Knowledge Discovery & Data Mining

— Analyzing Feature Relationships —

Instructor: Yong Zhuang

[*yong.zhuang@gvsu.edu*](mailto:yong.zhuang@gvsu.edu)

Outline

- Analyzing Feature Relationships
  - Introduction to Feature Analysis
  - Covariance (for numerical features)
  - Correlation Coefficient (for numerical features)
  - Chi-Square Test (for categorical features)

Why Analyze Relationships Between Features?

- Purpose: Understanding the relationships between features is key to improving predictive models, detecting patterns, and identifying significant associations.
- Key Reasons:
  - Identify Correlations: Determine how one feature may influence or be related to another.
  - Improve Model Performance: Feature relationships can inform better feature selection and model building.
  - Detect Patterns: Recognize trends and patterns within the data.
  - Hypothesis Testing: Verify if observed relationships in the data are statistically significant.

Covariance

Covariance is measure assessing how much two attributes change together.

Consider two numeric attributes A and B and a set of *n* real valued observations {(a1, b1), . . . , (an, bn)}. The mean values (also known as the expected values) of A and B, that is,

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n} \quad \text{and} \quad E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

Then the covariance between A and B is defined as

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

or

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

Covariance

- **Positive covariance**: If Cov(A,B) > 0, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If Cov(A,B) < 0 then if A is larger than its expected value, B is likely to be smaller than its expected value.

If A and B are independent, then $E(A \cdot B) = E(A) \cdot E(B)$. ➡ $Cov(A,B) = 0$

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- E(AllElectronics) =
- E(HighTech) =
- Cov(AllElectroncis, HighTech) =

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- E(AllElectronics) = 4
- E(HighTech) = 10.8
- Cov(AllElectroncis, HighTech)

$$= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Covariance

Example. This table presents a simplified example of stock prices observed at five time points for AllElectronics and HighTech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

- E(AllElectronics) = 4
- E(HighTech) = 10.8
- Cov(AllElectroncis, HighTech)

$$= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80$$

$$= 50.2 - 43.2 = 7.$$

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

stock prices for both companies rise together

Correlation Analysis (Numeric Data)

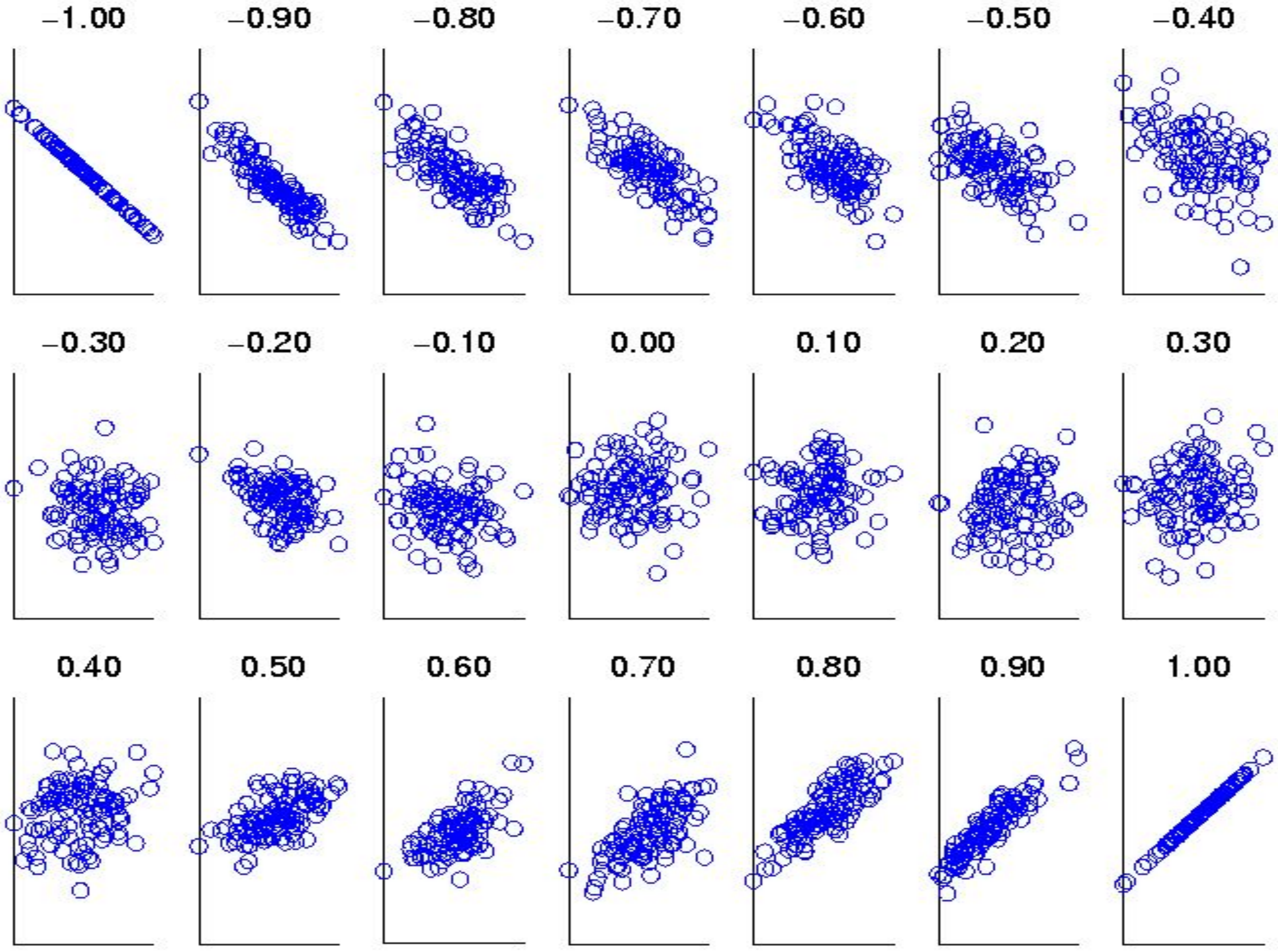
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B} = \frac{Cov(A,B)}{\sigma_A\sigma_B}$$

  where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- $-1 \le r_{A,B} \le +1$, If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher the value, the stronger the correlation
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the correlation coefficient from –1 to 1.

Correlation Coefficient (Numeric Data)

Example. $r_{AE,HT}$?

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

- Cov(AE, HT) = 7

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Correlation Coefficient (Numeric Data)

Example. $r_{AE,HT}$?

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

- Cov(AE, HT) = 7
- $\sigma AE = \sqrt{2} \approx 1.414$
- $\sigma HT = \sqrt{32.56} \approx 5.706$

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$r_{AE,HT} \approx 7/(1.414*5.706) \approx 0.868$

Correlation Analysis (Nominal Data)

For nominal data, a correlation relationship between two attributes, A and B, can be discovered by a χ^2 (chi-square) test.

Correlation Analysis (Nominal Data)

Suppose A has **c** distinct values, namely a1,a2, …, ac. B has **r** distinct values, namely b1,b2, …, br. The data tuples described by A and B can be shown as a **contingency table**, with the **c values of A making up the columns** and the **r values of B making up the rows**. Let (A_i ,B_j) denote the joint event that attribute A takes on value ai and attribute B takes on value bj , that is, where (A = ai ,B = bj). Each and every possible (A_i ,B_j) joint event has its own cell (or slot) in the table.

Correlation Analysis (Nominal Data)

The χ^2 value (also known as the Pearson χ^2 statistic) is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	
Nonfiction(nf)	50	1000	
Sum(col.)			

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j)

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	?
Nonfiction(nf)	50	1000	?
Sum(col.)	?	?	?

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250	200	450
Nonfiction(nf)	50	1000	1050
Sum(col.)	300	1200	1500

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 ($e_{f,g}$?)	200($e_{f,ng}$?)	450
Nonfiction(nf)	50($e_{nf,g}$?)	1000($e_{nf,ng}$?)	1050
Sum(col.)	300	1200	1500

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200($e_{f,ng}$?)	450
Nonfiction(nf)	50($e_{nf,g}$?)	1000($e_{nf,ng}$?)	1050
Sum(col.)	300	1200	1500

$$e_{f,g} = \frac{300 \times 450}{1500} = 90$$

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200(360)	450
Nonfiction(nf)	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

Correlation Analysis (Nominal Data)

Example. Suppose that a group of 1500 people was surveyed. Each person was asked whether his or her preferred type of reading material was fiction or nonfiction, and whether he or she liked playing video games.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

contingency table

	Game(g)	No game(ng)	Sum (row)
Fiction(f)	250 (90)	200(360)	450
Nonfiction(nf)	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Correlation Analysis (Nominal Data)

$\chi^2 = 507.93$

The χ^2 statistic tests the hypothesis that A and B are independent, that is, there is no correlation between them. The test is based on a significance level, with $(r - 1) \times (c - 1)$ degrees of freedom. Since in this example, $r = 2$ and $c = 2$, the degrees of freedom are $(2 - 1) \times (2 - 1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution).

Based on our computed value, we can reject the hypothesis that play_game and preferred_reading are independent, so they are **correlated**.

[Table of upper percentage points of the Chi-squared distribution]()

Summary

- Feature Analysis: Relationships
  - Introduction to Feature Analysis
  - Covariance (for numerical features)
  - Correlation Coefficient (for numerical features)
  - Chi-Square Test (for categorical features)