

数据科学的数学基础

第次理论作业

陈万祺 3220102895

2025 年 10 月 30 日

Problem 1

Prove that the distance induced by Rogers-Tanimoto similarity is a metric.

解：

由定义 Rogers-Tanimoto 相似度为

$$s_{RT}(p, q) = \frac{n - d_H(p, q)}{n + d_H(p, q)},$$

诱导距离定义为

$$d_{RT}(p, q) = \sqrt{s_{RT}(p, p) + s_{RT}(q, q) - 2s_{RT}(p, q)}.$$

由于 $s_{RT}(p, p) = 1$ 和 $s_{RT}(q, q) = 1$ ，我们有

$$d_{RT}(p, q) = \sqrt{2(1 - s_{RT}(p, q))} = \sqrt{2 \left(1 - \frac{n - d_H(p, q)}{n + d_H(p, q)} \right)} = \sqrt{2 \left(\frac{2d_H(p, q)}{n + d_H(p, q)} \right)} = \sqrt{\frac{4d_H(p, q)}{n + d_H(p, q)}}.$$

令 $h = d_H(p, q)$ ，则 $d_{RT}(p, q) = 2\sqrt{\frac{h}{n+h}}$ 。

现在证明 d_{RT} 满足度量性质：

1. 非负性：由于 $h \geq 0$ 且 $n + h > 0$ ，所以 $d_{RT}(p, q) \geq 0$ 。
2. 同一性：如果 $p = q$ ，则 $h = 0$ ，所以 $d_{RT}(p, q) = 0$ 。如果 $d_{RT}(p, q) = 0$ ，则 $\frac{4h}{n+h} = 0$ ，所以 $h = 0$ ，即 $p = q$ 。
3. 对称性：由于 $d_H(p, q) = d_H(q, p)$ ，所以 $d_{RT}(p, q) = d_{RT}(q, p)$ 。
4. 三角不等式：考虑函数 $f(x) = \sqrt{\frac{x}{n+x}}$ ，则 $d_{RT}(p, q) = 2f(h(p, q))$ 。由于 d_H 是度量，有 $h(p, r) \leq h(p, q) + h(q, r)$ 。函数 $f(x)$ 是凹函数且 $f(0) = 0$ ，因此对于凹函数有 $f(a+b) \leq f(a) + f(b)$ 。于是，

$$f(h(p, r)) \leq f(h(p, q) + h(q, r)) \leq f(h(p, q)) + f(h(q, r)),$$

所以

$$d_{RT}(p, r) = 2f(h(p, r)) \leq 2f(h(p, q)) + 2f(h(q, r)) = d_{RT}(p, q) + d_{RT}(q, r).$$

三角不等式成立。
因此， d_{RT} 是度量。

Problem 2

Show that the distance induced by Sorensen-Dice similarity is not a metric.

解：

由定义 Sorensen-Dice 相似度为

$$s_{SD}(p, q) = \frac{2p \cdot q}{\|p\|_1 + \|q\|_1},$$

诱导距离定义为

$$d_{SD}(p, q) = \sqrt{s_{SD}(p, p) + s_{SD}(q, q) - 2s_{SD}(p, q)}.$$

由于 $s_{SD}(p, p) = 1$ 和 $s_{SD}(q, q) = 1$ ，我们有

$$d_{SD}(p, q) = \sqrt{2(1 - s_{SD}(p, q))} = \sqrt{2 \left(1 - \frac{2p \cdot q}{\|p\|_1 + \|q\|_1} \right)} = \sqrt{2 \frac{\|p\|_1 + \|q\|_1 - 2p \cdot q}{\|p\|_1 + \|q\|_1}}.$$

注意 $\|p\|_1 + \|q\|_1 - 2p \cdot q = |A \Delta B| = d_H(p, q)$ ，其中 A 和 B 是对应的集合。所以

$$d_{SD}(p, q) = \sqrt{2 \frac{d_H(p, q)}{\|p\|_1 + \|q\|_1}}.$$

现在，我们通过反例证明 d_{SD} 不满足三角不等式。考虑三个集合： $A = \{1\}$ ， $B = \{1, 2\}$ ， $C = \{2\}$ ，对应的二进制向量为 p, q, r 。假设全集为 $\{1, 2\}$ ，则 $n = 2$ ，但这里我们直接计算相似度。

$|A \cap B| = 1, |A| = 1, |B| = 2$ ，所以 $s_{SD}(A, B) = \frac{2 \times 1}{1+2} = \frac{2}{3}$ ，于是 $d_{SD}(A, B) = \sqrt{2 \left(1 - \frac{2}{3} \right)} = \sqrt{\frac{2}{3}} \approx 0.816$ 。

$|B \cap C| = 1, |B| = 2, |C| = 1$ ，所以 $s_{SD}(B, C) = \frac{2 \times 1}{2+1} = \frac{2}{3}$ ，于是 $d_{SD}(B, C) = \sqrt{\frac{2}{3}} \approx 0.816$ 。

$|A \cap C| = 0, |A| = 1, |C| = 1$ ，所以 $s_{SD}(A, C) = \frac{2 \times 0}{1+1} = 0$ ，于是 $d_{SD}(A, C) = \sqrt{2(1 - 0)} = \sqrt{2} \approx 1.414$ 。

现在检查三角不等式：

$$d_{SD}(A, B) + d_{SD}(B, C) \approx 0.816 + 0.816 = 1.632 < 1.414 \approx d_{SD}(A, C),$$

违反三角不等式。因此， d_{SD} 不是度量。