MEASURES

BEIYU LIN

DATA QUALITY

- Data quality problems:
  - Noise and outliers (example in yellow box)
  - Missing values (in red box) – Impute missing value
  - Duplicate data (in green box)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

SIMILARITY AND DISSIMILARITY MEASURES

Similarity measure

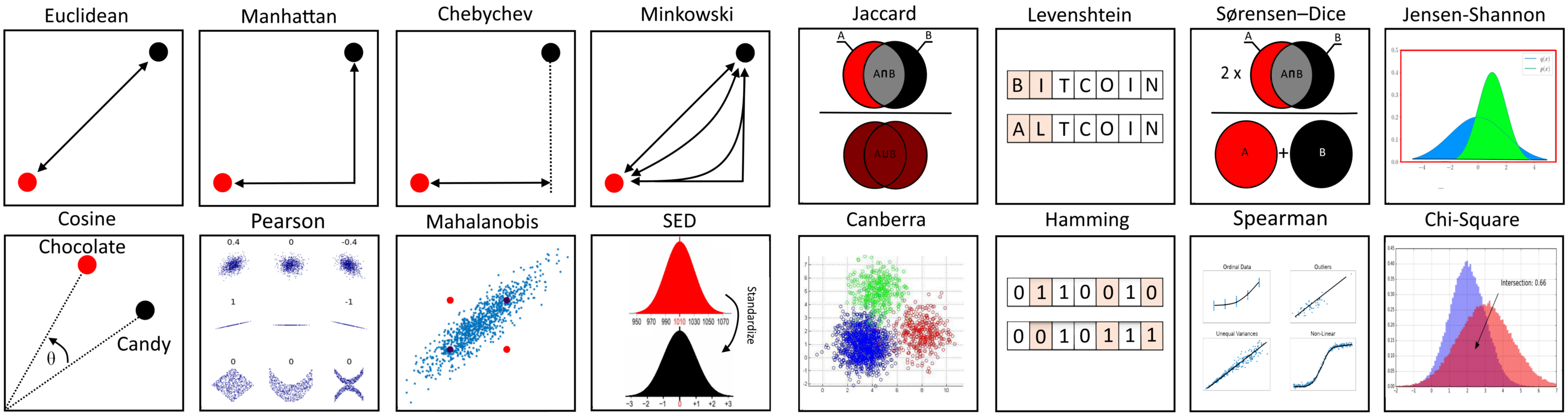
- how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity measure

- how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

SIMILARITY AND DISSIMILARITY MEASURES

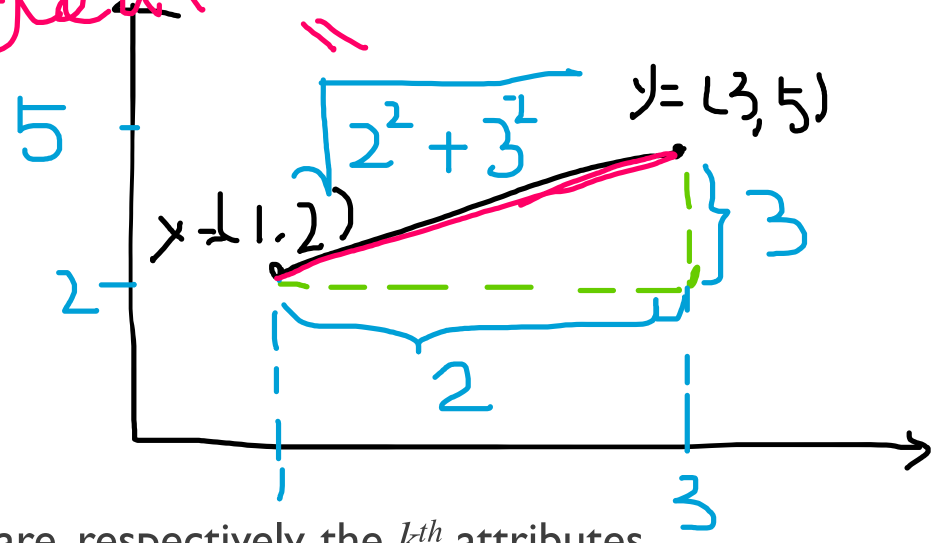


EUCLIDEAN DISTANCE

Euclidean

- Euclidean Distance

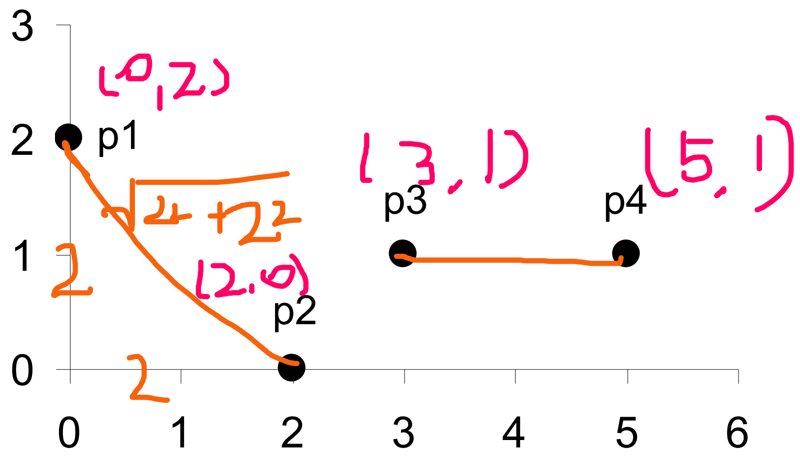
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$



where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

Standardization is necessary, if scales differ.

EUCLIDEAN DISTANCE



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

$d(p_2, p_1) = 2\sqrt{2}$
$= 2 * 1.4$
$= 2.8$

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

$d(p_3, p_4) = \sqrt{(3-5)^2 + (1-1)^2} =$

Euclidean Distance Matrix

EUCLIDEAN DISTANCE

- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

MINKOWSKI DISTANCE

- Minkowski Distance is a generalization of Euclidean Distance

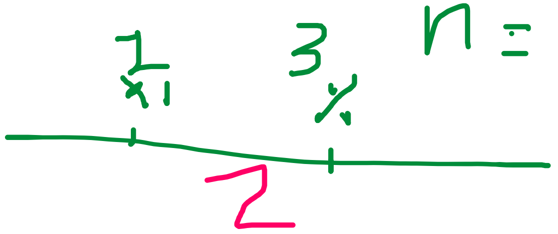
r v.s. n

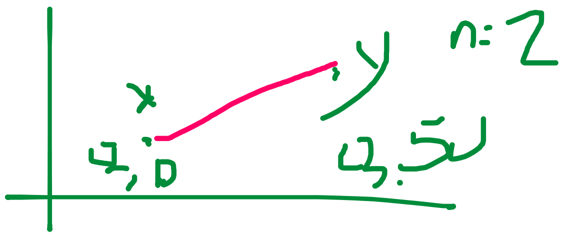
r = measure

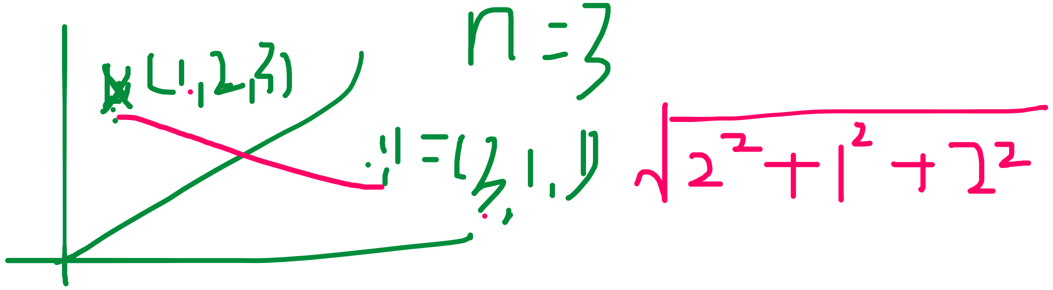
r=2

n = dimension of object

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

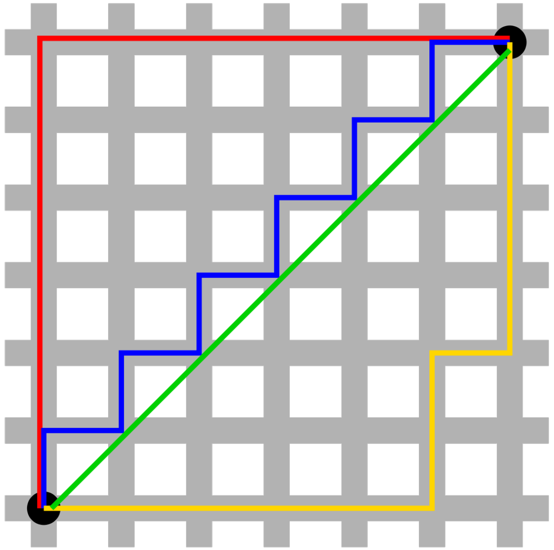


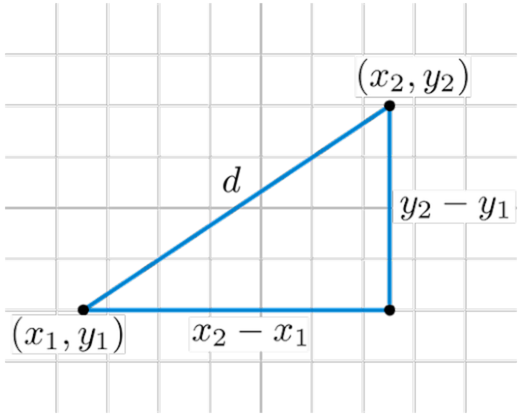




MINKOWSKI DISTANCE: EXAMPLES

- r = 1. City block (Manhattan, taxicab, L_1 norm) distance.

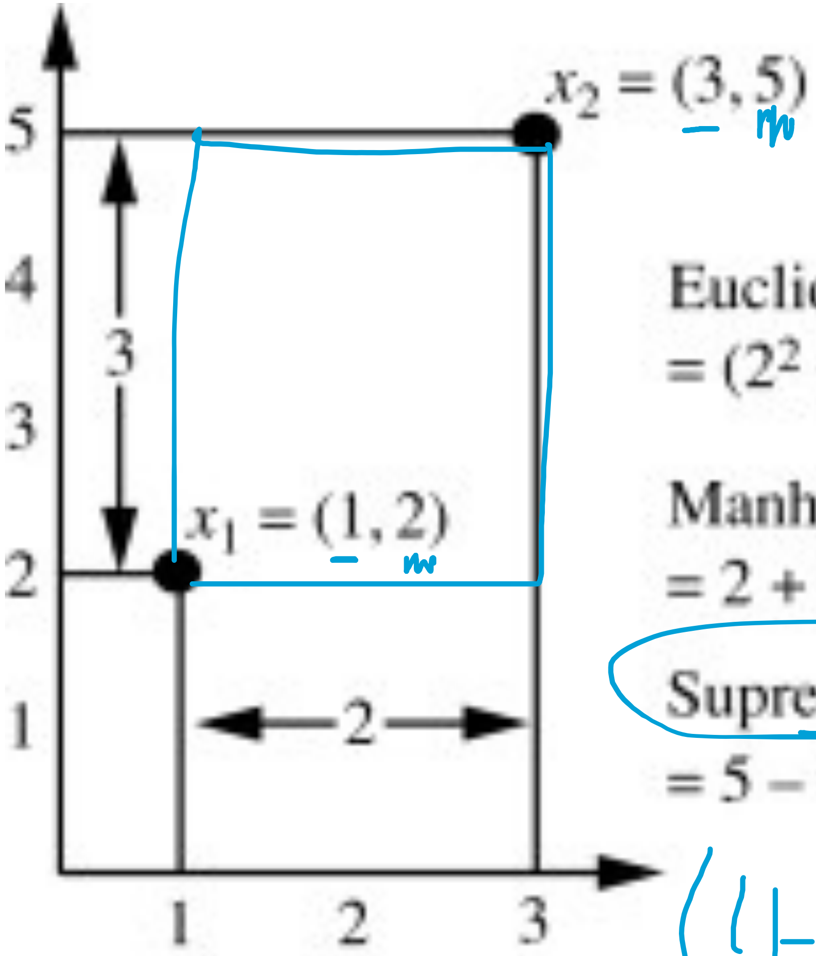




- r = 2. Euclidean distance

- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.

EXAMPLE FOR DIFFERENT DISTANCES



Euclidean distance $\Rightarrow r = 2$

$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance $=$ city block $r = 1$

$= 2 + 3 = 5$

Supremum distance $r = \infty = \max\{2, 3\} = 3$

$= 5 - 2 = 3$

$\left((1-3)^3 + (2-5)^3\right)^{\frac{1}{3}}$ $r = 3$

MINKOWSKI DISTANCE

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

n = 2
r = 1

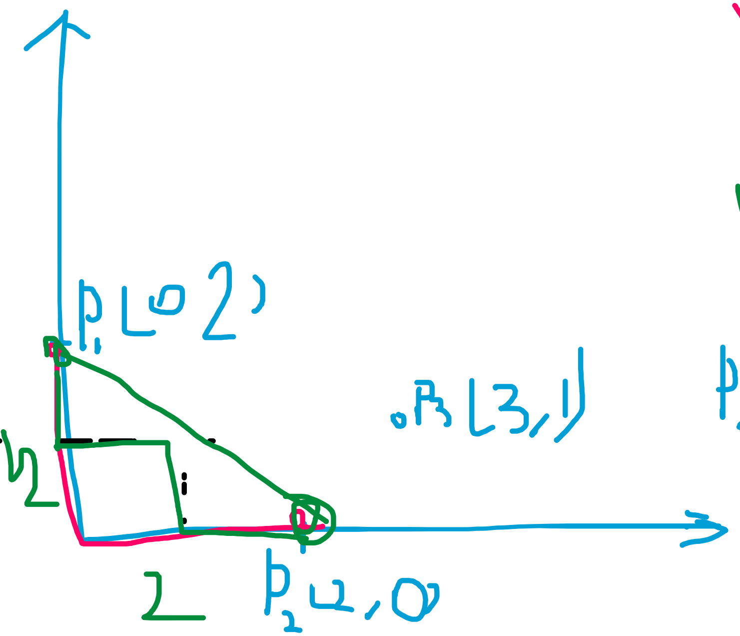
GPS
traj (lat, long)

$d(p_2, p_1) = 4 = d(p_1, p_2)$

$r = 2 \quad d = 2\sqrt{2}$

$r = \infty \quad d = \max$

$p_4(5,1) \quad \{|0-2|, |2-0|\}$
$= 2$



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

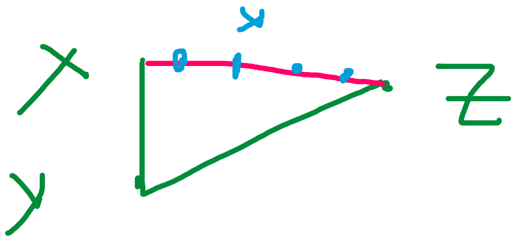
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

COMMON PROPERTIES OF A DISTANCE

- Properties of distances / dissimilarity, d(x, y), between points x and y.

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all *x* and *y* and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all **x** and **y**. (Symmetry)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points **x**, **y**, and **z**. (Triangle Inequality)



COMMON PROPERTIES OF A SIMILARITY

- Properties of Similarities

  1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.
  2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

SIMILARITY BETWEEN BINARY VECTORS

- If objects / data points, **x** and **y**, have only binary attributes

How many time

- Compute similarities using the following quantities

f_{01} = the number of attributes where **x** was 0 and **y** was 1

$x=0$

f_{10} = the number of attributes where **x** was 1 and **y** was 0

$y=1$

f_{00} = the number of attributes where **x** was 0 and **y** was 0

f_{11} = the number of attributes where **x** was 1 and **y** was 1

- Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

mutual presences $= f_{11}$

+ mutual absence $= f_{00}$

J = number of 11 matches / number of non-zero attributes

$= (f_{11}) / (f_{01} + f_{10} + f_{11})$

mutual presences $= f_{11}$

SMC VERSUS JACCARD: EXAMPLE

$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$f_{01} = 2$ (the number of attributes where **x** was 0 and **y** was 1)

$f_{10} = 1$ (the number of attributes where **x** was 1 and **y** was 0)

$f_{00} = 7$ (the number of attributes where **x** was 0 and **y** was 0)

$f_{11} = 0$ (the number of attributes where **x** was 1 and **y** was 1)

$$\text{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$