **x** 的距离(**x** 是否属于 S 并不重要。)

(1) 如果目标是发现点 **y** (**y**≠**x**)，指定距离 Ɛ 内的所有点，解释如何利用三角不等式和已经计算的到 x 的距离，来减少必须的距离计算数量。提示，三角不等式 d(**x** ,**z**) ≤ d(**x**, **y**) + d(**y**, **z**)可以写成 d(**x**, **y**) ≥ d(x, z)-d(y, z)。

(2) x 与 y 之间的距离对距离计算的数量有何影响。

10. 假定你可以从原来的数据点的集合中发现一个较小的子集 S’，使得数据集中每个点都至少到 S’中的一个点的距离不超过指定的 Ɛ，并且你还得到了 S’中每对点之间的距离矩阵。描述一种技术，使用这些信息，以最少的距离计算量，从数据集中计算到一个指定点距离不超过 β 的所有点的集合。