Chapter 2: Getting to Know Your Data

CSE 435:Data Mining

Md. Atikuzzaman
Department of Computer Science & Engineering
Green University of Bangladesh
atik@cse.green.edu.bd

Table of Contents

Data Objects and Attributes

What is a Data Object?

A **data object** represents an entity. In a dataset, these are the **rows**.

- *Also known as:* samples, instances, data points, tuples.
- *Examples:* A customer in a sales database, a patient in medical records.

What is an Attribute?

An **attribute** is a feature or characteristic of a data object. In a dataset, these are the **columns**.

- *Also known as:* dimensions, features, variables.
- *Examples:* `customer_ID`, age, `product_price`.

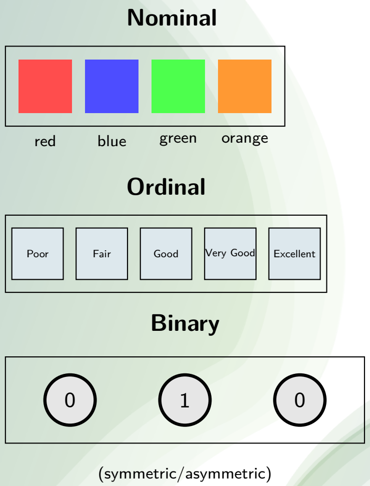
Customer Dataset

ID	age	price	segment	active
C001	22	799	Student	0
C002	35	1299	Regular	1
C003	29	499	Budget	0
C004	41	2199	Premium	0
C005	54	899	Regular	1
C006	31	1499	Premium	0
⋮	⋮	⋮	⋮	⋮

ID:	nominal (identifier)
age, price:	numeric (ratio)
segment:	nominal
active:	binary (asymmetric)

Qualitative (Categorical)

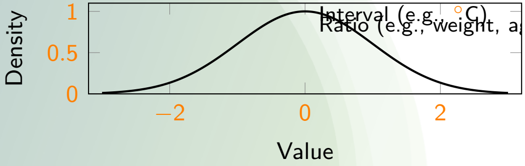
- **Nominal**: Distinct symbols, no order.
    - *Ex:* `eye_color`, `zip_codes`.
- **Ordinal**: Values have rank/sequence.
    - *Ex:* `drink_size` (small, medium, large).
- **Binary**: Two states (0/1).
    - *Symmetric:* both outcomes equally important.
    - *Asymmetric:* one outcome more important (e.g., positive test).



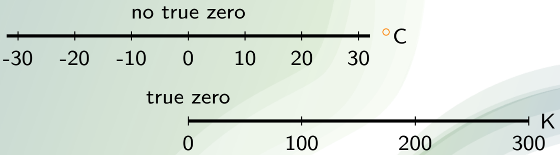
Quantitative (Numeric)

- **Interval-Scaled**: Ordered values, meaningful differences, *no true zero*.
  - *Ex:* Temperature in Celsius, calendar dates.
- **Ratio-Scaled**: True zero; ratios are meaningful.
  - *Ex:* `length`, `weight`, `salary`.

Numeric (distribution view)



Interval vs Ratio (number lines)

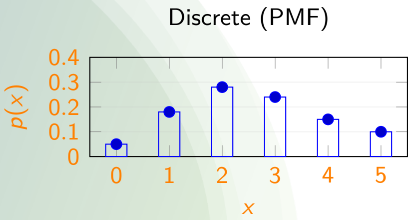


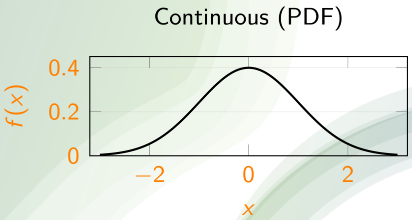
Discrete

- Takes values from a *countable* set (often integers).
- Examples: `num_items_bought`, `clicks`, `defects`.
- Described by a PMF $p(x) = \Pr(X = x)$ with $\sum_x p(x) = 1$.
- Typical summaries: frequency table, mode, entropy.

Continuous

- Takes values from an *uncountable* interval (real numbers).
- Examples: `temperature`, `weight`, `time`.
- Described by a PDF $f(x) \geq 0$ with $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
- $\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$; single points have prob. 0.





Measuring the Central Tendency — Mean

Mean (Average)

The sum of all values divided by the count of values. Sensitive to outliers.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- **Sample mean:** arithmetic average of n values.
- **Population mean:** average over all N population units.

Median

The middle value of a *sorted* dataset. Robust to outliers.

- If n is odd: the $\frac{n+1}{2}$-th value; if n is even: average of the two middle values.
- Median for grouped data:

$$\tilde{x} = L + \left(\frac{\frac{n}{2} - \mathrm{cf}}{f_m} \right) w$$

  - L : lower boundary of the *median class*
  - n : total frequency
  - cf : cumulative freq. *before* median class
  - f_m : freq. of median class
  - w : class width

Measuring the Central Tendency — Mode

Mode

The value that appears most frequently. A dataset can be **unimodal**, **bimodal**, or **trimodal**.

- **Grouped data (modal class interpolation):** estimate the peak inside the modal class.
- Useful empirical relation (for moderately skewed data):

$$\text{mean} - \text{mode} \approx 3\,(\text{mean} - \text{median})$$

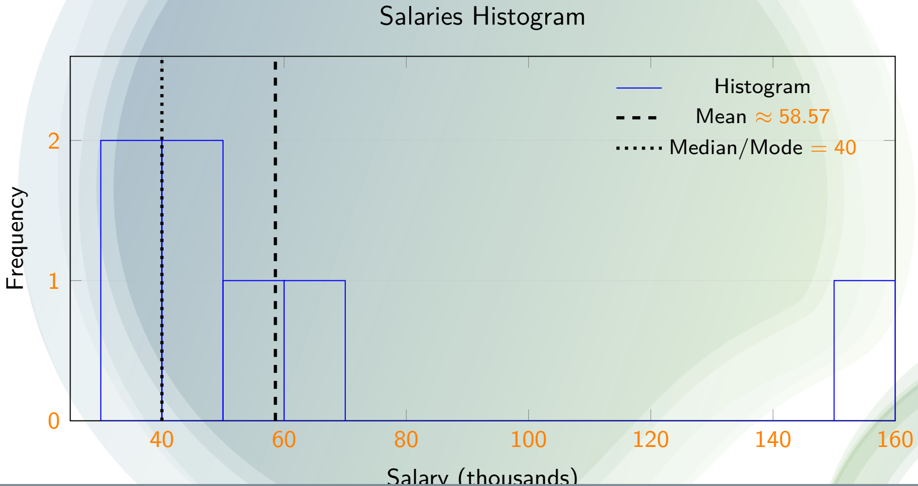
- Mode for grouped data:

$$\hat{m} = L + \left(\frac{d_1}{d_1 + d_2}\right) w$$

  - L : lower boundary of the *modal class*
  - w : class width
  - f_m : freq. of modal class
  - f_{m-1}, f_{m+1} : freqs. of adjacent classes
  - $d_1 = f_m - f_{m-1}, \quad d_2 = f_m - f_{m+1}$

Descriptive Statistics: Central Tendency (Salaries Example)

- Dataset (thousands): $\{30, 35, 40, 40, 55, 60, 150\}$
- Mean $\bar{x} = \frac{410}{7} \approx 58.57$; Median $= 40$; Mode $= 40$.



Measures of Data Dispersion

Range and Five-Number Summary

A summary of the distribution: **Minimum, Q1, Median (Q2), Q3, Maximum**.

Interquartile Range (IQR)

The range of the middle 50% of the data. Robust to outliers.

$$\mathrm{IQR} = Q_3 - Q_1$$

Measures of Data Dispersion

Variance (σ^2) and Standard Deviation (σ)

The variance is the average squared deviation from the mean; the standard deviation is its square root.

$$\underbrace{s^2}_{\text{sample variance}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \underbrace{s}_{\text{sample std. dev.}} = \sqrt{s^2}\,.$$

$$\underbrace{\sigma^2}_{\text{population variance}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2, \qquad \underbrace{\sigma}_{\text{population std. dev.}} = \sqrt{\sigma^2}\,.$$

Computational (shortcut) forms:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^{\,2}\right), \qquad \sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\,\mu^2\right).$$

Dispersion: Worked Example (Salaries)

Dataset (thousands): $\{30, 35, 40, 40, 55, 60, 150\}$

Five-Number Summary

- Min $=$ **30**
- Q_1 (25th pct) $=$ **35**
- Median (Q_2) $=$ **40**
- Q_3 (75th pct) $=$ **60**
- Max $=$ **150**

Interquartile Range

$$\mathrm{IQR} = Q_3 - Q_1 = 60 - 35 = \boxed{25}$$

Mean $\bar{x} = \frac{410}{7} \approx 58.57, \quad n = 7.$

Variance and Standard Deviation

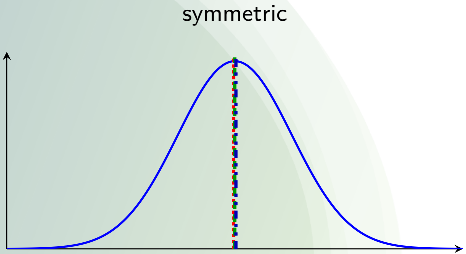
$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &\approx \frac{816.5 + 555.5 + 344.8 + 344.8 + 12.7 + 2.0 + 8359.2}{6} \\ &\approx 1739.25, \end{aligned}$$

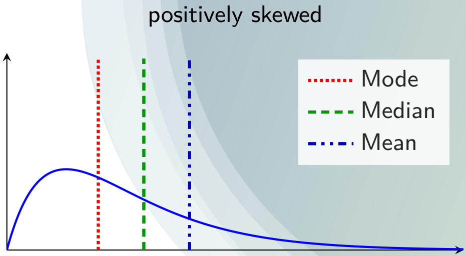
$$s = \sqrt{1739.25} \approx \boxed{41.7}.$$

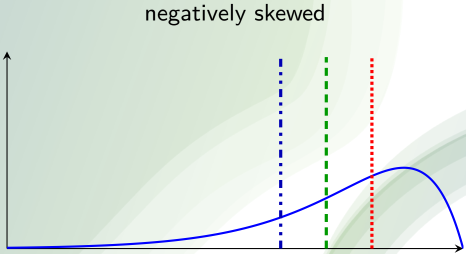
Note: The outlier **150** inflates s (high variability).

Median, Mean, Mode

- **Symmetric:** Mean $=$ Median $=$ Mode.
- **Positively skewed (right tail):** Mode $<$ Median $<$ Mean.
- **Negatively skewed (left tail):** Mean $<$ Median $<$ Mode.



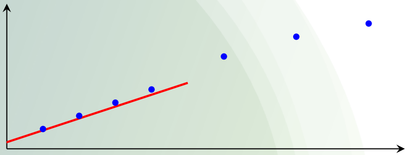


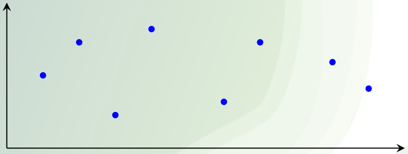


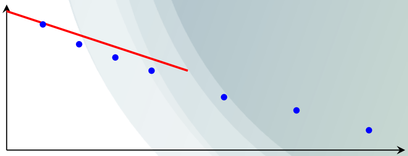
Correlation of Data

Correlation measures the linear relationship between two variables.

- **Positive Correlation:** As one variable increases, the other tends to increase. ($r \approx +1$)
- **Negative Correlation:** As one increases, the other decreases. ($r \approx -1$)
- **No Correlation:** No clear linear relationship. ($r \approx 0$)



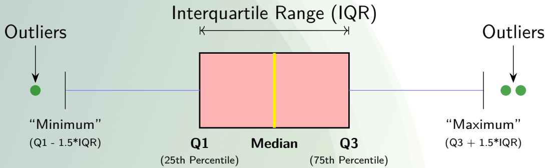




Boxplot Analysis

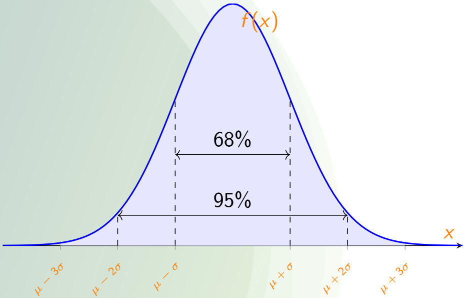
What a boxplot shows

- **Five-number summary:** min, Q_1, median (Q_2), Q_3, max.
- **Box** spans the interquartile range: $\mathrm{IQR} = Q_3 - Q_1$.
- **Whiskers:** extend to the most extreme points within $[\,Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}\,]$.
- **Outliers:** observations outside the whisker range (plotted as points).
- Quickly compares **center** (median), **spread** (IQR), and **skewness/outliers** across groups.



Key Characteristics

- The curve is bell-shaped and symmetric about the mean (μ).
- The **mean, median, and mode** are all equal and located at the center.
- The total area under the curve is equal to 1 (or 100%).
- The curve is described by its mean (μ) and standard deviation (σ).



Probability Density Function (PDF) The formula that defines the curve is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Why Visualize Data?

From Numbers to Insights

Data visualization turns raw data into charts and graphics so patterns, trends, and outliers jump out quickly.

- **Summarize** a dataset at a glance.
- **Reveal patterns/trends** that are hard to see in tables.
- **Spot outliers** and data quality issues.
- **Communicate** findings clearly to others.

"The greatest value of a picture is when it forces us to notice what we never expected to see." — John Tukey

Visualizing Distributions: The Histogram

Use Case

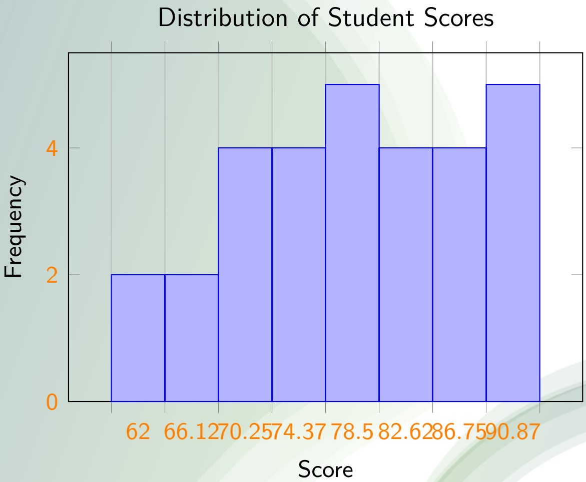
- Distribution of a **single continuous** variable.

Questions

- Symmetry? Skew? Unimodal/bimodal?

Examples

- Exam scores, ages, temperatures.



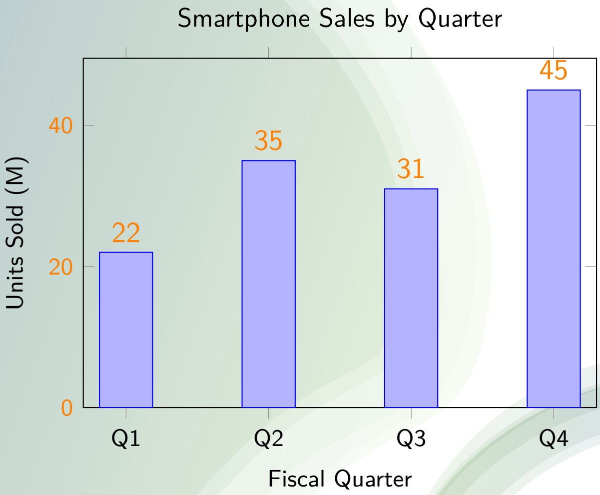
Comparing Categories: The Bar Chart

Use Case

- Compare a numeric value across **discrete categories**.

Examples

- Sales by quarter, population by country, feature importance.



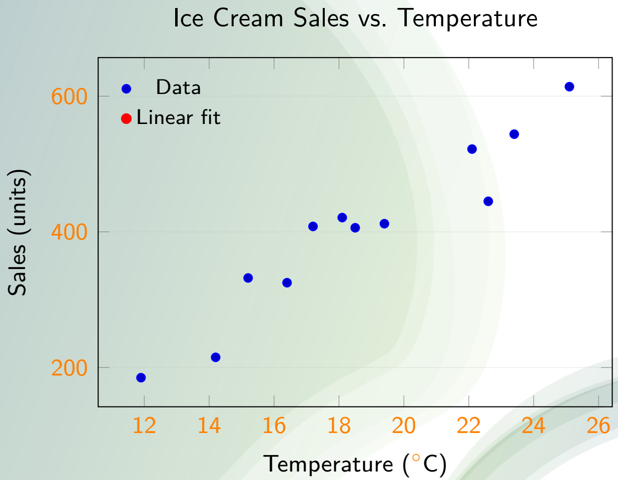
Exploring Relationships: The Scatter Plot

Use Case

- Relationship between **two continuous** variables.

Examples

- Ads spend vs. revenue; height vs. weight; temperature vs. sales.



Showing Proportions: The Pie Chart

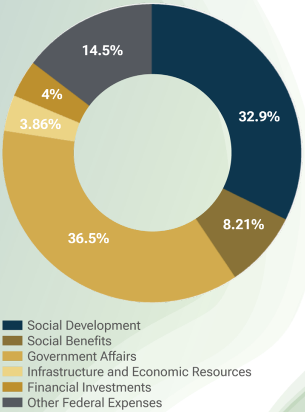
Use Case

- To show the **proportional composition** or **percentage share** of a whole.
- It's most effective with a small number of categories (usually 2-6).

Examples

- Market share of competing companies.
- Breakdown of a budget by department.
- Survey responses (e.g., "Agree", "Disagree", "Neutral").

Federal Budget



Similarity and Dissimilarity: The Basics

Core Concepts

- **Dissimilarity** (or **distance**) measures how *different* two data objects are. A low value means they are alike.
- **Similarity** measures how *alike* two data objects are. A high value means they are alike.

Relationship

Often, they are inverse concepts. A similarity measure $\text{sim}(\mathbf{x}, \mathbf{y})$ in the range $[0, 1]$ can be converted into a dissimilarity measure $d(\mathbf{x}, \mathbf{y})$ using: $d(\mathbf{x}, \mathbf{y}) = 1 - \text{sim}(\mathbf{x}, \mathbf{y})$

Why is this important?

- It's the foundation for many data mining tasks like **clustering**, **classification** (k-Nearest Neighbors), and **anomaly detection**.

Numeric Data: Minkowski Distance (L_p Norm)

General Formula

For two n-dimensional data points $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$, the Minkowski distance is: $d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^p\right)^{1/p}$

Three Common Cases:

- **p = 1: Manhattan Distance (L_1)**
  - The "city block" distance. You can only travel along grid lines. $d_1(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|$
- **p = 2: Euclidean Distance (L_2)**
  - The straight-line distance ("as the crow flies"). $d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$
- **p = ∞: Supremum Distance (L_∞)**
  - The maximum difference along any single dimension. $d_\infty(\mathbf{x}, \mathbf{y}) = \max_k |x_k - y_k|$

Minkowski Distance: Worked Example

Let's calculate the distance between two points in a 2D space: $\mathbf{x} = (2, 2), \quad \mathbf{y} = (5, 6)$

1. **Euclidean Distance ($p = 2$)**

$$\begin{aligned} d_2 &= \sqrt{(5-2)^2 + (6-2)^2} \\ &= \sqrt{3^2 + 4^2} \\ &= \sqrt{9 + 16} \\ &= \sqrt{25} = \mathbf{5} \end{aligned}$$

2. **Manhattan Distance ($p = 1$)**

$$\begin{aligned} d_1 &= |5-2| + |6-2| \\ &= 3 + 4 \\ &= \mathbf{7} \end{aligned}$$

3. **Supremum Distance ($p = \infty$)**

$$\begin{aligned} d_\infty &= \max(|5-2|, |6-2|) \\ &= \max(3, 4) = \mathbf{4} \end{aligned}$$

Numeric Data: Cosine Similarity

Concept

Measures the cosine of the angle (θ) between two non-zero vectors. It evaluates **orientation**, not magnitude, making it excellent for comparing documents or profiles.

Formula:

$$\text{sim}_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \sqrt{\sum_{k=1}^{n} y_k^2}}$$

Interpretation (for non-negative data):

- Result is in the range $[0, 1]$.
- $\text{sim}_{\cos} = 1 \implies$ Vectors point in the same direction (most similar).
- $\text{sim}_{\cos} = 0 \implies$ Vectors are orthogonal (unrelated).

Cosine Similarity: Worked Example

Consider two documents represented by term-frequency vectors: $\mathbf{x} = (3, 2)$ $\quad \mathbf{y} = (2, 3)$

Step 1: Calculate the dot product ($\mathbf{x} \cdot \mathbf{y}$) $\mathbf{x} \cdot \mathbf{y} = (3)(2) + (2)(3) = 6 + 6 = \mathbf{12}$

Step 2: Calculate the magnitude of each vector ($\|\mathbf{x}\|$ and $\|\mathbf{y}\|$)

$$\|\mathbf{x}\| = \sqrt{3^2 + 2^2} = \sqrt{9 + 4} = \sqrt{\mathbf{13}}$$

$$\|\mathbf{y}\| = \sqrt{2^2 + 3^2} = \sqrt{4 + 9} = \sqrt{\mathbf{13}}$$

Step 3: Calculate the cosine similarity

$$\text{sim}_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{12}{\sqrt{13} \times \sqrt{13}} = \frac{12}{13} \approx \mathbf{0.923}$$

Conclusion: The vectors are very similar in orientation.

Proximity for Binary Data

For binary vectors, we use a **contingency table** based on matching attributes.

		Object y		
		1	0	Total
Object x	1	q	r	$q+r$
	0	s	t	$s+t$
	Total	$q+s$	$r+t$	n

- q: number of attributes where $\mathbf{x}=1, \mathbf{y}=1$
- t: number of attributes where $\mathbf{x}=0, \mathbf{y}=0$

Simple Matching Coefficient (SMC)

- For **symmetric** variables (0 and 1 have equal weight, e.g., gender).

$$\text{SMC} = \frac{q+t}{q+r+s+t}$$

Jaccard Coefficient

- For **asymmetric** variables (0-0 matches are ignored, e.g., presence of a disease).

$$J = \frac{q}{q+r+s}$$

Binary Proximity: Worked Example (Part 1/3)

Problem Data

We want to calculate the proximity between Jack and Mary using their attributes.

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

Step 1: Convert to Binary Vectors We convert the attributes to a numerical binary format using the following mapping:

- **Symmetric Attributes (Gender):** M=1, F=0
- **Asymmetric Attributes (Fever, Cough, Tests):** Y/P=1 (presence), N=0 (absence)

This gives us the binary vectors:

- Jack (**x**): (Gender, Fever, Cough, Test-1, Test-2, Test-3, Test-4) $= (\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})$
- Mary (**y**): (Gender, Fever, Cough, Test-1, Test-2, Test-3, Test-4) $= (\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0})$

Binary Proximity: Worked Example (Part 2/3)

Recall Binary Vectors:

- Jack ($\mathbf{x}$): $(\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})$
- Mary ($\mathbf{y}$): $(\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0})$

Step 2: Create the Contingency Table By comparing the vectors attribute by attribute:

- q (attributes where $\mathbf{x} = 1, \mathbf{y} = 1$): 2 (Fever, Test-1)
- r (attributes where $\mathbf{x} = 1, \mathbf{y} = 0$): 1 (Gender)
- s (attributes where $\mathbf{x} = 0, \mathbf{y} = 1$): 1 (Test-3)
- t (attributes where $\mathbf{x} = 0, \mathbf{y} = 0$): 3 (Cough, Test-2, Test-4)

Binary Proximity: Worked Example (Part 3/3)

Step 3: Calculate Similarity Measures

Simple Matching Coefficient (SMC)

(Used for symmetric binary attributes, considers all matches/mismatches)

$$\begin{aligned} \text{SMC} &= \frac{q+t}{q+r+s+t} \\ &= \frac{2+3}{2+1+1+3} \\ &= \frac{5}{7} \approx \mathbf{0.714} \end{aligned}$$

The SMC considers the mismatch in Gender and the matches in absent symptoms (Cough, Test-2, Test-4) equally important.

Jaccard Coefficient

(Used for asymmetric binary attributes, ignores 0-0 matches)

$$\begin{aligned} J &= \frac{q}{q+r+s} \\ &= \frac{2}{2+1+1} \\ &= \frac{2}{4} = \mathbf{0.5} \end{aligned}$$

The Jaccard coefficient focuses only on shared presences (Fever, Test-1) and mismatches where at least one attribute is present. It ignores attributes where both are absent.

Why Standardize Numeric Data?

The Problem of Varying Scales

Many machine learning algorithms and distance metrics are sensitive to the scale of input features. An attribute with a large range (e.g., salary) can dominate and bias the outcome, while an attribute with a small range (e.g., age) might be treated as less important.

Example: Consider calculating the distance between two customers.

- **Customer A:** Age $= 25$, Salary $= \$50{,}000$
- **Customer B:** Age $= 30$, Salary $= \$60{,}000$

The difference in salary ($10,000) is numerically much larger than the difference in age (5). Without standardization, the salary attribute would almost completely determine the distance.

Goal of Standardization

To transform data attributes onto a common scale, ensuring that all features contribute more equally to the analysis, without distorting the differences in the ranges of values.

Method 1: Min-Max Normalization

Concept

This technique rescales a feature to a fixed range, typically $[0, 1]$. It preserves the relationships among the original data values.

Formula (to scale to [0, 1]): For a value v of an attribute A, the normalized value v' is:

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

- $\min_A$: The minimum value of attribute A.
- $\max_A$: The maximum value of attribute A.

Pros & Cons:

Pro: Guarantees all features will have the exact same scale. Useful for algorithms that require bounded inputs.

Con: Highly sensitive to outliers. A single extreme value can compress the rest of the data into a tiny sub-range.

Min-Max Normalization: Worked Example

Consider an 'Income' attribute (in thousands) with the following values:

$$\{25, 30, 45, 60, 150\}$$

Normalize the value $v = 45$ to the range $[0, 1]$.

Step 1: Find the min and max values

- $\text{min}_A = 25$
- $\text{max}_A = 150$

Step 2: Apply the formula

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} = \frac{45 - 25}{150 - 25} = \frac{20}{125} = \mathbf{0.16}$$

Result: The income of 45k is mapped to 0.16 on a [0, 1] scale. The extreme outlier (150) is mapped to 1.

Method 2: Z-Score Standardization

Concept

This technique transforms data to have a **mean of 0** and a **standard deviation of 1**. The resulting value is called a z-score.

Formula: For a value v, the standardized value v' is:

$$v' = \frac{v - \mu}{\sigma} \quad \text{or} \quad v' = \frac{v - \bar{x}}{s}$$

- μ or $\bar{x}$: The mean of the attribute.
- σ or s: The standard deviation of the attribute.

The resulting z-score tells us how many standard deviations a value is from the mean.

Pros & Cons:

Pro: Much less sensitive to outliers than min-max normalization. It is the default choice for many machine learning models.

Con: Does not map data to a specific bounded range.

Z-Score Standardization: Worked Example

Problem: Using the 'Income' data $\{25, 30, 45, 60, 150\}$, standardize the value $v = 45$.

Step 1: Calculate the Mean ($\bar{x}$)

$$\bar{x} = \frac{\sum x_i}{n} = \frac{310}{5} = \mathbf{62}$$

Step 2: Calculate Standard Deviation (s)

The sum of squared differences is $\sum(x_i - \bar{x})^2 = 10430$.

$$s^2 = \frac{10430}{4} = 2607.5$$

$$s = \sqrt{2607.5} \approx \mathbf{51.06}$$

Step 3: Apply Z-Score Formula

$$v' = \frac{v - \bar{x}}{s}$$

$$= \frac{45 - 62}{51.06}$$

$$= \frac{-17}{51.06}$$

$$\approx \mathbf{-0.333}$$

The income of 45k is **0.333** standard deviations **below** the mean.

References

[1] Jiawei Han, Micheline Kamber, & Jian Pei, *Data Mining: Concepts and Techniques*, 4th Edition, Morgan Kaufmann, 2012.

[2] David J. Hand, Heikki Mannila, & Padhraic Smyth, *Principles of Data Mining*, First Edition, A Bradford Book, 2001.

[3] Richard O. Duda, Peter E. Hart, & David G. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2001.