Lecture Summary: Correlation Coefficient and Scatter Plots

Source: Lecture 3.6.pdf

Key Points

- **Motivation:**
  - Covariance measures the relationship between two random variables, but its magnitude depends on the scale of the variables, making it challenging to interpret.
  - The correlation coefficient, $\rho(X, Y)$, normalizes covariance, providing a dimensionless measure that is easier to interpret.
- **Definition of Correlation Coefficient:**
  - Formula:
    $$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)},$$
    where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y.
  - Properties:
    * $\rho(X, Y) \in [-1, 1]$.
    * $\rho(X, Y) = 1$: Perfect positive correlation (linear relationship where $Y = aX + b$ with $a > 0$).
    * $\rho(X, Y) = -1$: Perfect negative correlation (linear relationship where $Y = aX + b$ with $a < 0$).
    * $\rho(X, Y) = 0$: No linear relationship between X and Y (uncorrelated).
- **Interpreting Correlation:**
  - ρ summarizes the trend between two variables:
    * Positive values: As X increases, Y tends to increase.
    * Negative values: As X increases, Y tends to decrease.
    * Near zero: No observable trend between X and Y.
  - If $|\rho(X, Y)| = 1$, Y is a linear function of X.
- **Scatter Plots for Visualizing Correlation:**
  - A scatter plot represents paired observations (X, Y) as points in a 2D space.
  - Patterns in the scatter plot indicate the type and strength of the correlation:
    * Tight linear alignment: $\rho \approx \pm 1$.
    * Broad linear trend: Moderate correlation ($0 < |\rho| < 1$).
    * Random scatter: $\rho \approx 0$.
  - Example:
    * Case 1: Points are widely spread, $\rho \approx 0$.
    * Case 2: Clear positive trend, $\rho > 0$ but not close to 1.
    * Case 3: Almost perfect positive correlation, $\rho \approx 1$.

    * Case 4: Clear negative trend, $\rho < 0$ but not close to -1.
    * Case 5: Strong negative correlation, $\rho \approx -1$.

- **Examples of Correlation Coefficient:**
  - Joint PMF of X, Y is used to compute $\rho(X, Y)$ systematically.
  - Variance and covariance calculations lead to:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

  - Example calculations demonstrate scenarios with positive, negative, and zero correlations.

- **Applications of Scatter Plots and Correlation:**
  - Scatter plots help identify potential relationships in large datasets, such as sports statistics (e.g., IPL cricket data).
  - Correlation quantifies trends for further analysis or modeling.

Simplified Explanation

Correlation Coefficient (ρ): Measures the strength and direction of a linear relationship between two variables: - 1: Perfect positive correlation. - -1: Perfect negative correlation. - 0: No linear correlation.

Scatter Plots: Visualize paired data to detect patterns: - Tight alignment: Strong correlation. - Random scatter: Weak or no correlation.

Example: For X, Y with $\mathrm{Cov}(X, Y) = -0.5$, $\sigma(X) = 1$, and $\sigma(Y) = 2$:

$$\rho(X, Y) = \frac{-0.5}{1 \cdot 2} = -0.25.$$

Conclusion

In this lecture, we:

- Defined and interpreted the correlation coefficient.
- Demonstrated scatter plots as a tool for visualizing relationships.
- Discussed practical applications and examples of correlation in data analysis.

Correlation is a powerful tool for summarizing trends and relationships in multivariable datasets.