numeric similarity 距离度量总结

1. Manhattan距离 (Manhattan Distance)

- **公式**:
  - $d(i,j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$
  - 其中，x_{ik} 和 x_{jk} 是对象 i 和 j 在第 k 个属性上的值。
- **实例**:
  点 $A(1,2)$ 和 $B(4,6)$:
  $d(A,B) = |1-4| + |2-6| = 3 + 4 = 7$。
- **使用场景**:
  适用于网格状路径（如城市街道）或需要计算绝对差值的场景。

2. Minkowski距离 (Minkowski Distance)

- **公式**:
  - $d(i,j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p\right)^{1/p}$
  - 其中 p 是阶数，当 $p = 1$ 时为Manhattan距离，当 $p = 2$ 时为Euclidean距离。
- **实例**:
  点 $A(1,2)$ 和 $B(4,6)$，当 $p = 3$:
  $d(A,B) = \left(|1-4|^3 + |2-6|^3\right)^{1/3} = (27+64)^{1/3} = \sqrt[3]{91}$。
- **使用场景**:
  可调节 p 值以适应不同的距离度量需求。

3. 欧式距离 (Euclidean Distance)

- **公式**:
  - $d(i,j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$
  - 是Minkowski距离的特例，当 $p = 2$。
- **实例**:
  点 $A(1,2)$ 和 $B(4,6)$:
  $d(A,B) = \sqrt{(1-4)^2 + (2-6)^2} = \sqrt{9+16} = 5$。
- **使用场景**:
  常用于几何空间中的距离计算，如图像处理和聚类分析。

4. Cosine相似度 (Cosine Similarity)

- **公式**:
  - $\mathrm{sim}(i,j) = \frac{\sum_{k=1}^{n} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} x_{jk}^2}}$
  - 用于衡量两个向量之间的夹角余弦值。
- **实例**:
  向量 $A = (1,2,3)$ 和 $B = (4,5,6)$:
  $\mathrm{sim}(A,B) = \frac{(1\cdot4)+(2\cdot5)+(3\cdot6)}{\sqrt{1^2+2^2+3^2} \cdot \sqrt{4^2+5^2+6^2}} = \frac{32}{\sqrt{14}\cdot\sqrt{77}}$。

- **使用场景**:
  适用于文本分析、推荐系统等需要比较向量方向而非大小的场景。

5. Supremum距离 (Supremum Distance)

- 定义
  - Supremum距离，也称为L∞距离或切比雪夫距离（Chebyshev Distance），是Minkowski距离的特例，当 $p \to \infty$ 时。
  - 它衡量两个点之间在所有维度上的最大差异。

- 公式
  - $d(i, j) = \lim_{p\to\infty} \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p\right)^{1/p} = \max_k |x_{ik} - x_{jk}|$
  - 其中，x_{ik} 和 x_{jk} 是对象 i 和 j 在第 k 个属性上的值。

示例

- 假设有两个点 $A(1, 2, 3)$ 和 $B(4, 6, 5)$:
  $d(A, B) = \max(|1 - 4|, |2 - 6|, |3 - 5|) = \max(3, 4, 2) = 4$ 。

使用场景

- Supremum距离适用于需要关注最大差异的场景，例如：
  - 检测多维数据中某一维度的极端变化。
  - 在棋盘游戏中，计算国王移动的最短步数（曼哈顿距离和欧式距离不适用）。
- 它对异常值敏感，因此在异常检测中可能有用。

距离度量的区别与联系

距离度量	特点	联系	使用场景
Manhattan距离		计算绝对差值之和，适合高维稀疏数据	Minkowski距离的特例，当 $p = 1$
欧式距离	计算直线距离，适合几何空间	Minkowski距离的特例，当 $p = 2$	几何空间计算，如图像处理、聚类分析
Minkowski距离		广义形式，可调节 p 值适应不同需求	包含Manhattan距离和欧式距离作为特例
Cosine相似度	衡量向量夹角余弦值，关注方向而非大小	与其他距离度量不同，适合高维稀疏数据	文本分析、推荐系统等需要比较向量方向的场景

Nominal Attribute Similarity

计算方法

- 对于名义属性（Nominal Attribute），相似性可以通过匹配的属性值数量来计算。
- **公式**:
  - $\text{sim}(i, j) = \frac{\text{matched attributes}}{\text{total attributes}}$

- 其中，matched attributes 表示对象 i 和 j 在相同状态下的属性数量。

示例

- 假设有两个对象 A 和 B，它们的属性如下：
  $A = \text{red, single, cat}$
  $B = \text{red, married, cat}$
  匹配的属性数量为 2，总属性数量为 3。
  $\mathrm{sim}(A, B) = \frac{2}{3} = 0.67$。$\mathrm{dist}(A, B) = 1 - \mathrm{sim}(A, B)$

使用场景

- 适用于分类数据，例如客户的婚姻状态、职业、宠物类型等。
- 常用于聚类分析和分类任务中，尤其是处理非数值型数据时。

Jaccard距离 (Jaccard Distance)

公式

- **Jaccard相似度**:
  $\mathrm{sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$
  其中，$|A \cap B|$ 是两个集合的交集大小，$|A \cup B|$ 是两个集合的并集大小。
- **Jaccard距离**:
  $\mathrm{dist}(A, B) = 1 - \mathrm{sim}(A, B)$

示例

- 假设集合 $A = 1, 2, 3$，集合 $B = 2, 3, 4, 5$：
  $|A \cap B| = 2$，$|A \cup B| = 5$
  $\mathrm{sim}(A, B) = \frac{2}{5} = 0.4$
  $\mathrm{dist}(A, B) = 1 - 0.4 = 0.6$

使用场景

- 适用于比较两个集合的相似性，尤其是稀疏数据或二元属性的场景。
- 常用于文本分析（如关键词集合的比较）、推荐系统、聚类分析等。

Binary Attribute Proximity

计算方法

- **Binary Attributes**: 只有两个状态（如 {0, 1} 或 {True, False}）。
- **对称二元属性**: 两种状态同等重要（如性别）。
  - **公式**:
    $\mathrm{sim}(i, j) = \frac{\text{matches}}{\text{total attributes}}$
    $\mathrm{dist}(i, j) = 1 - \mathrm{sim}(i, j)$
- **非对称二元属性**: 一种状态比另一种更重要（如医疗测试结果）。

- **Jaccard相似系数**:
  $\mathrm{sim(i, j)} = \frac{\mathrm{attributes\ where\ both\ are\ 1}}{\mathrm{attributes\ where\ at\ least\ one\ is\ 1}}$
  $\mathrm{dist(i, j)} = 1 - \mathrm{sim(i, j)}$

示例

- 假设有两个对象 A 和 B，它们的二元属性如下：
  $A = 1, 0, 1, 1$
  $B = 1, 1, 0, 1$
  - 对称相似度:
    $\mathrm{sim(A, B)} = \frac{2}{4} = 0.5$
    $\mathrm{dist(A, B)} = 1 - 0.5 = 0.5$
  - 非对称相似度（Jaccard）:
    $\mathrm{sim(A, B)} = \frac{2}{3} = 0.67$
    $\mathrm{dist(A, B)} = 1 - 0.67 = 0.33$

使用场景

- **对称属性**: 性别、是否已婚等。
- **非对称属性**: 医疗测试结果（阳性 vs 阴性）、故障检测（故障 vs 正常）等.

Ordinal Attribute Proximity

计算方法

- 对于有序属性（Ordinal Attribute），需要考虑值的顺序关系。
- **步骤**:
  1. 将每个值替换为其对应的**秩值**（Rank），例如：{low, medium, high} → {1, 2, 3}。
  2. 将秩值归一化到 $[0, 1]$，公式为：
     $\mathrm{z_i} = \frac{\mathrm{rank_i} - 1}{\mathrm{M} - 1}$
     其中，M 是属性的可能值总数。
  3. 使用数值属性的距离度量（如曼哈顿距离或欧式距离）计算归一化秩值之间的距离。

示例

- 假设属性为 {low, medium, high}，对应的秩值为 {1, 2, 3}，归一化后为 {0, 0.5, 1}。
- 比较 low 和 high 的距离：
  $\mathrm{d(low, high)} = |0 - 1| = 1$。
- 比较 medium 和 high 的距离：
  $\mathrm{d(medium, high)} = |0.5 - 1| = 0.5$。

使用场景

- 适用于有序但非数值型的数据，例如教育水平（高中、本科、硕士、博士）、满意度（非常不满意、不满意、中立、满意、非常满意）等。
- 常用于分类、聚类和排序任务中.