Fundamentals of Data Science

Semester B 20-21

Tutorial 3

1. Let **x** and **y** be two data objects where each object has an L_2 length of 1. The relationship between the Euclidean distance and cosine similarity is derived as follows:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sqrt{\sum_{u=1}^{n}(x_u - y_u)^2} \\ &= \sqrt{\sum_{u=1}^{n}(x_u^2 - 2x_u y_u + y_u^2)} \\ &= \sqrt{1 - 2\cos(\mathbf{x},\mathbf{y}) + 1} \\ &= \sqrt{2(1 - \cos(\mathbf{x},\mathbf{y}))} \end{aligned}$$

2. If **x** and **y** are binary vectors, the similarity measure $S(\mathbf{x},\mathbf{y})$ can be simplified as follows:

$$S(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}\cdot\mathbf{y}} = \frac{f_{11}}{(f_{10} + f_{11}) + (f_{01} + f_{11}) - f_{11}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

which corresponds to the Jaccard coefficient.

3. a. Suppose we denote the first attribute as *x*, and the second attribute as *y*.

$$\bar{x} = \frac{2+9+7+5}{4} = 5.75, \quad \bar{y} = \frac{19+6+15+12}{4} = 13$$

$$\sigma_x^2 = \frac{1}{3}[(2-5.75)^2 + (9-5.75)^2 + (7-5.75)^2 + (5-5.75)^2] = 8.917$$

$$\sigma_y^2 = \frac{1}{3}[(19-13)^2 + (6-13)^2 + (15-13)^2 + (12-13)^2] = 30$$

$$\text{covariance}(x, y) = \frac{1}{3}[(2-5.75)(19-13) + (9-5.75)(6-13) + (7-5.75)(15-13) + (5-5.75)(12-13)]$$
$$= -14$$

The covariance matrix is given by $\begin{bmatrix} 8.917 & -14 \\ -14 & 30 \end{bmatrix}$

b. The correlation coefficient between the two attributes is

$$r_{xy} = \frac{-14}{\sqrt{8.917}\sqrt{30}} = -0.86$$