Distance and Similarity

1. Euclidean Distance (歐式距離非尤拉距離)

$$d = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2} \quad (1)$$
$$d = \sqrt{(a-b)(a-b)^T} \quad (2)$$

2. Manhattan Distance (曼哈頓距離)

$$d = |x_1 - x_2| + |y_1 - y_2| \quad (1)$$
$$d = \sum_{k=1}^{n} |x_{1k} - x_{2k}| \quad (2)$$

計算複雜度最小

3. Chebyshev Distance (切比雪夫距離)

$$d = \max(|x_1 - x_2|, |y_1 - y_2|) \quad (1)$$

$$d = \max_i(|x_{1i} - x_{2i}|) \quad (2)$$

$$d = \lim_{k\to\infty}(\sum_{i=1}^{n} |x_{1i} - x_{2i}|^k)^{1/k} \quad (3)$$

4. Minkowski Distance (閔可夫斯基距離)

$$d = \sqrt[p]{\sum_{k=1}^{n} |x_{1k} - x_{2k}|^p}$$

1.2.3.4. 有以下的缺點：

a. 將各個分量的量網(單位)當作一樣

b. 沒有考慮各個分量的分布可能是不同的

Example : 二維樣本(身高,體重)

5. Standardized Euclidean distance (標準化歐式距離)

$X^* = \frac{X-m}{s}$ (標準後的值 = (標準化前的值-分量的均值) / 分量的標準差)

$$d = \sqrt{\sum_{k=1}^{n}(\frac{x_{1k}-x_{2k}}{s_k})^2}$$

改善數據各維分量的分布不一樣

6. Mahalanobis Distance (馬氏距離)

$$\mathrm{D(x)} = \sqrt{(X-\mu)^T S^{-1}(X-\mu)} \quad (1)$$

$$\mathrm{D}(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1}(X_i - X_j)} \quad (2)$$

$$\mathrm{D}(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)} \quad (3)$$

改善單位問題，排除變量之間的相關性的干擾

7. Hamming distance (漢明距離)

   兩個等長字串s_1和s_2，將其中一個變為另一個所需要的最小替換次數，

   Example : “1111”和”1001”的距離為 2

   若為兩向量，則距離為兩個向量不同的分量所佔的百分比

   Example : [0 0] 和 [1 0], d = 0.5

   [1 0] 和 [0 2], d = 1

8. Correlation coefficient & Pearson Correlation Coefficient (皮爾森相關係數)

$$\rho = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-E_x)(Y-E_y))}{\sqrt{D(X)}\sqrt{D(Y)}} \text{ (相關係數)}$$

$$= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\cdot\sqrt{n\sum y^2 - (\sum y)^2}}$$

9. Correlation distance (相關距離)

$$D_{xy} = 1 - \rho_{xy} \text{ (相關距離)}$$

10. Cosine (夾角餘弦)

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{{x_1}^2 + {x_2}^2}\sqrt{{y_1}^2 + {y_1}^2}} \quad (1)$$

$$\cos\theta = \frac{a\cdot b}{|a||b|} \quad (2)$$

$$\cos\theta = \frac{\sum_{k=1}^{n} x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n} {x_{1k}}^2}\sqrt{\sum_{k=1}^{n} {x_{2k}}^2}} \quad (3)$$

衡量兩個向量方向的差距

11. Jaccard similarity coefficient (杰卡德相似係數)

$$\mathrm{J}(\mathrm{A},\mathrm{B}) = \frac{|A \cap B|}{|A \cup B|}$$

Example: A [0 1 1 1], B [1 0 1 1]

P : A & B 都是 1 的個數

Q: A 是 1，B 是 0 的個數

R : A 是 0，B 是 1 的個數

S : A & B 都是 0 的個數

$$\mathrm{J}(\mathrm{A},\mathrm{B}) = \frac{p}{p+q+r}$$

12. Jaccard distance (杰卡德相似係數)

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

13. Adjusted Cosine Similarity (調整餘弦相似度)

$$\cos\theta = \frac{(a-\mu_{ab})\cdot(b-\mu_{ab})}{|a-\mu_{ab}||b-\mu_{ab}|}$$

除了分辨個體在維度間的差異，也衡量每個維數值的差異

Example: X 和 Y 的評分內容 X[1,2] Y[4,5], similarity = 0.98

但是 X 卻是不太喜歡，Y 則是極度喜歡，但是 X 和 Y 卻極度相似

X 和 Y 的平均值是 3，X[-2,-1] Y[1,2], similarity = -0.8

如此 X 和 Y 就不相似