2-data

data = data objects + their attributes

attribute = property / characteristis of an object

object = collection of attributes

attribute values = numbers/symbols assigned to an attribute for a particular object

↳ same attribute can be mapped to different attribute values (ex: height in feet or meters)

	attribute type	description	operations	examples	transformation
categorical (qualitative)	nominal	labeled	$=, \neq$	color of.., types of .., zip codes, tc no, list of popular...	any permutation of vals (assign all again)
	ordinal	labaled + order	$= \neq$ $> <$	street numbers, school letter grades, political orientation	new_val = f(old_val) f → monotonic increasin (preserving order)
numerical (quantative)	interval	labaled + order + equal interval	$= \neq$ $> <$ $+ -$	date, temp (in celcius / fahren) test scores, time in clock, iq (bcs of no zero)	new_v = a x old_v + b
	ratio	labeled + order + equal interval + true zero	$= \neq$ $> <$ $+ -$ $\times \div$	weight, height, age, time it takes, income, temp in kelvin	new_val = a x old_val

even the group does not include 0, it is still ratio → ages of the people in the university

discrete attribute = finite/ countably infinite set of val → counts, zip codes, binary var

continuous attribute = real number val → temp, height, weight

symmetric attribute = all its vals are equally valuable

asymmetric attribute = only presence (non-zero val) is regarded as important

important characteristic of data

↳ dimentionality = number of attributes

↳ sparsity = only presence counts

↳ resolution = pattern depends on the scale

↳ size

types of dataset

↳ record data = each object has fixed set of attributes

↳ data matrix = if data objects have same fixed set of numeric attributes

↳ document data = each object is a component - attribute vector →

	A	B	C	D	E
doc1 -	1	1	2	0	4
doc2 -	3	1	1	2	1

→ each document becomes term vector
↳ each term is a component / att of vec
↳ ex: frequency of words in documents

↳ transaction data = each transaction involves set of items → set of purchased products (no quantity only existance)

↳ graph = word wide web, moleculer structures

↳ ordered = sequence of transactions → avg montly temp of world → spational (related to location) + temporal (rel to time)

data quality problems = noise, outliers, missing values, duplicate data, wrong / fake data

↳ for objects, it is an extraneous object; for attributes, it is modification of original values

outlier: case1 = outliers are noise that interferes with data analysis, ignore/remove

case 2 = outliers are the goal of analysis, ex: cancer cells

similarity - dissimilarity measures = proximity

distance = dissimilarity $s \in [0,1]$ $d \in [0,k)$

↳ nominal if $x=y$ sim=1 dis=0, $x \neq y$ sim=0 dis=1

↳ ordinal dis = $|x-y| / (n-1)$ values mapped to int 0 to n-1 sim = 1-d

↳ interval or ratio $d = |x-y|$ $s = -d$, $s = 1/(1+d)$, $s = e^{-d}$, $s = 1 - \frac{d - min_d}{max_d - min_d}$

• minkowski distance = $d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$ r = parameter, n = number of attributes

↳ r=1 → manhattan dist → ex: used to find the distance between binary vectors → hamming distance

↳ r=2 → euclidean dist } n can be anything, n and r are independent

1 0 0 1 1
1 1 0 0 1 } ②

↳ r=∞ → supremum (max dist between any component of the vectors)

• mahalanobis distance = difference between a point and a distribution

common properties of distance metrics

↳ $d(x,y) \geq 0$, $d(x,y) = 0 \Leftrightarrow x=y$

↳ $d(x,y) = d(y,x)$

↳ $d(x,z) \leq d(x,y) + d(y,z)$

common properties of similarity metrics

↳ $s(x,y) = 1$ only if $x=y$ (cosine does not hold)

↳ $s(x,y) = s(y,x)$

• similarity between binary vectors

↳ f_{00}: # att x=0, y=0

f_{10}: # att x=1, y=0

:

SMC - simple matching = # matches / # attributes $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

J - Jaccard coefficients = # f_{11} / # non-zero attributes $= f_{11} / (f_{01} + f_{10} + f_{11})$

• focusing only existance of attributes

x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1 } $f_{00} = 7$ $f_{11} = 0$ $f_{10} = 1$ $f_{01} = 2$ SMC = 7/10 J = 0 → more meaningful mostly

• cosine similarity = $\cos(d_1, d_2) = \langle d_1, d_2 \rangle / \|d_1\| \cdot \|d_2\|$

↳ d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 1 0 0 0 0 0 0 1 0 2 }

$\langle d_1, d_2 \rangle = 3.1 + 2.0 + \ldots + 0.2 = 5$
$\|d_1\| = \sqrt{3^2 + 2^2 \ldots 0^2} = 6.481$
$\|d_2\| = 2.499$ } $\cos(d_1, d_2) = 0.315$

$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x}, \mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \, s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\mathrm{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12)$$

$$\mathrm{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\mathrm{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

↳ not affected

- **x** = (1, 2, 4, 3, 0, 0, 0), **y** = (1, 2, 3, 4, 0, 0, 0)
- **y_s = y * 2** (scaled version of y), **y_t = y + 5** (translated version)

Measure	(x , y)	(x , y_s)	(x , y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

→ since cos is angle btw vectors, scaling does not affect

information based measures → information relates possible outcomes of an event

↳ the more certain an outcome, the less information that it contains

↳ throwing a dice (1/6) contains more information than flipping a coin (1/2)

entrophy = between 0 and $\log_2 n$

- a variable (event), X,
- with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
- each outcome having probability, $p_1, p_2 \ldots, p_n$
- the entropy of X, $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

ex: coin → $H = -0.5 \log(0.5) - 0.5 \log(0.5) = 1$

max entrophy = $\log_2 n$ → when all outcomes have equal probability

mutual information = max = $\log_2(\min(n_x, n_y))$

Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

maximal information coefficient = to compute the mutual info of two continuous variables

general approach for combining similarities

• $s_k(x,y) \in [0,1]$

• $\delta_k = 0$ if k^{th} attribute is asymmetric, or one of the objects has a missing val for kth att } indicator variable

↳ $\delta_k = 1$ else

• $\mathrm{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$

weighted similarity

$$similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$$

weighted distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

* weights must be non-negative