# Cluster Analysis in Data Mining: Course Overview

# What Is Cluster Analysis?

- When flying over a city, one can easily identify fields, forests, commercial areas, and residential areas based on their features, without anyone's explicit "training"—This is the power of cluster analysis
- This course will systematically study cluster analysis methods and help answer the following:
  - What are the different proximity measures for effective clustering?
  - Can we cluster a massive number of data points efficiently?
  - Can we find clusters of arbitrary shape? At multiple levels of granularity?
  - How can we judge the quality of the clusters discovered by our system?

# The Value of Cluster Analysis

- What is the value of cluster analysis?
  - Cluster analysis helps you partition massive data into groups based on its features
  - Cluster analysis will often help subsequent data mining processes such as pattern discovery, classification, and outlier analysis
- What roles does cluster analysis play in the Data Mining Specialization?
  - You will learn various scalable methods to find clusters from massive data
  - You will learn how to mine different kinds of clusters effectively
  - You will also learn how to evaluate the quality of the clusters you find
  - Cluster analysis will help with classification, outlier analysis, and other data mining tasks

# Broad Applications of Cluster Analysis

- Data summarization, compression, and reduction
  - Examples: Image processing or vector quantization
- Collaborative filtering, recommendation systems, or customer segmentation
  - Finding like-minded users or similar products
- Dynamic trend detection
  - Clustering stream data and detecting trends and patterns
- Multimedia data analysis, biological data analysis, and social network analysis
  - Examples: Clustering video/audio clips or gene/protein sequences
- A key intermediate step for other data mining tasks
  - Generating a compact summary of data for classification, pattern discovery, and hypothesis generation and testing
  - Outlier detection: Outliers are those "far away" from any cluster

# Major Reference Readings for the Module

- **Textbook**
  - Han, J., Pei, J. & Tong, H. (2022). *Data mining: Concepts and techniques (4th ed.)*. Morgan Kaufmann
  - **Chapters most related to the course**
    - Chapter 2: Data, Measurements, and Data Preprocessing (Section 2.3: Similarity and Distance Measures)
    - Chapter 8: Cluster Analysis
    - Chapter 9: Advanced Cluster Analysis

- Other references will be listed at the end of each lecture video

# Cluster Analysis: An Introduction

# Cluster Analysis: An Introduction

- What Is Cluster Analysis?
- Applications of Cluster Analysis
- Cluster Analysis: Requirements and Challenges
- Cluster Analysis: A Multi-Dimensional Categorization
- An Overview of Typical Clustering Methodologies
- An Overview of Clustering Different Types of Data
- An Overview of User Insights and Clustering
- Summary

# Session 1: What Is Cluster Analysis?

# What Is Cluster Analysis?

- **What is a cluster?**
  - A cluster is a collection of data objects which are
    - Similar (or related) to one another within the same group (i.e., cluster)
    - Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)
- **Cluster analysis** (or ***clustering, data segmentation***, …)
  - Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible
- Cluster analysis is **unsupervised learning** (i.e., no predefined classes)
  - This contrasts with *classification* (i.e., *supervised learning*)
- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms

# Session 2: Applications of Cluster Analysis

# Cluster Analysis: Applications

- A key intermediate step for other data mining tasks
  - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
  - Outlier detection: Outliers—those "far away" from any cluster
- Data summarization, compression, and reduction
  - Ex. Image processing: Vector quantization
- Collaborative filtering, recommendation systems, or customer segmentation
  - Find like-minded users or similar products
- Dynamic trend detection
  - Clustering stream data and detecting trends and patterns
- Multimedia data analysis, biological data analysis and social network analysis
  - Ex. Clustering images or video/audio clips, gene/protein sequences, etc.

# Session 3: Cluster Analysis: Requirements and Challenges

# Considerations for Cluster Analysis

- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Similarity measure**
  - Distance-based (e.g., Euclidean, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- **Clustering space**
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

# Requirements and Challenges

- **Quality**
  - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
  - Discovery of clusters with arbitrary shape
  - Ability to deal with noisy data
- **Scalability**
  - Clustering all the data instead of only on samples
  - High dimensionality
  - Incremental or stream clustering and insensitivity to input order
- **Constraint-based clustering**
  - User-given preferences or constraints; domain knowledge; user queries
- **Interpretability and usability**

# Session 4: Cluster Analysis: A Multi-Dimensional Categorization

# Cluster Analysis: A Multi-Dimensional Categorization

- **Technique-Centered**
  - Distance-based methods
  - Density-based and grid-based methods
  - Probabilistic and generative models
  - Leveraging dimensionality reduction methods
  - High-dimensional clustering
  - Scalable techniques for cluster analysis
- **Data Type-Centered**
  - Clustering numerical data, categorical data, text data, multimedia data, time-series data, sequences, stream data, networked data, uncertain data
- **Additional Insight-Centered**
  - Visual insights, semi-supervised, ensemble-based, validation-based

# Session 5: An Overview of Typical Clustering Methodologies

# Typical Clustering Methodologies (I)

- **Distance-based methods**
  - Partitioning algorithms: K-Means, K-Medians, K-Medoids
  - Hierarchical algorithms: Agglomerative vs. divisive methods
- **Density-based and grid-based methods**
  - Density-based: Data space is explored at a high-level of granularity and then post-processing to put together dense regions into an arbitrary shape
  - Grid-based: Individual regions of the data space are formed into a grid-like structure
- **Probabilistic and generative models:** Modeling data from a generative process
  - Assume a specific form of the generative model (e.g., mixture of Gaussians)
  - Model parameters are estimated with the Expectation-Maximization (EM) algorithm (using the available dataset, for a maximum likelihood fit)
  - Then estimate the generative probability of the underlying data points

# Typical Clustering Methodologies (II)

- **High-dimensional clustering**
  - Subspace clustering: Find clusters on various subspaces
    - Bottom-up, top-down, correlation-based methods vs. δ-cluster methods
  - Dimensionality reduction: A vertical form (i.e., columns) of clustering
      - Columns are clustered; may cluster rows and columns together (co-clustering)
    - Probabilistic latent semantic indexing (PLSI) then LDA: Topic modeling of text data
      - A cluster (i.e., topic) is associated with a set of words (i.e., dimensions) and a set of documents (i.e., rows) simultaneously
    - Nonnegative matrix factorization (NMF) (as one kind of co-clustering)
      - A nonnegative matrix A (e.g., word frequencies in documents) can be approximately factorized two non-negative low rank matrices U and V
    - Spectral clustering: Use the *spectrum* of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions

# Session 6: An Overview of Clustering Different Types of Data

# Clustering Different Types of Data (I)

- **Numerical data**
  - Most earliest clustering algorithms were designed for numerical data
- **Categorical data** (including binary data)
  - Discrete data, no natural order (e.g., sex, race, zip-code, and market-basket)
- **Text data**: Popular in social media, Web, and social networks
  - Features: High-dimensional, sparse, value corresponding to word frequencies
  - Methods: Combination of k-means and agglomerative; topic modeling; co-clustering
- **Multimedia data**: Image, audio, video (e.g., on Flickr, YouTube)
  - Multi-modal (often combined with text data)
  - Contextual: Containing both behavioral and contextual attributes
    - Images: Position of a pixel represents its context, value represents its behavior
    - Video and music data: Temporal ordering of records represents its meaning

# Clustering Different Types of Data (II)

- **Time-series data**: Sensor data, stock markets, temporal tracking, forecasting, etc.
  - Data are temporally dependent
  - Time: contextual attribute; data value: behavioral attribute
  - Correlation-based online analysis (e.g., online clustering of stock to find stock tickers)
  - Shape-based offline analysis (e.g., cluster ECG based on overall shapes)
- **Sequence data**:  Weblogs, biological sequences, system command sequences
  - Contextual attribute: Placement (rather than time)
  - Similarity functions: Hamming distance, edit distance, longest common subsequence
  - Sequence clustering: Suffix tree; generative model (e.g., Hidden Markov Model)
- **Stream data**:
  - Real-time, evolution and concept drift, single pass algorithm
  - Create efficient intermediate representation, e.g., micro-clustering

# Clustering Different Types of Data (III)

- **Graphs and homogeneous networks**
  - Every kind of data can be represented as a graph with similarity values as edges
  - Methods: Generative models; combinatorial algorithms (graph cuts); spectral methods; non-negative matrix factorization methods
- **Heterogeneous networks**
  - A network consists of multiple typed nodes and edges (e.g., bibliographical data)
  - Clustering different typed nodes/links together (e.g., NetClus)
- **Uncertain data**: Noise, approximate values, multiple possible values
  - Incorporation of probabilistic information will improve the quality of clustering
- **Big data**: Model systems may store and process very big data (e.g., weblogs)
  - Ex. Google's MapReduce framework
    - Use *Map* function to distribute the computation across different machines
    - Use R*educe* function to aggregate results obtained from the Map step

# Session 7: An Overview of User Insights and Clustering

# User Insights and Interactions in Clustering

- **Visual insights**: One picture is worth a thousand words
  - Human eyes: High-speed processor linking with a rich knowledge-base
  - A human can provide intuitive insights; HD-eye: visualizing HD clusters
- **Semi-supervised insights**: Passing user's insights or intention to system
  - User-seeding: A user provides a number of labeled examples, approximately representing categories of interest
- **Multi-view and ensemble-based insights**
  - Multi-view clustering: Multiple clusterings represent different perspectives
  - Multiple clustering results can be ensembled to provide a more robust solution
- **Validation-based insights**: Evaluation of the quality of clusters generated
  - May use case studies, specific measures, or pre-existing labels

# Summary

# Summary: Cluster Analysis—An Introduction

- What Is Cluster Analysis?
- Applications of Cluster Analysis
- Cluster Analysis: Requirements and Challenges
- Cluster Analysis: A Multi-Dimensional Categorization
- An Overview of Typical Clustering Methodologies
- An Overview of Clustering Different Types of Data
- An Overview of User Insights and Clustering

# Recommended Readings

- Major Reference Books on Cluster Analysis
  - Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011 (Chapters 10 & 11)
  - Charu Aggarwal and Chandran K. Reddy (eds.). Data Clustering: Algorithms and Applications. CRC Press, 2014
  - Mohammed J. Zaki and Wagner Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- Reference paper for this lecture
  - Charu Aggarwal. An Introduction to Clustering Analysis. *in* Aggarwal and Reddy (eds.). Data Clustering: Algorithms and Applications (Chapter 1). CRC Press, 2014

# Similarity Measures for Cluster Analysis

# 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

# 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, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

# Session 2: Distance on Numeric Data: Minkowski Distance

# Data Matrix and Dissimilarity Matrix

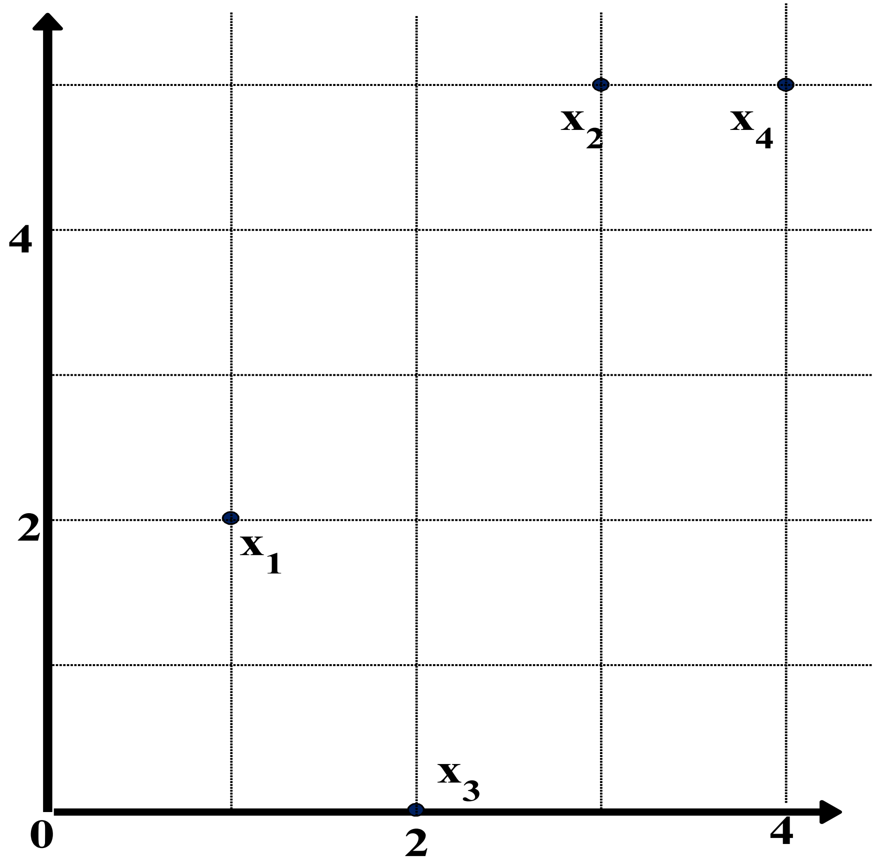
- Data matrix
  - A data matrix of n data points with *l* dimensions

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

- 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

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

  where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ 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 ≠ j, and d(i, i) = 0 (Positivity)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ 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

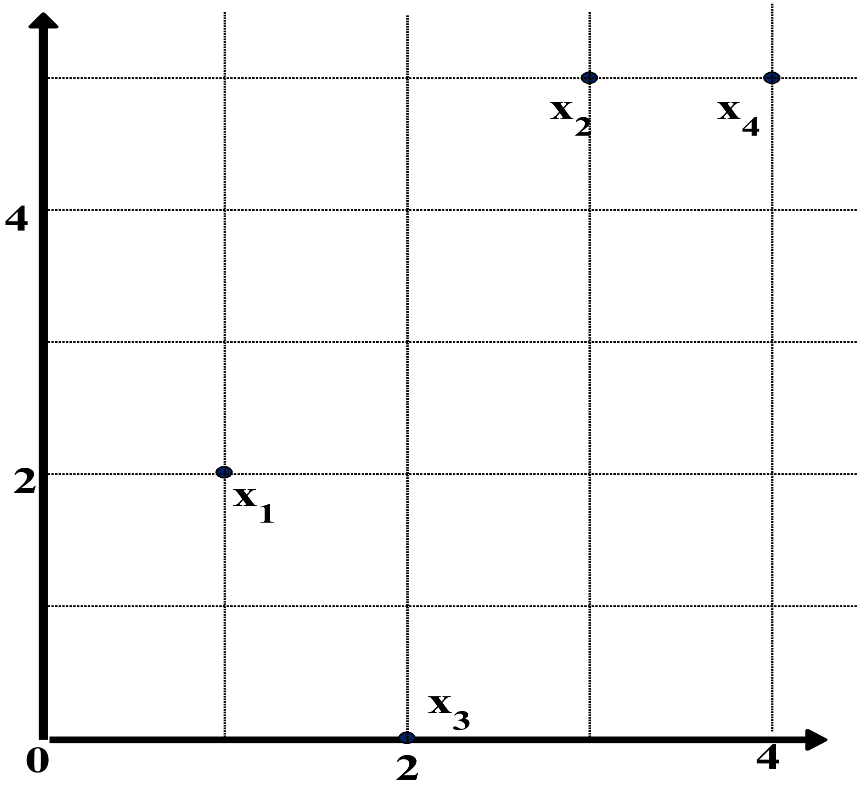
$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- $p \rightarrow \infty$: ($L_{max}$ norm, $L_\infty$ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

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

# 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 ($L_\infty$)

| $L_\infty$ | 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 $j$

| Object $i$ | 1 | 0 | sum |
|---|---|---|---|
| 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

| Jack | 1 | 0 | $\sum_{row}$ |
|---|---|---|---|
| 1 | 2 | 0 | 2 |
| 0 | 1 | 3 | 4 |
| $\sum_{col}$ | 3 | 3 | 6 |

Jim

| Jack | 1 | 0 | $\sum_{row}$ |
|---|---|---|---|
| 1 | 1 | 1 | 2 |
| 0 | 1 | 3 | 4 |
| $\sum_{col}$ | 2 | 4 | 6 |

Mary

| Jim | 1 | 0 | $\sum_{row}$ |
|---|---|---|---|
| 1 | 1 | 1 | 2 |
| 0 | 2 | 2 | 4 |
| $\sum_{col}$ | 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: $r_{if} \in \{1, ..., M_f\}$
  - 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(freshman, senior) = 1, d(junior, 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

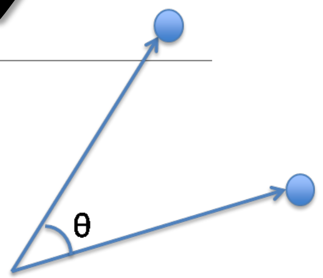
| *Document* | *team* | *coach* | *hockey* | *baseball* | *soccer* | *penalty* | *score* | *win* | *loss* | *season* |
|---|---|---|---|---|---|---|---|---|---|---|
| Document1 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| Document2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| Document3 | 0 | 7 | 0 | 2 | 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: $$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$ $$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

  where • 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) $d_2$ = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

  - First, calculate vector dot product

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

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

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

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

# 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:*

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

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

    μ is the mean, and μ = 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: $\sigma^2 = \mathrm{var}(X) = E[(\mathrm{X}-\mu)^2] = E[\mathrm{X}^2]-\mu^2 = E[\mathrm{X}^2]-[E(x)]^2$

- Sample variance is the average squared deviation of the data value $x_i$ from the sample mean $\hat{\mu}$

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

# 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 $\mu_2$

- Sample covariance between $X_1$ and $X_2$: $\hat{\sigma}_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$
- Sample covariance is a generalization of the sample variance:

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

- **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: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
  - $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14)/5 - 4 \times 9.6 = 4$
- 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

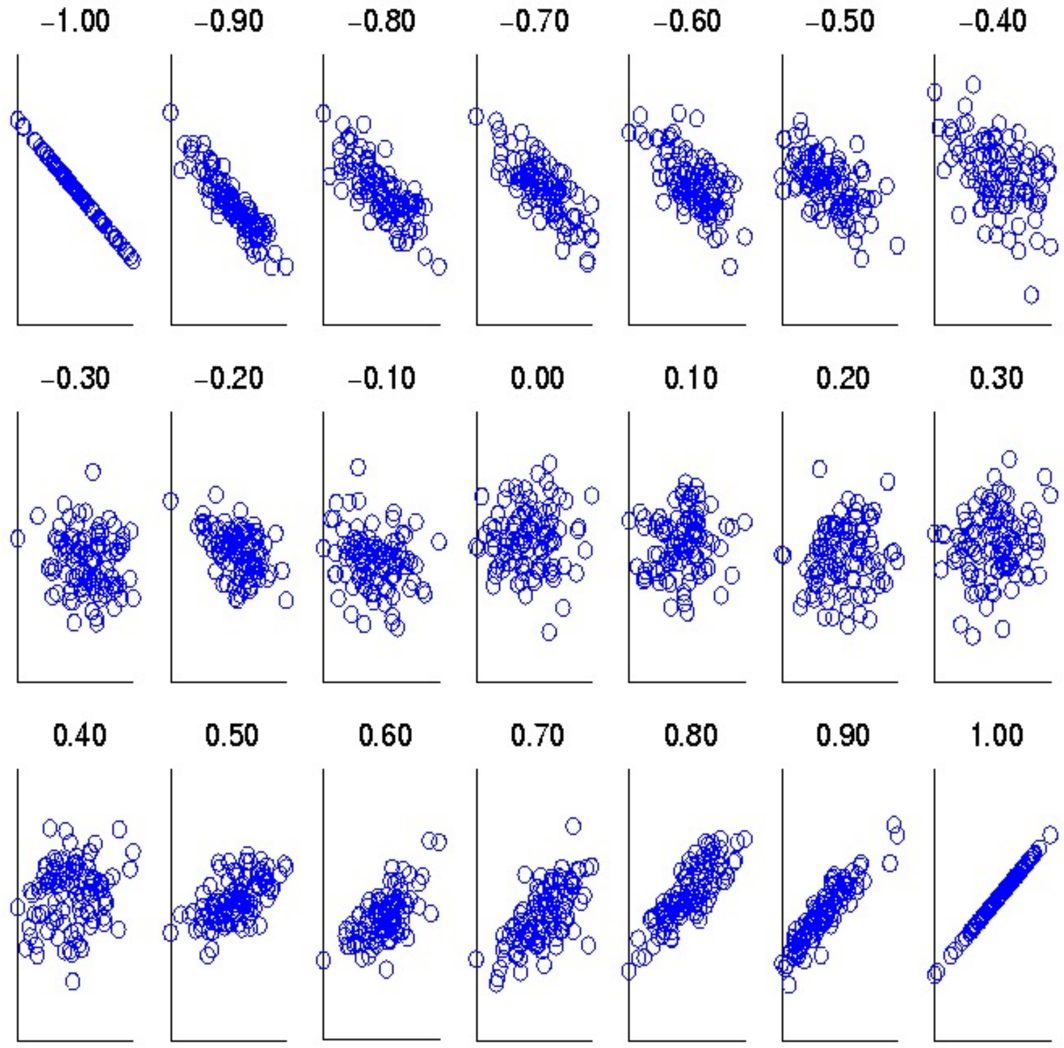
$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^{\ 2} \sigma_2^{\ 2}}}$$

- **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, $\mu_1$ and $\mu_2$ are the respective means of $X_1$ and $X_2$ , $\sigma_1$ and $\sigma_2$ 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

$$\Sigma = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = E[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}(X_1-\mu_1 \quad X_2-\mu_2)]$$

$$= \begin{pmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Generalizing it to *d* dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \mathbf{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

# 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

# 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