Data basics

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Outline

- Types of datasets
- Data objects and attributes
- Distance and similarity
- Data normalization

Types of Datasets

1. Record Data

- records in a relational database

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Types of Datasets

- 2. Graphs and networks

Image credit: Medium

Social Networks

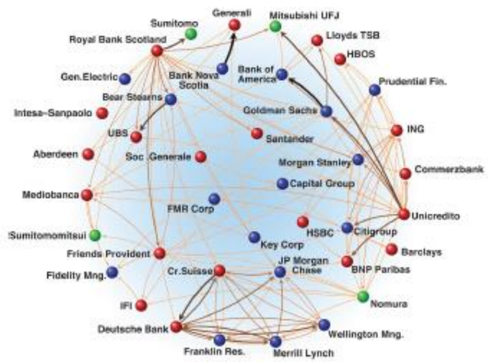
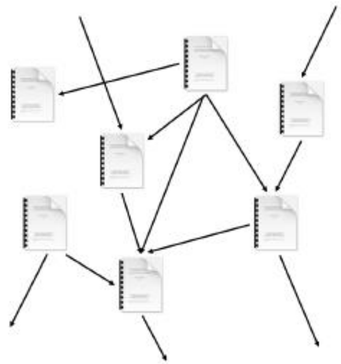


Image credit: Science

Economic Networks

Image credit: Lumen Learning

Communication Networks



Citation Networks

Image credit: Missoula Current News

Internet

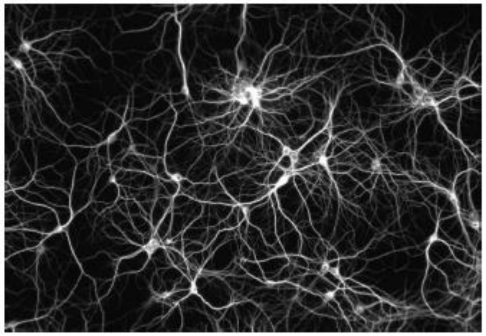


Image credit: The Conversation

Networks of Neurons

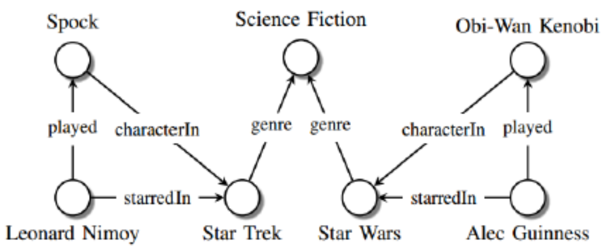


Image credit: Maximilian Nickel et al

Knowledge Graphs

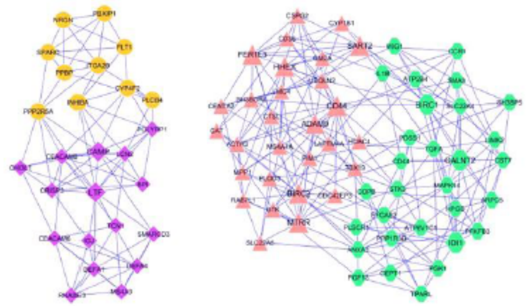


Image credit: ese.wustl.edu

Regulatory Networks

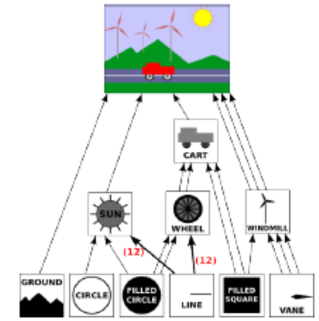


Image credit: math.hws.edu

Scene Graphs

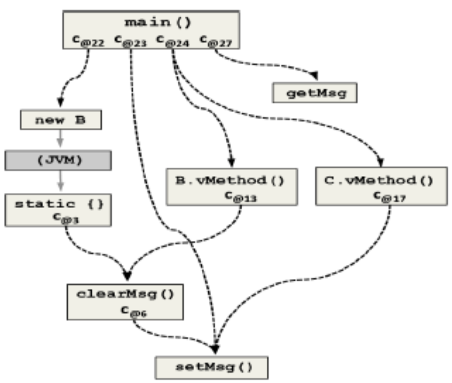


Image credit: ResearchGate

Code Graphs

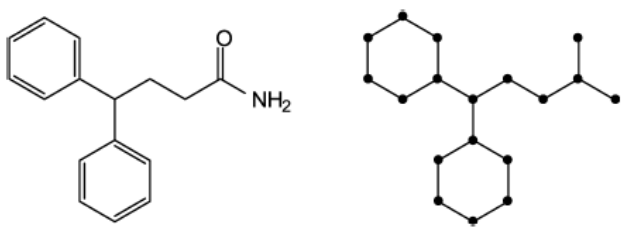


Image credit: MDPI

Molecules

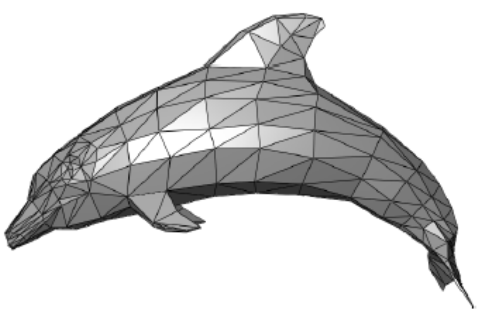


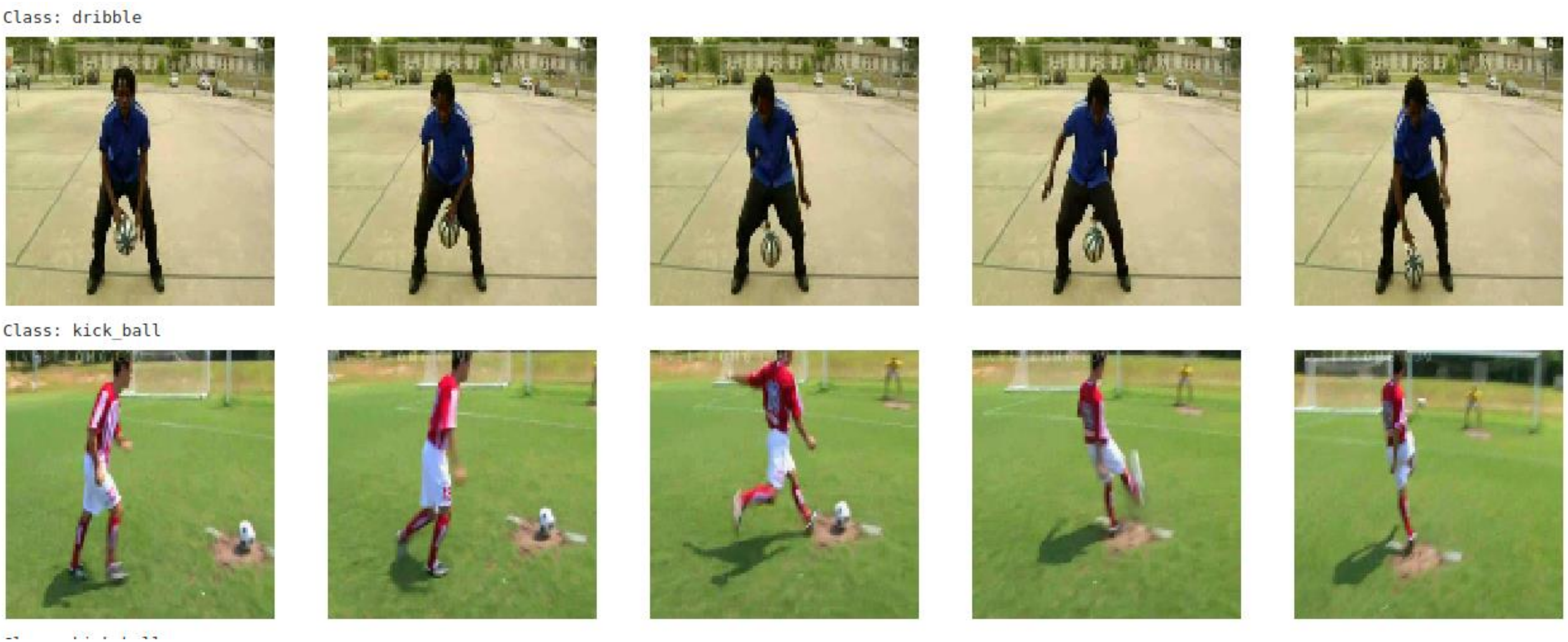
Image credit: Wikipedia

3D Shapes

Types of Datasets

- 3. Multimedia data

Images



Videos / image sequences

Types of Datasets

- 4. Text data
  - Twitter/Facebook posts
  - News
  - Wikipedia texts
  - Shopping item comments
  - Books
  - Transcripts
  - Emails
  - Documents
  - …

The Daily News

21st July 1969 THE WORLD'S FAVOURITE NEWSPAPER - Since 1879

The Eagle Has Landed!

The American flag is placed on the moon

American astronaut Neil Armstrong has become the first person to step on the surface of the moon.

Armstrong used the phrase 'the eagle has landed' to let Houston know that the lunar module had actually landed on the moon.

As Armstrong put his foot on the moon he declared: "That's one small step for man, one giant leap for mankind." The craft landed on an area of the moon called 'the sea of tranquility'.

Armstrong was joined on the moon by Edwin 'Buzz' Aldrin. A third astronaut, called Michael Collins, stayed aboard the Columbia mothership and orbited the moon.

The two moon bound astronauts stayed on the moon for a total of around 21 hours. They spent most of this time inside the lunar module.

Story continued on the next page.

Data objects and attributes

- Data objects are also called samples, examples, instances, data points, records, tuples, ….
- Data attributes are also called dimensions, features, variables, channels, …
- Data sets are made up of data objects/samples
- Data objects are described by attributes/features

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Each row represents a data sample;
Each column represents a data attribute.

Attribute Types

- **Numeric:** real numbers (continuous values)
  - *Prices, $500, $200; Image pixel intensity: [0 255]*
  - *temperature, height, or weight*

- **Nominal:** (Categorical) categories, states, or "names of things" (discrete values)
  - *Hair_color = {black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes

- **Binary** (discrete values)
  - Nominal attribute with only 2 states (0 and 1)
  - {true, false},
  - {1, 0} indicates one item exists or not

- **Ordinal** (discrete values)
  - Values have a meaningful order (ranking)
    but magnitude between successive values is not known
  - *Size = {small, medium, large}*, grades={A, B, C, D},

- Convert raw data into numeric values
  - For many data mining (machine learning) tasks, e.g., clustering, classification, regression.
  - Nominal ->numeric: use one-hot encoding
  - Ordinal->numeric: use numbers to indicate ranking (1,2,3,…)
  - Binary -> numeric: convert to {0, 1}

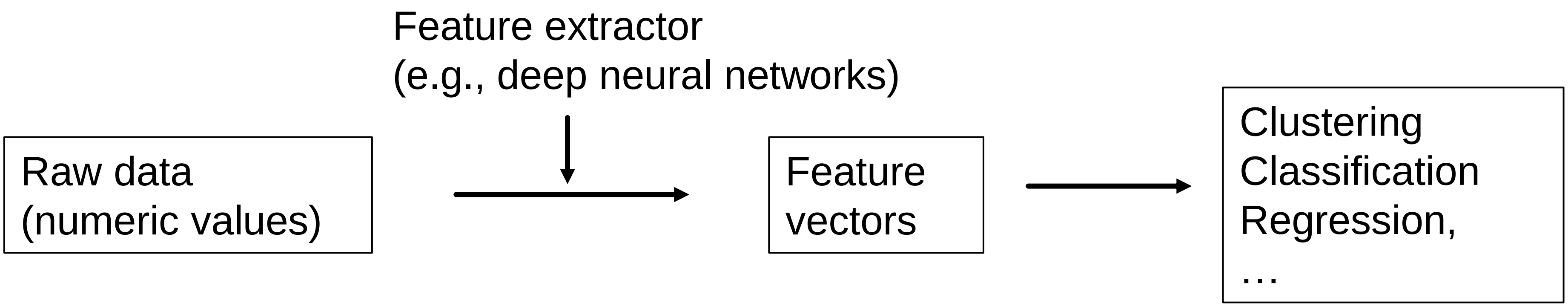
Example: convert nominal values to numeric values using one-hot encoding

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

A typical pipeline for applying data mining techniques:



Feature vectors: one data sample is represented by one feature vector.
e.g., $\mathbf{x} = [x_{1,}\ x_{2,}\ x_{3,}\ x_{4,\ \ldots}]$

- Vector norm
  - also called vector magnitude, the length of the vector

1. Lp-norm (general):

Let $p \geq 1$ be a real number. The p-norm (also called ℓ_p-norm) of vector $\mathbf{x} = (x_1, \ldots, x_n)$ is[9]

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

2. L1-norm:

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i| .$$

3. L2-norm:

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

Also called Euclidean norm, vector length

Distance

Given two vectors: $\mathbf{x}_i$ and $\mathbf{x}_j$ (they have l dimensions)

1. Calculate the vector difference (residual vector): $\mathbf{r} = \mathbf{x}_i - \mathbf{x}_j$
2. apply vector norm on the vector difference:

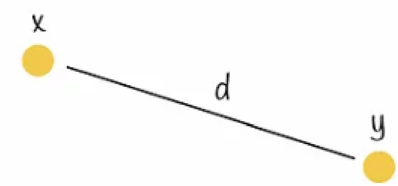
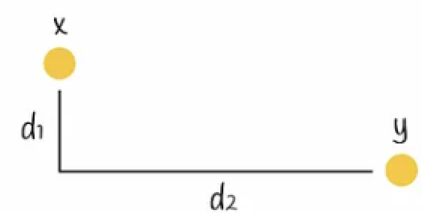
$$d_p(i, j) = \| \mathbf{r} \|_p = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_p$$

- p = 1: (L_1 norm) Manhattan distance (L1 distance)

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

- p = 2: (L_2 norm) Euclidean distance (L2 distance)

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

Metric	Formula	Interpretation
Euclidean distance	$d = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$	
Manhattan distance	$d = \sum_{i}^{n} \lvert x_i - y_i \rvert$	 $d = d_1 + d_2$

https://towardsdatascience.com/similarity-search-knn-inverted-file-index-7cab80cc0e79

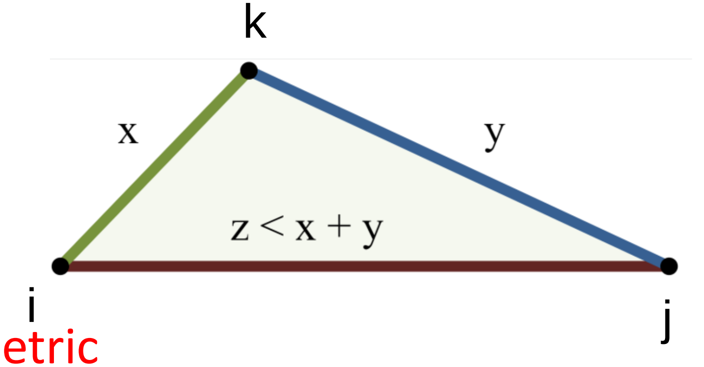
Distance

- Minkowski distance (defined by vector norm):

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

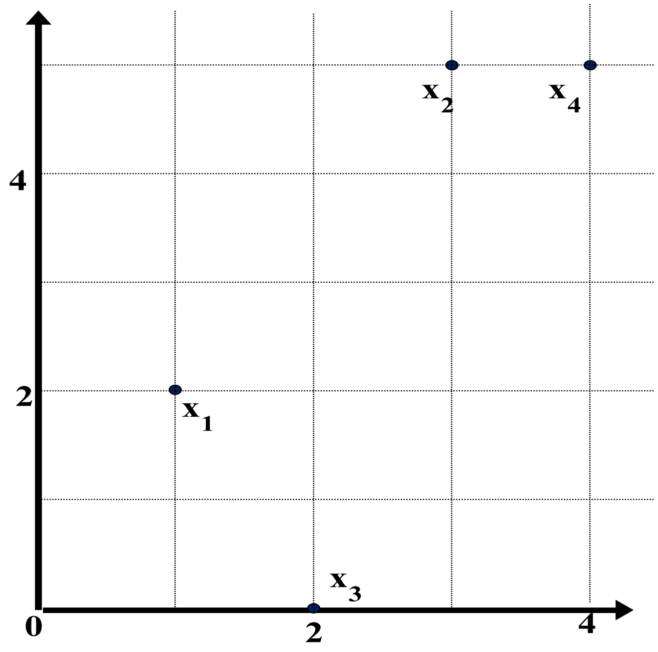
  where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two l-dimensional data objects, and p is the order
  (defined based on L-p norm)

- It has the 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

- Data matrix
  - Describe a data set (data samples)
  - E.g, a data matrix of n data points with d dimensions
  - Each row indicates a feature vector

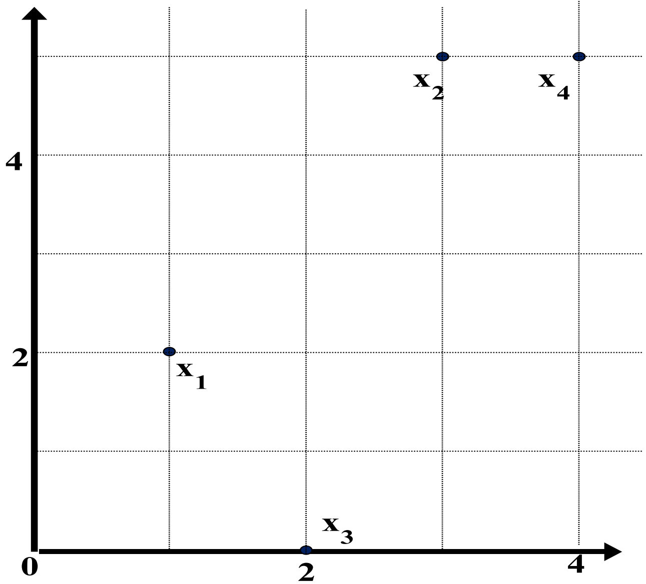


Data Matrix
n=4 (data points), d=2 (dimensions)

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

Distance example

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



Manhattan distance (L_1)

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

Euclidean distance (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

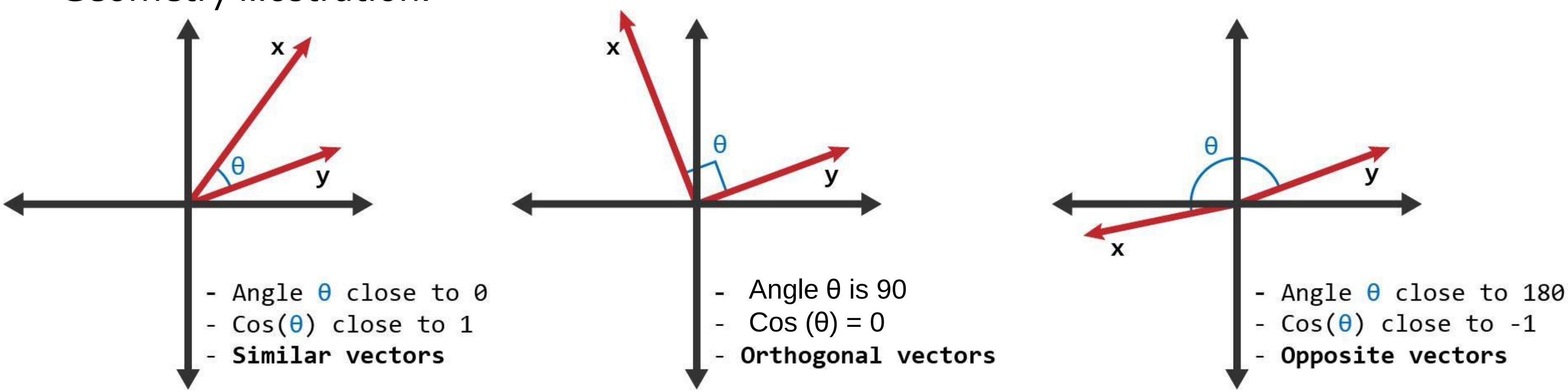
Similarity

- Cosine similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Cosine distance = 1- cosine similarity

Geometry illustration:



https://www.learndatasci.com/glossary/cosine-similarity/

Similarity example

- $D1 = [1, 1, 1, 1, 1, 0, 0]$
- $D2 = [0, 0, 1, 1, 0, 1, 1]$

First, we calculate the dot product of the vectors:

$$D1 \cdot D2 = 1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 0 \times 1 + 0 \times 1 = 2$$

Second, we calculate the magnitude (L2 norm) of the vectors:

$$\|D1\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2} = \sqrt{5}$$

$$\|D2\| = \sqrt{0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{4}$$

Finally:

$$similarity(D1, D2) = \frac{D1 \cdot D2}{\|D1\|\|D2\|} = \frac{2}{\sqrt{5}\sqrt{4}} = \frac{2}{\sqrt{20}} = 0.44721$$

We can further calculate the angle between the vectors:

$$cos(\theta) = 0.44721$$

$$\theta = \arccos(0.44721) = 63.435$$

Data normalization

- Data normalization
  - The goal of normalization is to transform attributes/features to be on a similar scale.
    - Algorithms may bias to the features which have a larger magnitude.
    - E.g., L2-distance will be dominated by large attributes

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

Examples:
x1= [100, 0.1]
x2= [120, 0.01]
x3= [120, 0.1]

$\longrightarrow$ d(x1, x2) ≈ d(x1, x3)

- 1. Max-Min Normalization
  - rescale to a value in [0, 1]

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- 2. Z-score normalization
  - Also called standardization
  - rescale to ensure the mean and the standard deviation to be 0 and 1
  - More robust to outlier

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

σ = population standard deviation
N = the size of the population
X_i = each value from the population
μ = the population mean

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html
https://www.nextdatalab.com/standard-deviation/

Example

Input dataset

user	Age	Salary
1	40	100000
2	32	80000
3	21	43000
4	24	51000
5	35	70000

Age Mean	30.4
Age Std	7.829432
Salary mean	68800
Salary std	22818.85

Age min	21
Age max	40
Salary min	43000
Salary max	100000

Z-score normalization

user	Age	Salary
1	1.2261426	1.367291
2	0.2043571	0.490822
3	-1.200598	-1.13064
4	-0.817428	-0.78006
5	0.5875267	0.052588

Max-Min Normalization

user	Age	Salary
1	1	1
2	0.5789474	0.649123
3	0	0
4	0.1578947	0.140351
5	0.7368421	0.473684