Data Preprocessing (part 2)

Class Project: Any questions?

Class “homework” (recap from previous lecture)

Please research one-sentence definitions of different types of social relationships:

1. Kinship
2. Friendship
3. Acquaintances/Social Contacts
4. Familiar Strangers
5. Encounters
6. Strangers

Dimensionality Reduction

Purpose

1. Avoid curse of dimensionality
2. Reduce amount of time and memory required by data mining algorithms
3. Allow data to be more easily visualized
4. May help to eliminate irrelevant features or reduce noise

Techniques

1. Principal Component Analysis (PCA)
2. Singular Value Decomposition
3. Others: supervised and non-linear techniques

Techniques for Feature Selection

1. Brute-force approach
   Try all possible feature subsets as input to data mining algorithm

2. Embedded approaches
   Feature selection occurs naturally as part of the data mining algorithm

3. Filter approaches
   Features are selected before data mining algorithm is run

4. Wrapper approaches
   Use the data mining algorithm as a black-box to find best subset of attributes

Discretization

Discretization is the process of converting a continuous attribute into an ordinal attribute

A potentially infinite number of values are mapped into a small number of categories

Discretization is commonly used in classification

Many classification algorithms work best if both the independent and dependent variables have only a few values

Discretization

Discretization means breaking continuous numbers into smaller groups or categories. It helps in making data easier to understand and use in machine learning. Here are some ways to do it:

Discretization

Equal Width Bins (N intervals with the same width)

1. Divide the data into **equal-sized groups** based on range.
2. Formula: **W = (max value - min value) / number of groups**

Example: If test scores range from **0 to 100**, and we want **5 groups**, we divide it into:

- 0-20
- 21-40
- 41-60
- 61-80
- 81-100

Problem: This method doesn't work well if there are very high or very low values (**outliers**).

Discretization

Equal Frequency Bins (N intervals with the same number of data points)

Each group has about **the same number of values** instead of the same width.

Example: If you have 100 students, and you want **4 groups**, each group should have **25 students** based on their scores.

Better for: Data that is unevenly spread (works better with outliers).

Discretization

Clustering-Based Bins

Groups data points that are naturally **similar to each other**.

Example: Instead of setting fixed intervals, the computer finds **clusters** in the data, like small groups of students who scored similarly.

Best for: Complicated data that doesn’t fit into simple intervals.

Discretization

When we have a set of numbers (like student grades, heights, or temperatures), sometimes we need to **change them** in a way that makes comparisons easier. This is called **attribute transformation**.

1. Attribute Transformation (Changing Values in a Useful Way)

A **function** is used to change all the numbers in a set to new values.

Examples:

- $\mathbf{x^2}$ → Squaring the numbers (e.g., 3 → 9, 4 → 16)
- **log(x)** → Taking the logarithm (e.g., 10 → 1, 100 → 2)
- $\mathbf{|x|}$ → Absolute value (e.g., -5 → 5)

Discretization

2. Normalization (Making Different Data Comparable)

Normalization is a way to **adjust numbers** so they fit within the same scale.

Example:

Min-Max Normalization Formula

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Key Terms:

- X_{norm} → The scaled value after normalization.
- X → The original data value.
- X_{min}, X_{max} → The smallest and largest values in the dataset.
- This formula **scales data** between 0 and 1 to improve performance in machine learning.

Discretization

3. Standardization (Making Data Follow a Common Pattern)

It **subtracts the average (mean)** and **divides by how spread out the data is (standard deviation)**

Example:

Z-Score (Standard Score) Formula

$$Z = \frac{X - \mu}{\sigma}$$

Key Terms:

- Z → Z-score (how far a value is from the mean in standard deviations).
- X → The original data value.
- μ → The mean (average) of the dataset.
- σ → The standard deviation (spread of the data).
- This formula **standardizes data** to compare values from different datasets.

Euclidean Distance

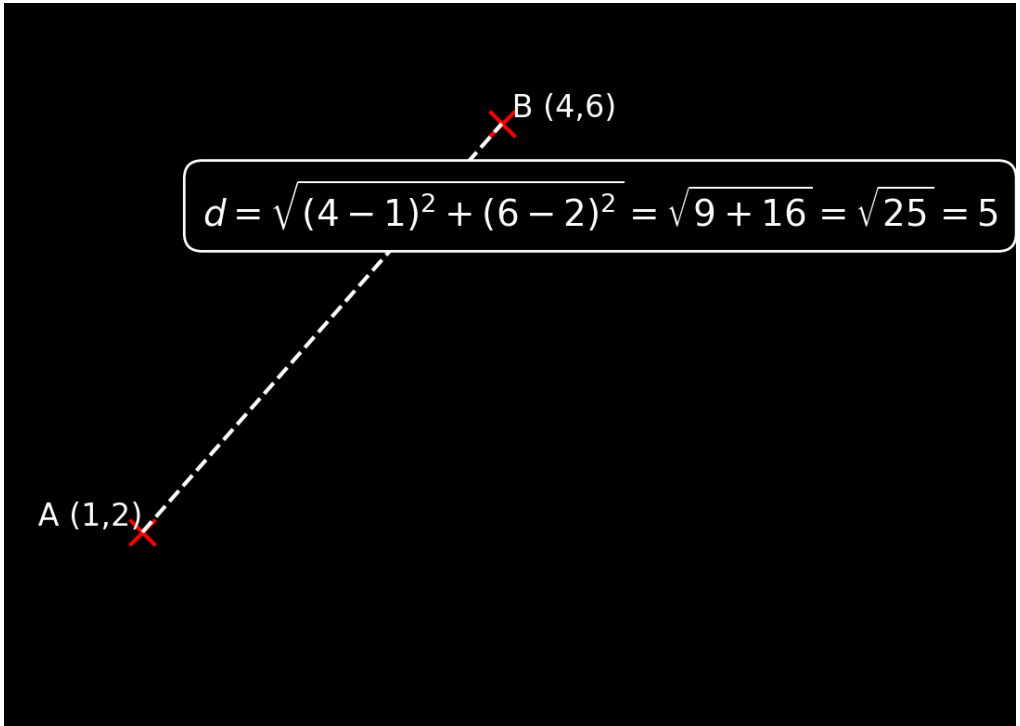
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

d → The Euclidean distance (the straight-line distance between two points).

(x1,y1) → Coordinates of the first point.

(x2,y2) → Coordinates of the second point.

Euclidean Distance



Minkowski Distance

Minkowski Distance is a generalization of Euclidean Distance

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

d → Minkowski Distance (the calculated distance between two points).
n → Number of dimensions (for 2D, n=2n = 2n=2, for 3D, n=3n = 3n=3, etc.).
xi,y_i → Coordinates of the two points.
p → The **Minkowski Power Parameter**, which determines the type of distance:

- **p=1→ Manhattan Distance** (sum of absolute differences).
- **p=2→ Euclidean Distance** (straight-line distance).
- **p→∞ → Chebyshev Distance** (maximum coordinate difference).

Cosine similarity

Cosine Similarity measures how similar two vectors (documents) are.

$$\cos(\theta) = \frac{d_1 \cdot d_2}{||d_1|| \times ||d_2||}$$

Example:

$d_1 = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0]$

$d_2 = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$

$d_1 \cdot d_2 = 5$

$||d_1|| = 6.481,\ ||d_2|| = 2.245$

$$\cos(d_1, d_2) = \frac{5}{6.481 \times 2.245} = 0.315$$

Cosine similarity (if you insist…)

Given Vectors (Representing Two Documents):

$$d_1 = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0]$$

$$d_2 = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

Step 1: Compute the Dot Product

Multiply corresponding values and add them up:

$$(3 \times 1) + (2 \times 0) + (0 \times 0) + (5 \times 0) + (0 \times 0) + (0 \times 0) + (0 \times 0) + (2 \times 1) + (0 \times 0) + (0 \times 2)$$

$$= 3 + 0 + 0 + 0 + 0 + 0 + 0 + 2 + 0 + 0 = 5$$

Step 2: Compute the Norm of Each Vector

For d_1:

$$||d_1|| = \sqrt{(3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)}$$

$$= \sqrt{9 + 4 + 0 + 25 + 0 + 0 + 0 + 4 + 0 + 0} = \sqrt{42} = 6.481$$

For d_2:

$$||d_2|| = \sqrt{(1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)}$$

$$= \sqrt{1 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 0 + 4} = \sqrt{6} = 2.245$$

Step 3: Compute Cosine Similarity

$$\cos(d_1, d_2) = \frac{5}{6.481 \times 2.245}$$

Data correlation

We use a **measure** to check how **two sets of data** are related to each other.

Why is it useful?

Data correlation (why is it useful)

1. **Find relationships in data** – Helps understand how data points are connected.
2. **Useful in early analysis** – Helps in the **exploration phase** to understand patterns.
3. **Feature correlation** – If two features (data columns) are **strongly related**, we should **remove one** to make analysis easier.
4. **Better performance** – Removing unnecessary data **improves** the speed and accuracy of algorithms.

Data correlation

Example:

- Imagine you are analyzing students' performance in **Math and Physics**. If students who score high in **Math** also score high in **Physics**, these two subjects are **correlated**.
- If the correlation is very strong, we might **remove one** of these features to **simplify our analysis** without losing important information.

Pearson's Correlation Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}}$$

Key Terms:

- r → Pearson's correlation coefficient (value between -1 and 1).
- x_i, y_i → Individual data points.
- $\bar{x}, \bar{y}$ → Mean (average) of x and y values.
- $\sum$ → Summation (adding up all values).
- The formula measures **how strongly two variables are related**.

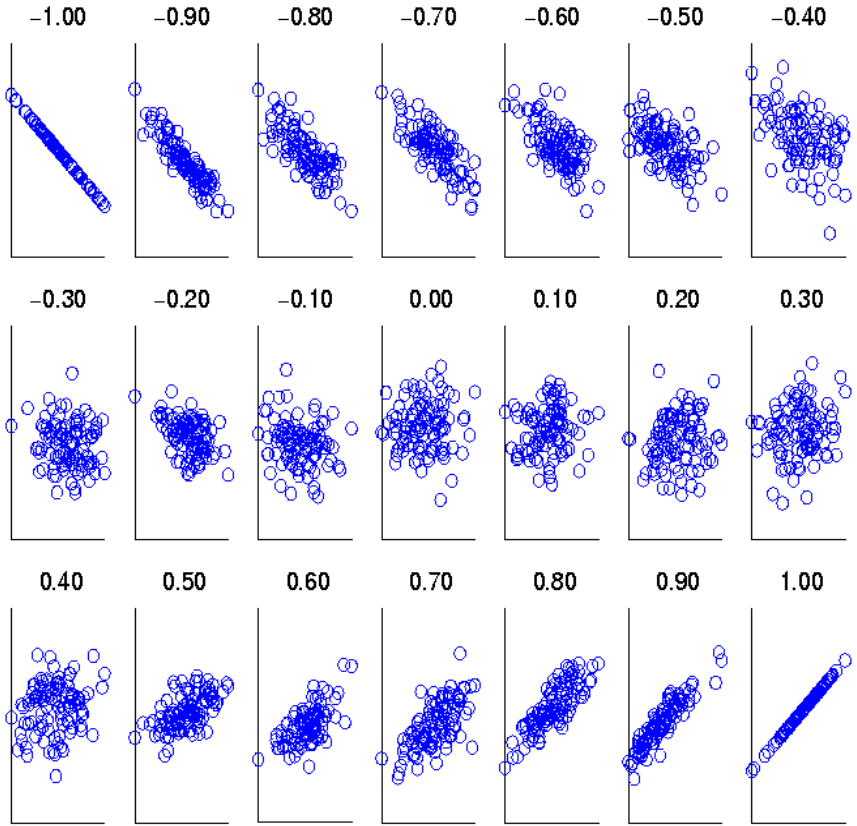
Example Calculation:

$$x = [1, 2, 3, 4, 5]$$
$$y = [2, 4, 5, 4, 5]$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}}$$

$$r = 0.775$$

Scatter plots showing the similarity from –1 to 1.



Drawback of Correlation

What Correlation Measures
Correlation tells us how **strongly two sets of numbers are related**. A correlation of **1** means they go up together, and a correlation of **-1** means when one goes up, the other goes down. A correlation of **0** means no clear relationship.

What's the problem?

Drawback of Correlation

We have two sets of numbers:
x-values: (-3, -2, -1, 0, 1, 2, 3)
y-values: (9, 4, 1, 0, 1, 4, 9)

If we plot these points, we see they follow a **U-shape** (parabola).

Data correlation

The Problem Here: When we calculate the correlation between **x and y**, the answer comes out as **0**. This suggests **no relationship** between x and y.

Why is This a Drawback? Even though the numbers clearly follow a U-shape, correlation **only measures linear relationships (straight-line trends)**. Since our data follows a curve, correlation **fails to capture** the real relationship.

Key Takeaway: Correlation is **not useful** when the relationship between variables is **non-linear** (curved). Other methods, like regression or scatter plots, help better understand these types of patterns.