

The background of the slide is a collage of abstract data visualizations. It features several network graphs with nodes and edges in various colors (red, green, blue, orange). There are also scatter plots with points of different colors (orange, blue, brown) and some pixelated or noisy patterns. The overall aesthetic is technical and data-driven.

Lecture 2. Similarity Measures for Cluster Analysis

Lecture 2. Similarity Measures for Cluster Analysis

- Basic Concept: Measuring Similarity between Objects
- Distance on Numeric Data: Minkowski Distance
- Proximity Measure for Symmetric vs. Asymmetric Binary Variables
- Distance between Categorical Attributes, Ordinal Attributes, and Mixed Types
- Proximity Measure between Two Vectors: Cosine Similarity
- Correlation Measures between Two Variables: Covariance and Correlation Coefficient
- Summary

The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements: a network of thin, light-colored lines forming a mesh or web; numerous small, colored dots (green, blue, orange) scattered across the field; and a grid of small white plus signs. In the upper left, there's a faint, horizontal band of more complex, possibly mathematical or symbolic, patterns. A large, white, semi-transparent trapezoidal shape serves as a backdrop for the title text.

Session 1: Basic Concepts: Measuring Similarity between Objects

What Is Good Clustering?

- ❑ A good clustering method will produce high quality clusters which should have
 - ❑ **High intra-class similarity:** Cohesive within clusters
 - ❑ **Low inter-class similarity:** Distinctive between clusters
- ❑ **Quality function**
 - ❑ There is usually a separate “quality” function that measures the “goodness” of a cluster
 - ❑ It is hard to define “similar enough” or “good enough”
 - ❑ The answer is typically highly subjective
- ❑ There exist many similarity measures and/or functions for different applications
- ❑ Similarity measure is critical for cluster analysis

Similarity, Dissimilarity, and Proximity

□ **Similarity measure or similarity function**

- A real-valued function that quantifies the similarity between two objects
- Measure how two data objects are alike: The higher value, the more alike
- Often falls in the range $[0,1]$: 0: no similarity; 1: completely similar

□ **Dissimilarity (or distance) measure**

- Numerical measure of how different two data objects are
- In some sense, the inverse of similarity: The lower, the more alike
- Minimum dissimilarity is often 0 (i.e., completely similar)
- Range $[0, 1]$ or $[0, \infty)$, depending on the definition

□ **Proximity** usually refers to either similarity or dissimilarity

The background of the slide is a complex, abstract composition. It features a grid of small, light-colored plus signs (+) overlaid on a dark, reddish-brown background. A network of thin, light-colored lines connects various points, creating a web-like structure. In the upper left, there is a small, semi-transparent rectangular area containing a grid of small, colorful dots (green, blue, yellow, and red) arranged in a pattern that resembles a data visualization. The overall aesthetic is technical and data-driven.

Session 2: Distance on Numeric Data: Minkowski Distance

Data Matrix and Dissimilarity Matrix

□ Data matrix

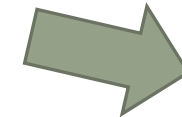
- A data matrix of n data points with l dimensions



$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

□ Dissimilarity (distance) matrix

- n data points, but registers only the distance $d(i, j)$ (typically metric)



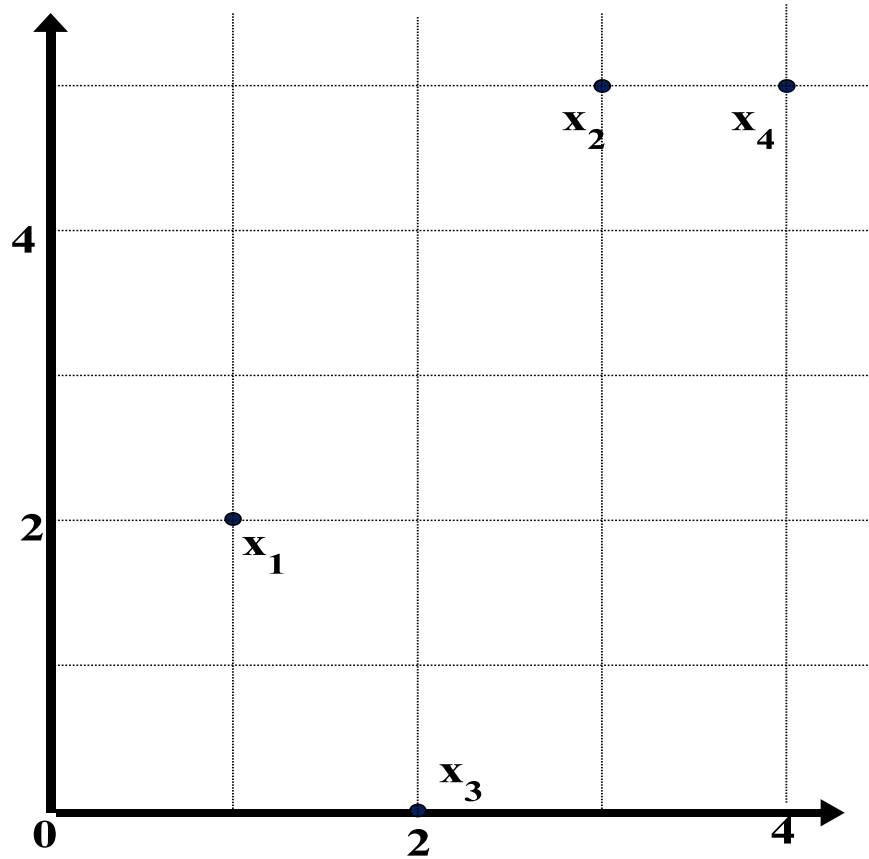
- Usually symmetric, thus a triangular matrix

- **Distance functions** are usually different for real, boolean, categorical, ordinal, ratio, and vector variables

- Weights can be associated with different variables based on applications and data semantics

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

Dissimilarity Matrix (by **Euclidean Distance**)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	2.24	5.1	0	
$x4$	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- **Minkowski distance**: A popular distance measure

$$d(i, j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{il} - x_{jl}|^p}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{il})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jl})$ are two l -dimensional data objects, and p is the order (the distance so defined is also called L- p norm)

- Properties

- $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
- $d(i, j) = d(j, i)$ (Symmetry)
- $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)

- A distance that satisfies these properties is a **metric**
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

□ $p = 1$: (L_1 norm) **Manhattan (or city block) distance**

□ E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|$$

□ $p = 2$: (L_2 norm) **Euclidean distance**

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

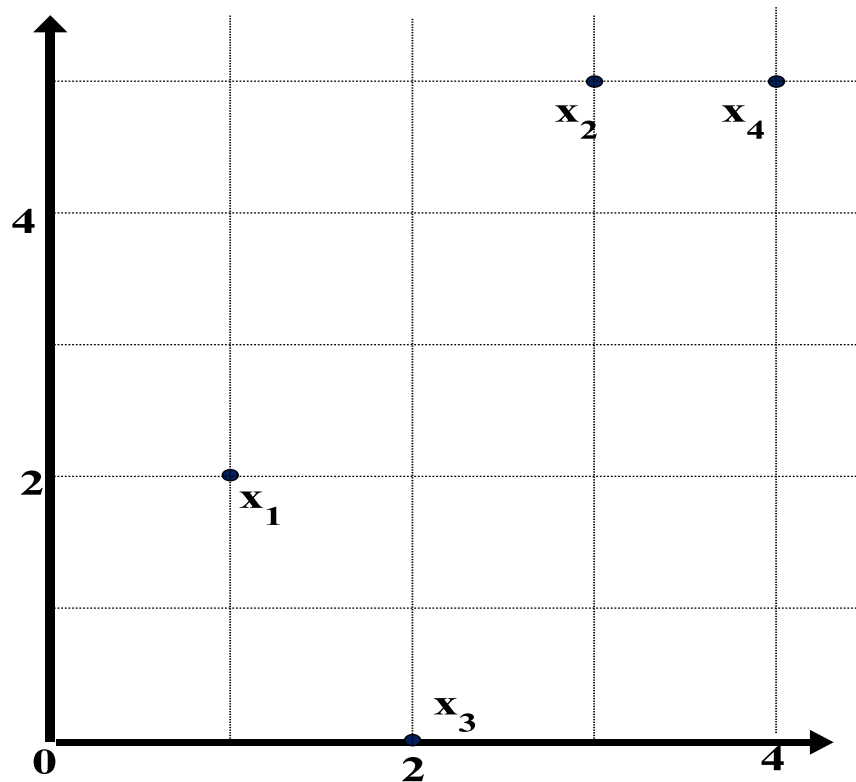
□ $p \rightarrow \infty$: (L_{\max} norm, L_{∞} norm) **“supremum” distance**

□ The maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{p \rightarrow \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^l |x_{if} - x_{jf}|$$

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

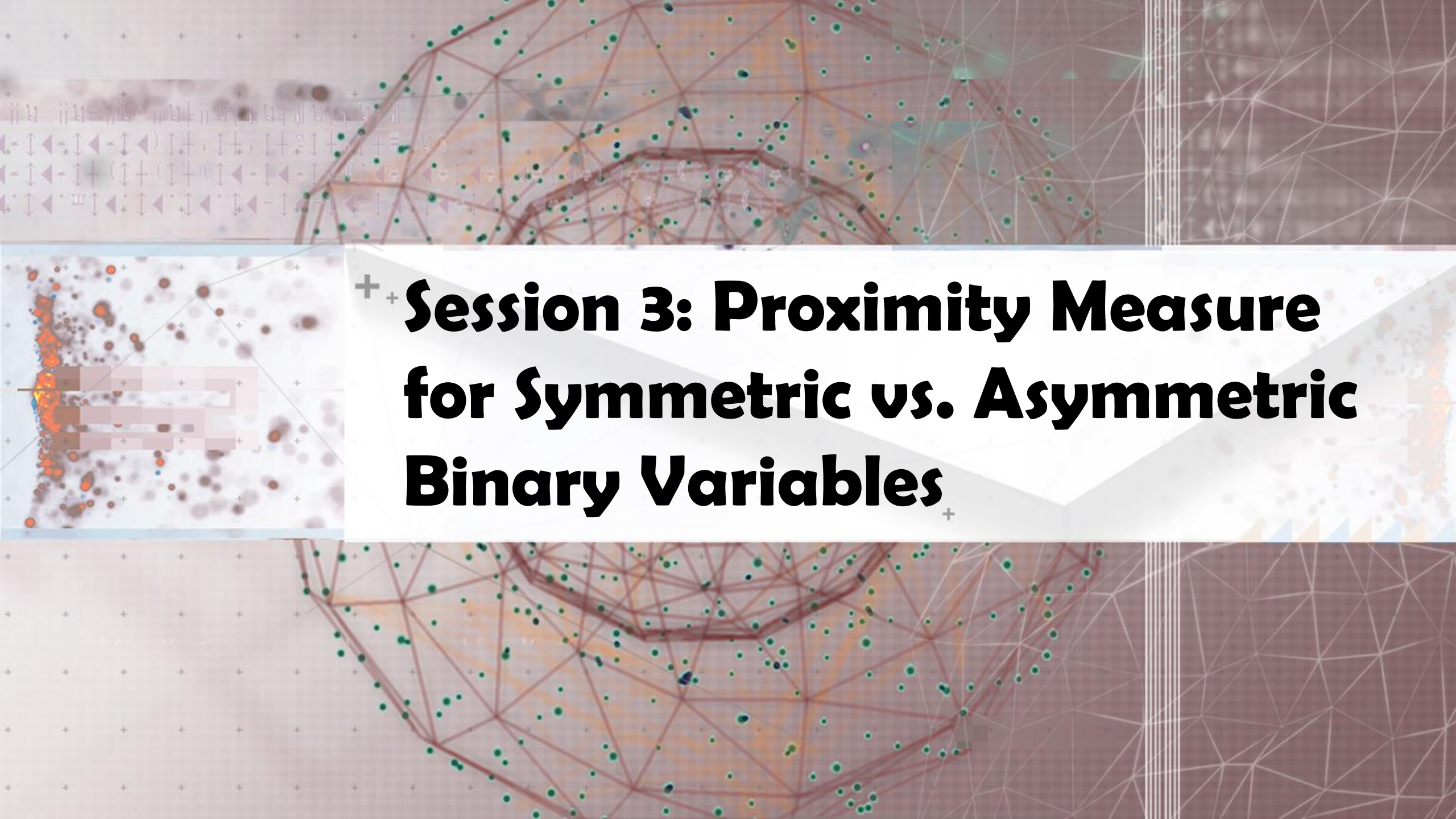
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0



+ Session 3: Proximity Measure for Symmetric vs. Asymmetric Binary Variables

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
sum		$q + s$	$r + t$	p

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard coefficient is the same as “coherence”: (a concept discussed in Pattern Discovery)

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

Example: Dissimilarity between Asymmetric Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

Distance:
$$d(i, j) = \frac{r + s}{q + r + s}$$

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

		Mary		
		1	0	Σ_{row}
Jack	1	2	0	2
	0	1	3	4
	Σ_{col}	3	3	6

		Jim		
		1	0	Σ_{row}
Jack	1	1	1	2
	0	1	3	4
	Σ_{col}	2	4	6

		Mary		
		1	0	Σ_{row}
Jim	1	1	1	2
	0	2	2	4
	Σ_{col}	3	3	6



Session 4: Distance between Categorical Attributes, Ordinal Attributes, and Mixed Types

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes

- Example: Color (red, yellow, blue, green), profession, etc.

- Method 1: Simple matching

- m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes

- Creating a new binary attribute for each of the M nominal states

Ordinal Variables

- ❑ An ordinal variable can be discrete or continuous
- ❑ Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- ❑ Can be treated like interval-scaled
 - ❑ Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, \dots, M_f\}$
 - ❑ Map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
 - ❑ Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
 - ❑ Then distance: $d(\text{freshman}, \text{senior}) = 1$, $d(\text{junior}, \text{senior}) = 1/3$
 - ❑ Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A dataset may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

$$d(i, j) = \frac{\sum_{f=1}^p w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p w_{ij}^{(f)}}$$

- If f is numeric: Use the normalized distance
- If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
- If f is ordinal
 - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
 - Treat z_{if} as interval-scaled



Session 5: Proximity Measure between Two Vectors: Cosine Similarity

Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

<i>Document</i>	<i>team</i>	<i>coach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$$

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity:

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$$

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \quad d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

- First, calculate vector dot product

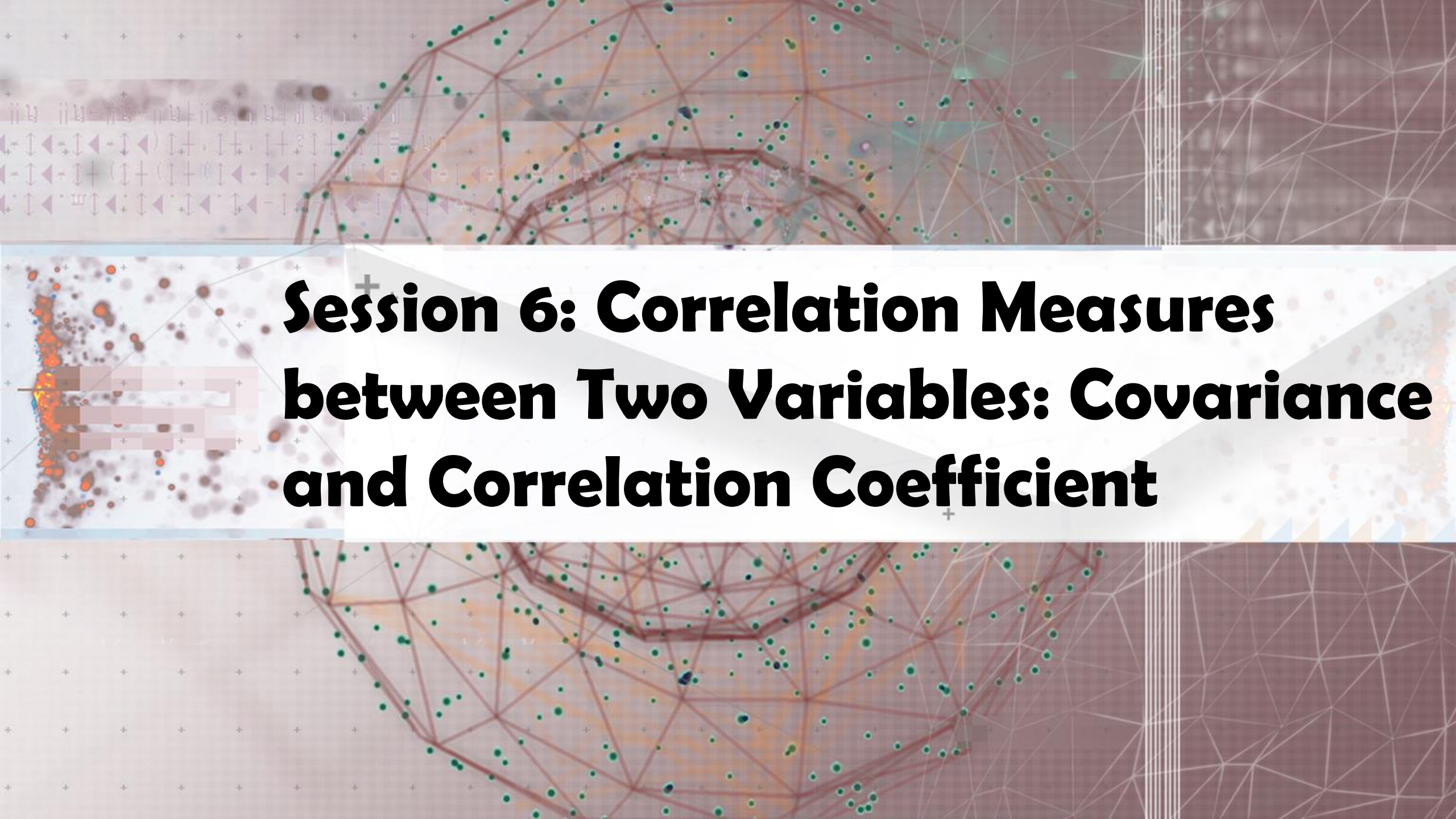
$$d_1 \bullet d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

- Then, calculate $\|d_1\|$ and $\|d_2\|$

$$\|d_1\| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\|d_2\| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

- Calculate cosine similarity: $\cos(d_1, d_2) = 25 / (6.481 \times 4.12) = 0.94$

The background features a complex network of thin, light-colored lines forming a web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. A prominent, thicker red line or path winds through the center of the image. The overall aesthetic is technical and data-driven.

Session 6: Correlation Measures between Two Variables: Covariance and Correlation Coefficient

Variance for Single Variable

- The variance of a random variable X provides a measure of how much the value of X deviates from the mean or expected value of X :

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- where σ^2 is the variance of X , σ is called *standard deviation*

μ is the mean, and $\mu = E[X]$ is the expected value of X

- That is, variance is the expected value of the square deviation from the mean

- It can also be written as: $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$

- Sample variance is the average squared deviation of the data value x_i from the sample mean

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$



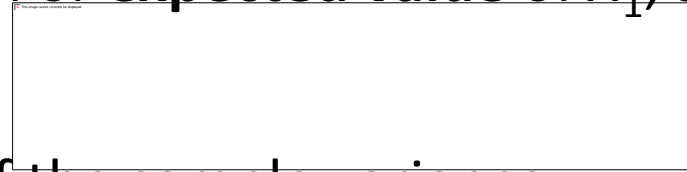
Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1 ; similarly for μ_2

- Sample covariance between X_1 and X_2 :



- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 = \hat{\sigma}_1^2$$

- **Positive covariance:** If $\sigma_{12} > 0$

- **Negative covariance:** If $\sigma_{12} < 0$

- **Independence:** If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true

- Some pairs of random variables may have a covariance 0 but are not independent
- Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

□ Suppose two stocks X_1 and X_2 have the following values in one week:

□ $(2, 5), (3, 8), (5, 10), (4, 11), (6, 14)$

□ Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

□ Covariance formula

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

□ Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$

□ $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$

□ $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$

□ $\sigma_{12} = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$

□ Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

- ❑ **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

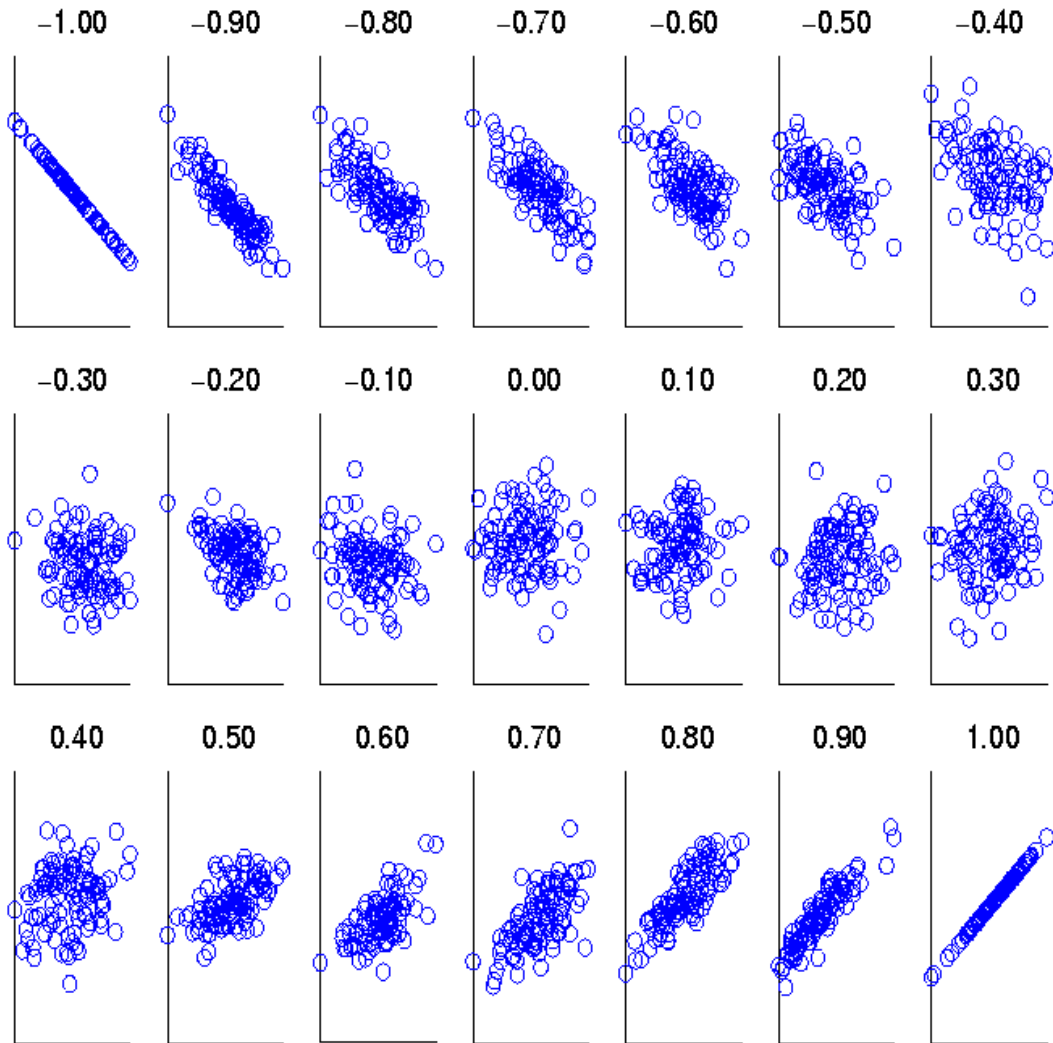
$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- ❑ **Sample correlation** for two attributes X_1 and X_2 : $\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$

where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2 , σ_1 and σ_2 are the respective standard deviation of X_1 and X_2

- ❑ If $\rho_{12} > 0$: A and B are positively correlated (X_1 's values increase as X_2 's)
 - ▢ The higher, the stronger correlation
- ❑ If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- ❑ If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: $[-1, 1]$
- A set of scatter plots shows sets of points and their correlation coefficients changing from -1 to 1

Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 X 2 covariance matrix as

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & X_2 - \mu_2 \end{pmatrix}\right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}$$

- In general, considering d numerical attributes, we have,

$$D = \begin{matrix} & \begin{matrix} X_1 & X_2 & \dots & X_d \end{matrix} \\ \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix} & \Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nd}^2 \end{pmatrix} \end{matrix}$$

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle geometric pattern. Above and below this banner are sections with a dark, reddish-brown background, overlaid with a network of thin, light-colored lines and small, colorful dots (green, blue, yellow). On the left side, there is a vertical strip with a light blue background, containing a grid of small, colorful dots (orange, red, blue, purple) and a larger, semi-transparent rectangular area with a grid pattern. The overall aesthetic is modern and technical, suggesting a focus on data or science.

Session 7: Summary

Summary: Similarity Measures for Cluster Analysis

- ❑ Basic Concept: Measuring Similarity between Objects
- ❑ Distance on Numeric Data: Minkowski Distance
- ❑ Proximity Measure for Symmetric vs. Asymmetric Binary Variables
- ❑ Distance between Categorical Attributes, Ordinal Attributes, and Mixed Types
- ❑ Proximity Measure between Two Vectors: Cosine Similarity
- ❑ Correlation Measures between Two Variables: Covariance and Correlation Coefficient
- ❑ Summary

Recommended Readings

- ❑ L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990
- ❑ Mohammed J. Zaki and Wagner Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- ❑ Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- ❑ Charu Aggarwal and Chandran K. Reddy (eds.). Data Clustering: Algorithms and Applications. CRC Press, 2014