DATA AND DISTANCE MEASURES

Courtesy:
Jiawei Han, Micheline Kamber, and Jian Pei
Anjali Jivani

TYPES OF DATA SETS

RECORD

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

GRAPH AND NETWORK

- World Wide Web
- Social or information networks
- Molecular Structures

ORDERED

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

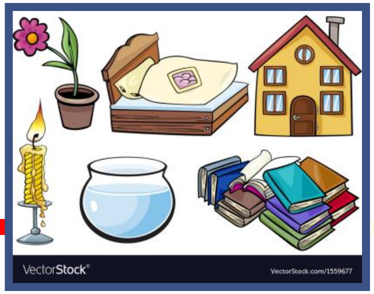
SPATIAL, IMAGE AND MULTIMEDIA:

- Spatial data: maps
- Image data:
- Video data:

IMPORTANT CHARACTERISTICS

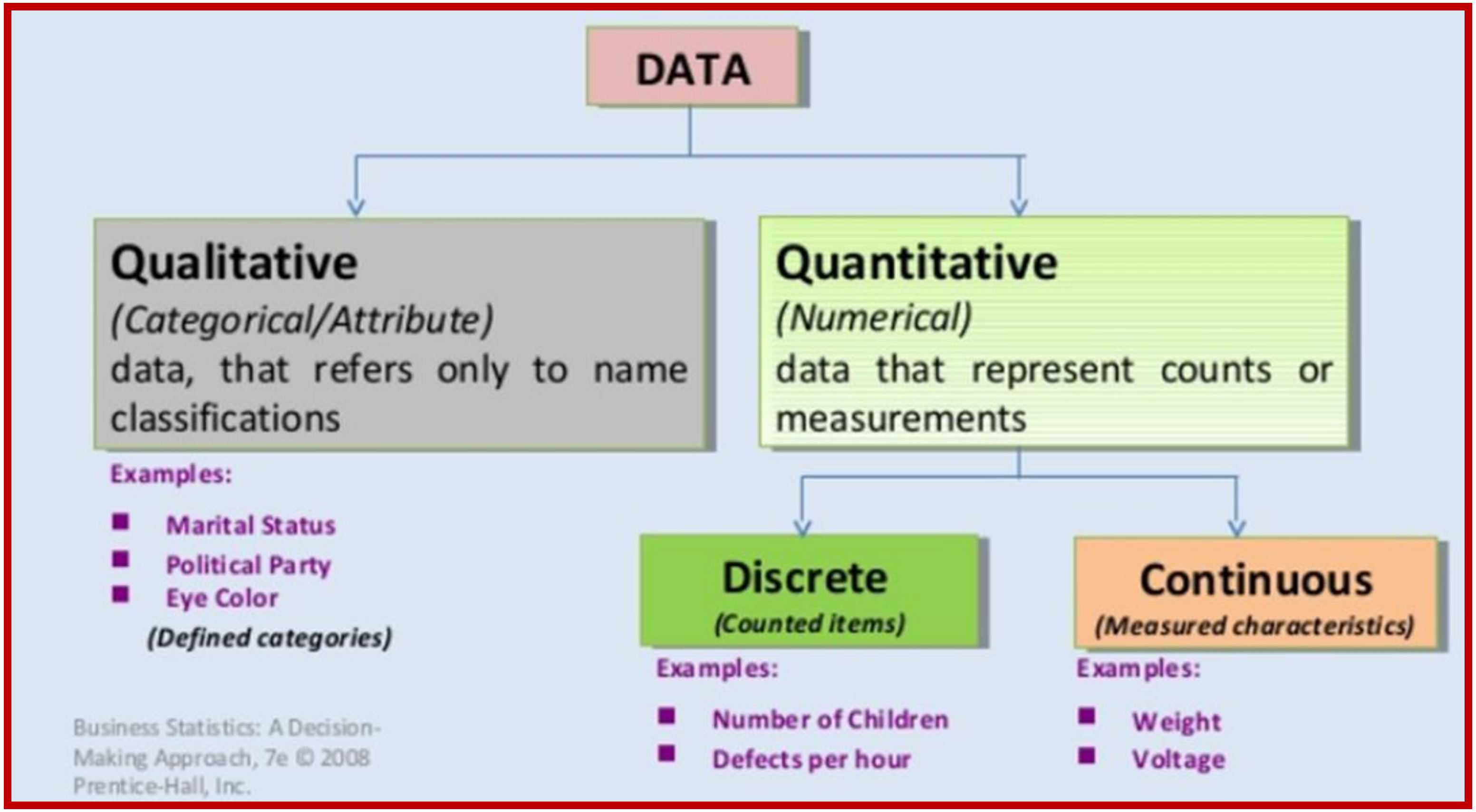
- DIMENSIONALITY
  - Curse of dimensionality
- SPARSITY
  - Only presence counts
- RESOLUTION
  - Patterns depend on the scale
- DISTRIBUTION
  - Centrality and dispersion

DATA OBJECTS

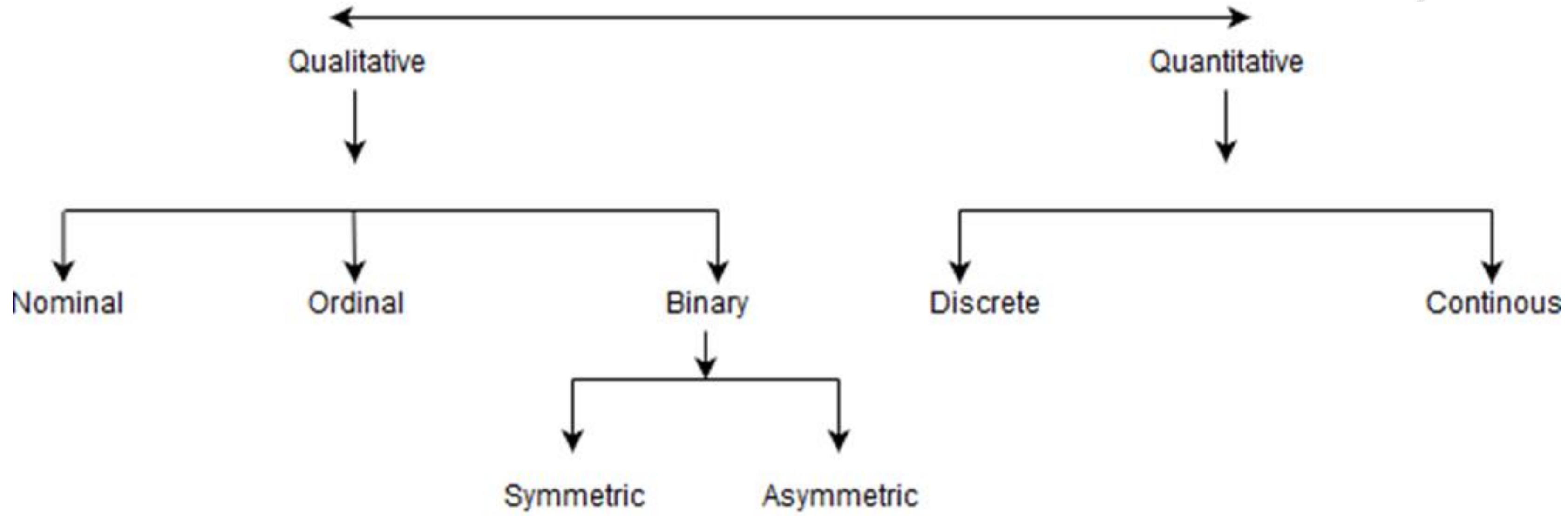


- **Data sets are made up of data objects.**
- **A data object represents an entity.**
- **Examples:**
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- **Also called *samples , examples, instances, data points, objects, tuples*.**
- **Data objects are described by attributes.**
- **Database rows -> data objects; columns ->attributes.**

ATTRIBUTES / FEATURES



ATTRIBUTE TYPES



- **Nominal**
  - categories, states, or “names of things”
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings, designation

ATTRIBUTE TYPES

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - E.g. gender
  - Asymmetric binary: outcomes not equally important.
    - medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Numeric:

- **Integer or real-valued**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

ATTRIBUTE TYPES

- **Discrete**
  - Discrete data have finite values it can be numerical and can also be in categorical form.
  - These attributes has finite or countable infinite set of values.
    - E.g. Number of people living in your town, number of students who take statistics, pin codes, etc.
- **Continuous**
  - Continuous data have infinite no of states.
  - Continuous data is of float type. There can be many values between 2 and 3.
    - E.g. height, weight, etc.

Attribute	Values
Gender	Male , Female

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

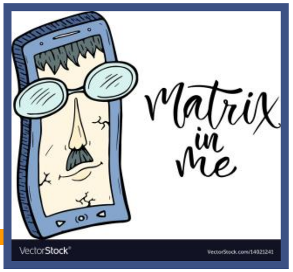
Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 ....etc

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

SIMILARITY AND DISSIMILARITY

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity (e.g., distance)**
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity refers to a similarity or dissimilarity**
  - Helps in identifying objects which are similar to each other
  - Used especially in clustering, classification,…

DATA MATRIX AND DISSIMILARITY MATRIX



Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

DISTANCE MATRIX

$$\begin{array}{cl} \left[\begin{array}{ccccc} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{array}\right] & \begin{array}{l} \rightarrow \text{Object 1} \\ \rightarrow \text{Object 2} \\ \\ \\ \rightarrow \text{Object n} \end{array} \end{array}$$

$$\begin{bmatrix} 0 & d(1,2) & d(1,3) & & \\ d(2,1) & 0 & d(2,3) & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

$d(x,x) = 0$
$d(x,y) = d(y,x)$

- There are n number of objects with p number of attributes.
- The distance matrix stores the distance between every object with every other.
- Since the distance between two objects say x and y, d(x,y) is same as d(y,x), we consider only the lower triangular matrix.

STANDARDIZING NUMERIC DATA

- Data can be transformed to convert it to unit less data and to suit the data mining algorithm. One popular method is the z-score normalization.

$$z = \frac{x - \mu}{\sigma} \qquad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - “-” negative when the raw score is below the mean, “+” when above

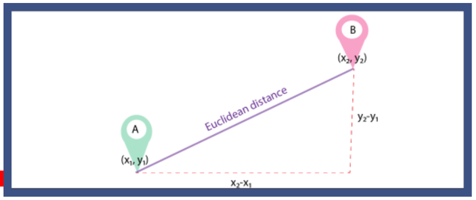
- An alternative way: Calculate the mean absolute deviation (instead of σ)

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|) \text{ i.e. } \frac{\sum (xi - mf)}{n}$$

Where, $m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$. $\qquad z_{if} = \frac{x_{if} - m_f}{s_f}$

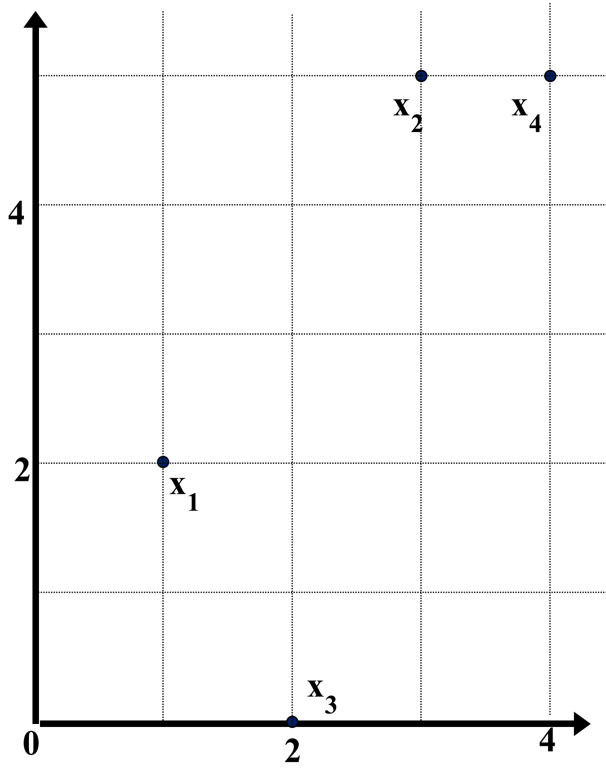
- Using mean absolute deviation is more robust than using standard deviation

THE EUCLIDEAN DISTANCE



The most popular distance measure for interval-scaled variables is the Euclidean Distance.

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$



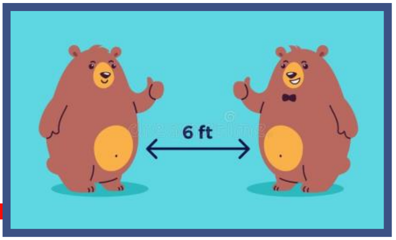
Data Matrix

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

Dissimilarity Matrix
(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

THE MINKOWSKI DISTANCE



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

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm).

- **Properties:**
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - 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**
- **One may use a weighted formula to combine their effects**

$$d(i,j) = \sqrt{(w1|x_{i_1} - x_{j_1}|^2 + w2|x_{i_2} - x_{j_2}|^2 + \ldots + wp|x_{i_p} - x_{j_p}|^2)}$$

DISTANCE MEASURES

h **= 1: “Manhattan” (city block, L_1 norm) distance**

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

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

h **= 2: “Euclidean” (L_2 norm) distance**

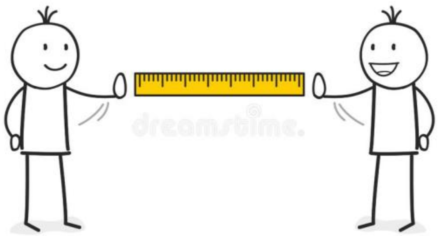
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

h **$\rightarrow \infty$. “Supremum” (L_{max} norm, L_∞ norm) distance.**

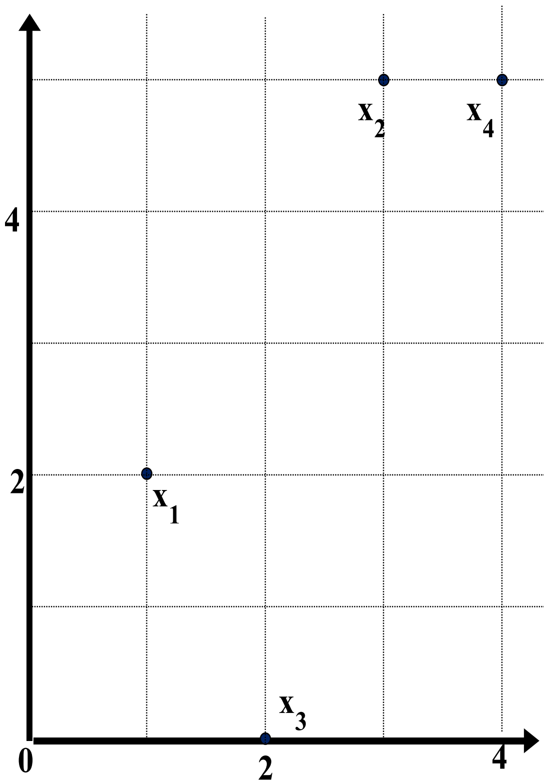
- This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

EXAMPLES OF DISTANCE



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_{max})

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

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

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:**

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

BINARY ATTRIBUTES

All attributes are binary:

	A1	A2	A3	A4	A5	A6	A7
i	1	1	0	0	0	1	0
j	1	0	1	0	1	0	1

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

q = 1, r = 2, s = 3, t = 1

Symmetric Binary d(i,j) = (r + s) / (q + r + s + t)
= (2+3) / 7
= 5/7 = 0.71

Asymmetric Binary d(i,j) = (r + s) / (q + r + s)
= (2+3) / 6
= 5/6 = 0.83

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

Object *j*

Object *i*

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- **Jaccard coefficient (*similarity* measure for *asymmetric* binary variables): 1 - d(i, j), where d(i, j) is distance for asymmetric binary**

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

- **Note: Jaccard coefficient is the same as "coherence":**

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

EXAMPLE OF 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 to be considered for distance)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

Jack

Mary	1	0
1	q = 2	r = 1
0	s = 0	t = 3

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

d(Jack, Mary) = (1 + 0) / (2 +0 +1)
= 0.33

EXAMPLE OF 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

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

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

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

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

- Can take 2 or more states, e.g. hair colour - red, black, brown, grey (generalization of a binary attribute)
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

Name	Hair colour	Skin Colour	Eyes Colour	Country
Jack	Black	Fair	Blue	Germany
Jim	Black	Dark	Brown	India

  - d(Jack, Jim) = (4 – 1)/4 = 3/4 = 0.75

NOMINAL ATTRIBUTES EXAMPLE

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the *M* nominal states

Name	Hair black	Hair red	Hair brown	Skin Fair	Skin Dark
Jack	1	0	0	1	0
Jim	1	0	0	0	1

  - In this way all possible values of nominal are converted to binary. In this case it is symmetric binary.
  - Use the distance measure of symmetric binary

PROXIMITY MEASURE FOR ORDINAL ATTRIBUTES

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank

    $$r_{if} \in \{1, \ldots, M_f\}$$

  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f* th variable by

    $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Example:

Qualification:

SSC,HSC,UG,PG,PHD

1. Assign ranks: (rif)

   SSC – 1

   HSC – 2

   UG – 3

   PG – 4

   PHD – 5

2. Find zif

   zif = (1-1)/(5-1) = 0 for SSC

   zif = (3-1)/(5-1) = 0.5 for UG

   zif = (5-1)/(5-1) = 1 for PHD

All values lie between (0,1)

Now use Euclidean, Manhattan,...

ATTRIBUTES OF MIXED TYPES

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal

	Fever (Asymmetric Binary)	Cough (Asymmetric Binary)	Height (Numeric)	Weight (Numeric)	Gender (Symmetric Binary)	Skin Colour (Nominal)
i	Y	N	165	64	Female	Fair
j	N	N	150	Null	Female	Dark
δ	1	0	1	0	1	1

- One may use a weighted formula to combine their effects

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

  - **$d_{ij}^{(f)}$** is the distance between object i and j for the ***f*** ***th*** attribute.
  - **δ** is called the **indicator** and can take values 1 or 0.
  - δ takes the value 0 only when:
    - There is a missing value for an attribute
    - The attribute is asymmetric binary and both i and j have ‘N’ or 0 values

MIXED TYPES

	Fever (Asymmetric Binary)	Cough (Asymmetric Binary)	Height (Numeric)	Weight (Numeric)	Gender (Symmetric Binary)	Skin Colour (Nominal)
i	Y	N	165	64	Female	Fair
j	N	N	150	Null	Female	Dark
δ	1	0	1	0	1	1

- *f* is binary or nominal:
  - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- *f* is numeric: use the normalized distance

$$d_{ij}^{(f)} = \frac{|xif - x_{jf}|}{\max_{hf} - \min_{hf}}$$

where, $\max_{hf}$ is the maximum value over all non-missing values of f and $\min_{hf}$ is the minimum value over all non-missing values of f.

- *f* is ordinal
  - Compute ranks r_{if} and calculate z_{if}

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Treat z_{if} as numeric and find the distance.

ATTRIBUTES OF MIXED TYPES

	Fever (Asymmetric Binary)	Cough (Asymmetric Binary)	Height (Numeric)	Weight (Numeric)	Gender (Symmetric Binary)	Skin Colour (Nominal)
i	Y	N	165	64	Female	Fair
j	N	N	150	Null	Female	Dark
δ	1	0	1	0	1	1

$$d(i, j) = \frac{(1 * dij^{fr}) + (0 * d_{ij}^{cg}) + (1 * d_{ij}^{ht}) + (0 * d_{ij}^{wt}) + (1 * d_{ij}^{gd}) + (1 * d_{ij}^{sk})}{4}$$

1. $d_{ij}^{fr} = 1$ (fever is asymmetric binary and both are diff.)
2. $d_{ij}^{ht} = \frac{|165 - 150|}{200 - 75} = \frac{15}{125} = 0.12$

$$d_{ij}^{(f)} = \frac{|xif - xjf|}{\max_{hf} - \min_{hf}}$$

3. $d_{ij}^{gd} = 0$ (gender is symmetric binary and both are same)
4. $dij^{sk} = 1$ (skin colour is nominal and both are diff.)

$$d(i, j) = \frac{(1*1)+(0*0)+(1*0.12)+(0*0)+(1*1)}{4} = 0.53$$

EXAMPLE (MIX TYPES)

- Find the distance between the following cars and find which are most similar and which are most different:

	Petrol/diesel	Color	Weight	Size	Average (per km)	Popular	Price (in lacs)
Honda (i)	P	Silver	150	M	14	Y	10
Toyota (j)	D	White	null	L	20	Y	16
Audi (k)	P	Black	350	L	15	N	28

- Petrol/diesel (symmetric binary)
- Color (nominal) – white, black, blue, silver, red, grey
- Weight (numeric) – max. 500 and min. 100
- Size (ordinal) – VS, S, M, L, VL
- Average (numeric) – max. 25 and min. is 6
- Popular (asymmetric binary)
- Price (numeric) – max. 50 and min. 3

EXAMPLE (MIX TYPES)

	Petrol/diesel	Color	Weight	Size	Average (per km)	Popular	Price (in lacs)
Honda (i)	P	Silver	150	M	14	Y	10
Toyota (j)	D	White	null	L	20	Y	16
Audi (k)	P	Black	350	L	15	N	28

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

d(honda, audi) = $\frac{A}{B} = \frac{A}{1+1+1+1+1+1+1} = \frac{A}{7}$

- Petrol/diesel (symmetric binary)
  $d^{pet/dei}$ **=0** (as they match)
- Color (nominal)
  d^{color} **=1** (as they don't match)
- Weight (numeric) – max. 500 and min. 100
  $d^{weight} = \frac{|150-350|}{500-100} = 0.5$
- Size (ordinal) – VS, S, M, L, VL
  1. Assign ranks:
     VS – 1, S – 2, M – 3, L – 4, VL – 5
  2. $Z_M = \frac{3-1}{5-1} = 0.5$, $Z_L = \frac{4-1}{5-1} = 0.75$
  3. $d^{size} = \frac{|0.5-0.75|}{1-0} = 0.25$
- Average (numeric) – max. 25 and min. is 6
  $d^{average} = \frac{|14-15|}{25-6} = 0.053$
- Popular (asymmetric binary)
  $d^{popular}$ **=1** (as they don't match)

EXAMPLE (MIX TYPES)

	Petrol/diesel	Color	Weight	Size	Average (per km)	Popular	Price (in lacs)
Honda (i)	P	Silver	150	M	14	Y	10
Toyota (j)	D	White	null	L	20	Y	16
Audi (k)	P	Black	350	L	15	N	28

Price (numeric) – max. 50 and min. 3

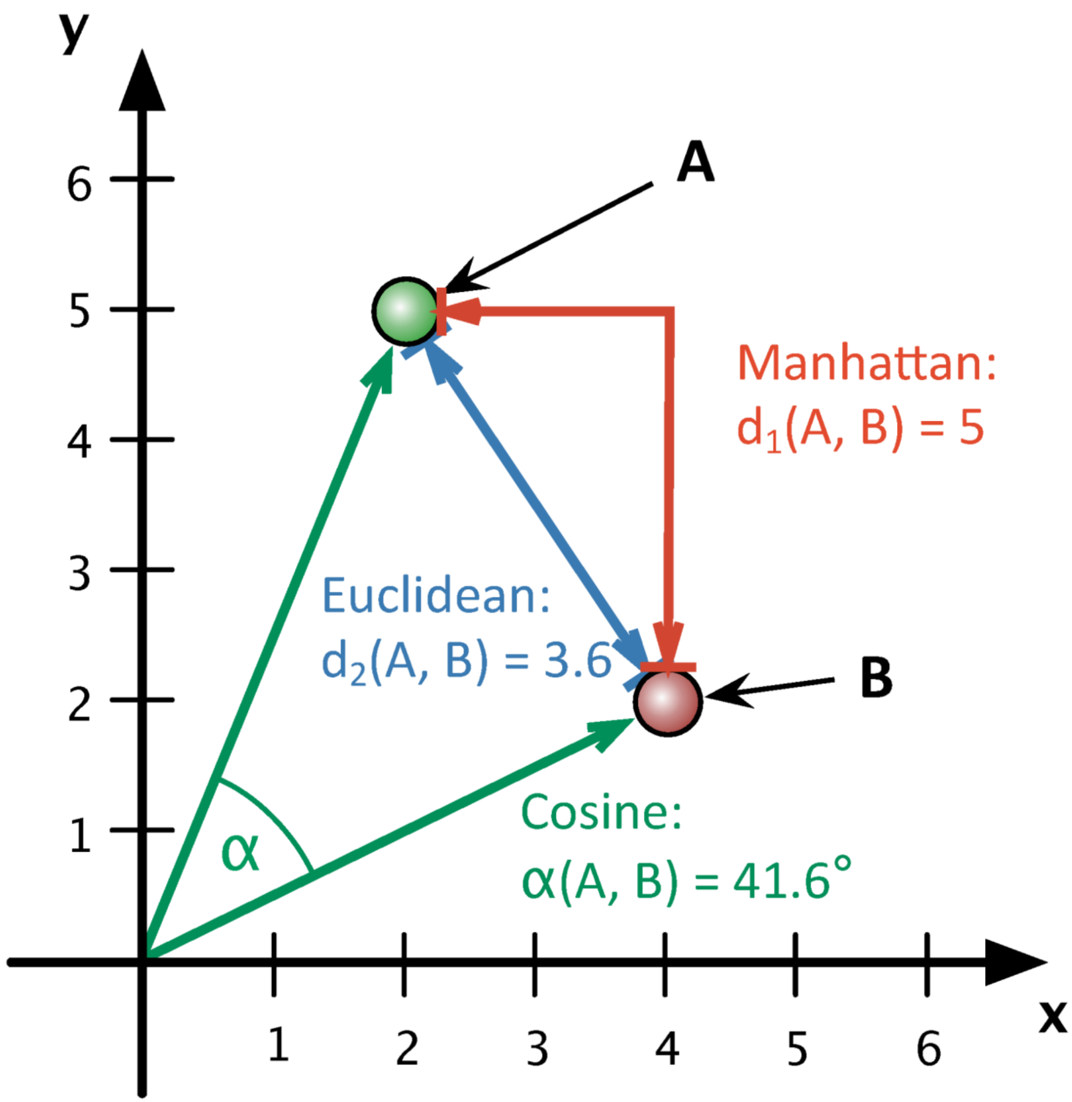
$$d^{price} = \frac{|10 - 28|}{50 - 3} = 0.383$$

- d(honda, audi) = $\frac{A}{B} = \frac{A}{1+1+1+1+1+1+1} = \frac{A}{7}$
- A = 1 x 0 + 1 x 1 + 1 x 0.5 + 1 x 0.25 + 1 x 0.053 + 1 x 1 + 1 x 0.383 = 3.186

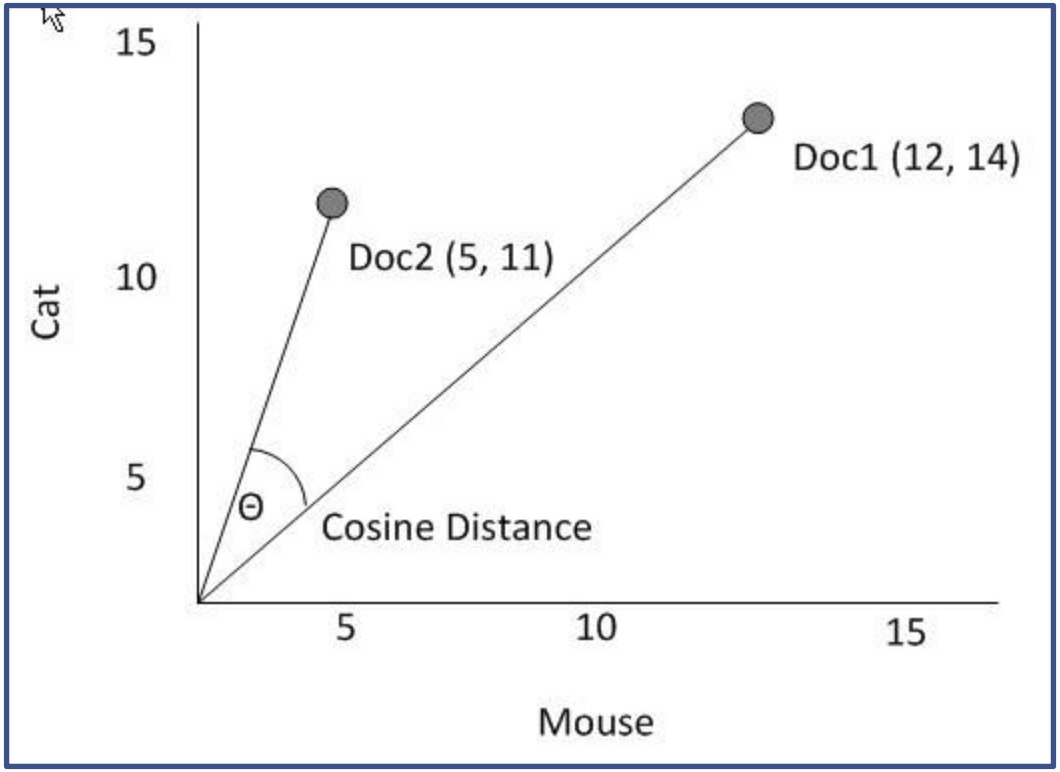
d(honda, audi) = $\frac{A}{B} = \frac{3.186}{7}$ = 0.455

- Similarly find d(honda, toyota) and d(toyota, audi). The distance value which is smallest shows the two cars which are most similar and the largest distance shows the two cars which are least similar.

COSINE SIMILARITY



COSINE SIMILARITY



$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

COSINE SIMILARITY

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) 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

- Applications: information retrieval, text mining, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

  $$\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1||\ ||d_2||)\ ,$$

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

EXAMPLE OF COSINE SIMILARITY

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

- 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)

$d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25
$||d_1||$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}$=$(42)^{0.5}$ = 6.481
$||d_2||$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}$=$(17)^{0.5}$ = 4.12
cos(d_1, d_2) = 25 / (6.481 x 4.12) = 25 / 26.702

cos(d_1, d_2) = 0.94

SUMMARY

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

END