Lecture 27: Application – Data Correlation

Correlation is a measure of how closely two variables are dependent.

Definition

The *mean* μ_X of a data set $X = \{x_1, \ldots, x_n\}$ is the average of the values in the data set.

$$\mu_X = \frac{1}{n}(x_1 + \cdots + x_n).$$

The *correlation of variables X and Y* is

$$\operatorname{corr}(X, Y) = \cos(\theta(X - \mu_X, Y - \mu_Y)), \quad \text{where}$$

$X - \mu_X = |x_1 - \mu_X, \ldots, x_n - \mu_X\rangle$ and $Y - \mu_Y = |y_1 - \mu_Y, \ldots, y_n - \mu_Y\rangle$.

Use $\quad \cos(\theta(\mathbf{u}, \mathbf{v})) = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}$ to compute the correlation.

A value close to 1 indicates the values a highly correlated and a value close to -1 indicates the values are not at all correlated.

Example E3. Suppose the data set is assignment and exam marks for 7 students.

Student	Assignment Mark	Exam Mark
S1	99	100
S2	80	82.5
S3	79	79
S4	75.5	82.5
S5	87.5	91
S6	67	67.5
S7	76	68

The mean assignment mark is

$$\mu_A = \frac{1}{7}(99 + 80 + 79 + 75.5 + 87.5 + 67 + 76) = 80.5.$$

The mean exam mark is

$$\mu_E = \frac{1}{7}(100 + 82.5 + 79 + 82.5 + 91 + 67.5 + 68) = 81.5.$$

Then

$$
\begin{aligned}
A - \mu_A &= |18.5, -0.5, -1.5, -5.5, 7, -13.5, -4.5\rangle, \\
E - \mu_E &= |18.5, 1, -2.5, 1, 9.5, -14, -13.5\rangle
\end{aligned}
$$

and the correlation between the assignment marks and the exam marks is

$$
\begin{aligned}
\mathrm{corr}(A, E) &= \cos(\theta(A - \mu_A, E - \mu_E)) \\
&= \frac{\langle A - \mu_A, E - \mu_E \rangle}{\|A - \mu_!\| \cdot \|E - \mu_E\|} \;= \frac{656.75}{(24.92)(28.62)} \approx 0.92.
\end{aligned}
$$