Data Mining :: Unit-2

(Distances and Similarity Measures)

Er. Dinesh Baniya Kshatri
(Lecturer)

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

Distance and Similarity

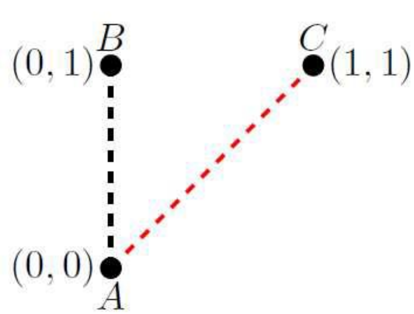
- Distance / Dissimilarity
  - Quantify the difference of two objects
  - The value is usually in the interval $[0, \infty]$
  - Lower values mean that the objects are more similar
- Similarity
  - Quantify the alikeness of two objects
  - The value is usually in the interval $[0, 1]$
  - Lower values mean that the objects are less similar

Usefulness of Distance and Similarity

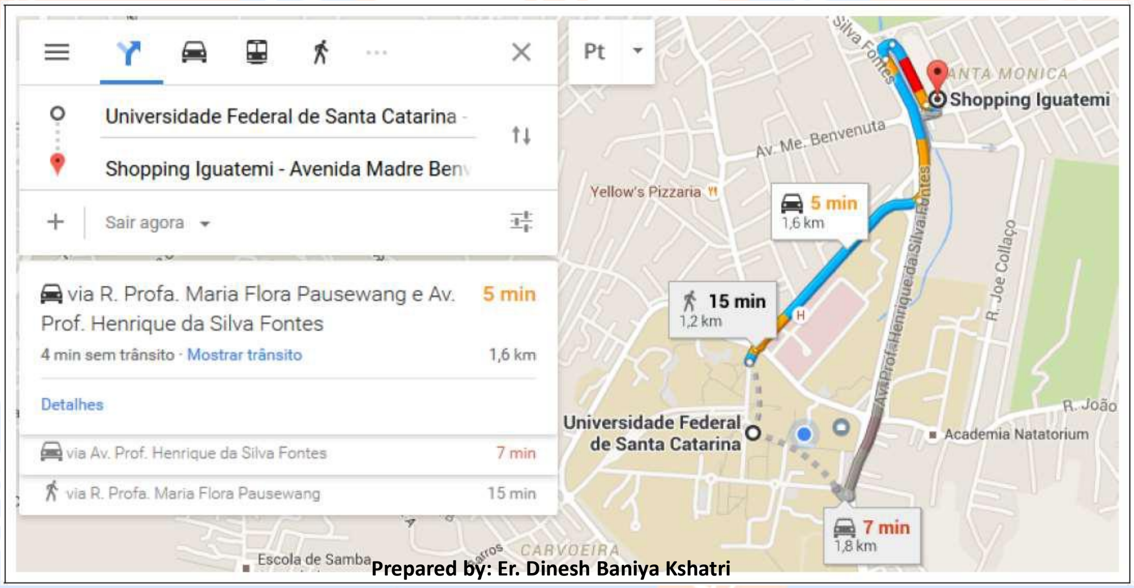
- Distance and Similarity measures are useful for several applications:
  - Calculate the distance between two points in a plane
  - Calculate the distance between two locations
  - Find the restaurants that are near a location
  - Search systems (e.g., a search in Google)
  - Given an image return the most similar images (e.g., Google Images)
  - Identify similar customers in a company database
  - ...

Distance between two points in a plane

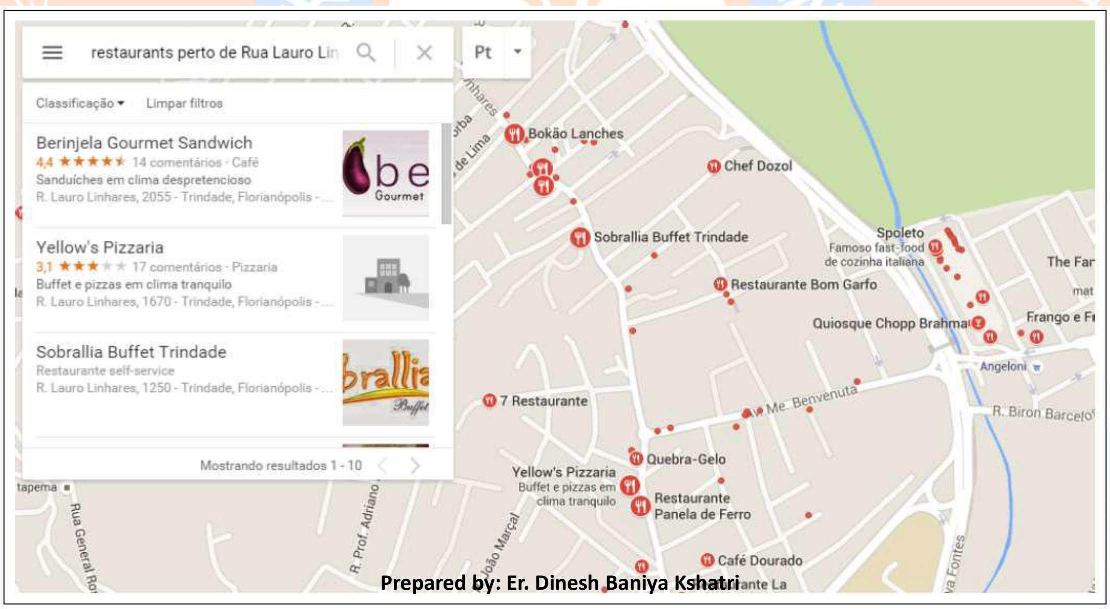
- Distance between two points in a plane
- Euclidean Distance:
  - $d(A, B) = 1$
  - $d(A, C) = \sqrt{2}$



Distance between Two Locations



Restaurants near a Location



Textual Similarity

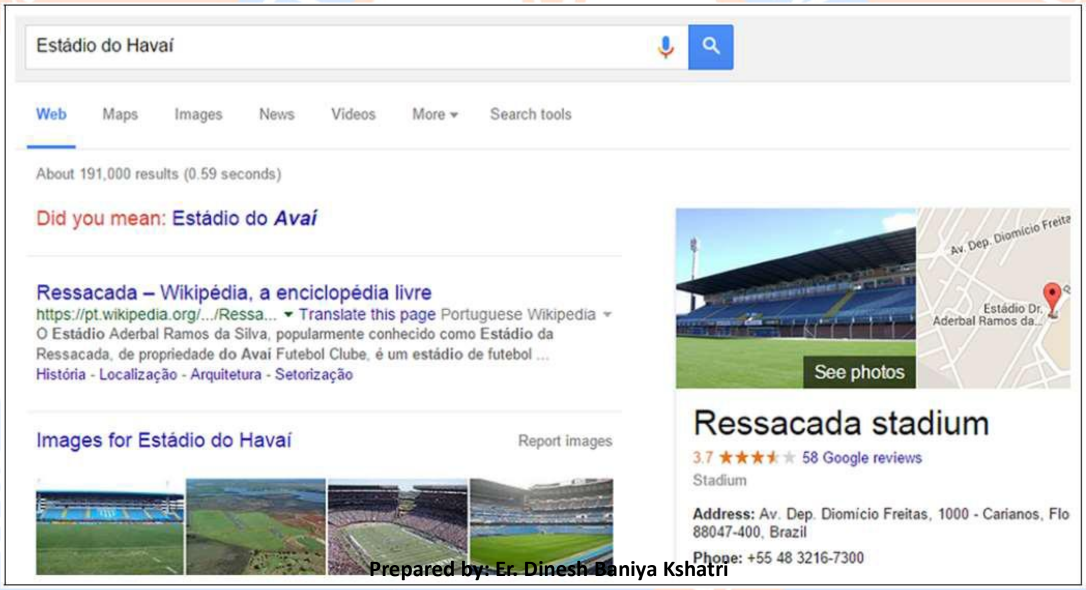
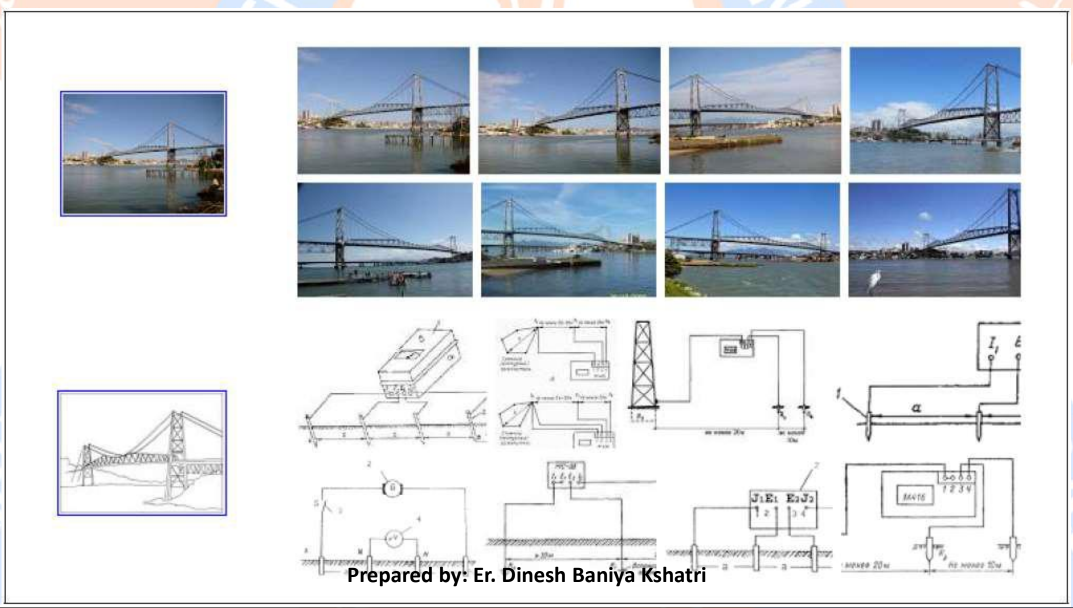


Image Similarity



Applications – Similarity and Distance

- For many different problems we need to quantify how close two objects are.
- Examples:
  - For an item bought by a customer, find other similar items
  - Group together the customers of a site so that similar customers are shown the same ad.
  - Group together web documents so that you can separate the ones that talk about politics and the ones that talk about sports.
  - Find all the near-duplicate mirrored web documents.
  - Find credit card transactions that are very different from previous transactions.
- To solve these problems we need a definition of similarity, or distance.
  - The definition depends on the type of data that we have

Distance

- Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Higher when two objects are different
- Minimum distance is 0, when comparing an object with itself.
- Upper limit varies

Distance Metric

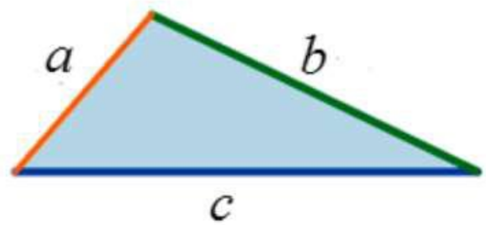
- A distance function d is a distance metric if :
  1. $d(x,y) \geq 0$. (non-negativity)
  2. $d(x,y) = 0$ iff $x = y$. (identity)
  3. $d(x,y) = d(y,x)$. (symmetry)
  4. $d(x,y) \leq d(x,z) + d(z,y)$ (triangle inequality).

Triangle Inequality

- Triangle inequality guarantees that the distance function is well-behaved.
  - The direct connection is the shortest distance
- It is useful also for proving useful properties about the data.

Triangle Inequality Theorem

The sum of the lengths of any two sides of a triangle is greater than the length of the third side.



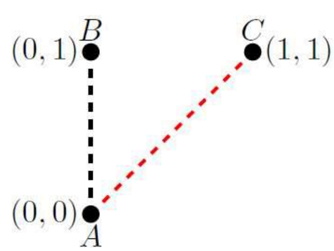
$$a + b > c$$

$$a + c > b$$

$$b + c > a$$

Euclidean Distance (A,B)

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$



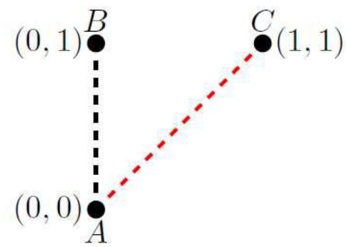
$$d(A,B) = \sqrt{(A.x - B.x)^2 + (A.y - B.y)^2} \qquad \text{(ED1)}$$

$$d(A,B) = \sqrt{(0-0)^2 + (0-1)^2} \qquad \text{(ED2)}$$

$$d(A,B) = \sqrt{(0)^2 + (1)^2} = 1 \qquad \text{(ED3)}$$

Euclidean Distance (A,C)

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$



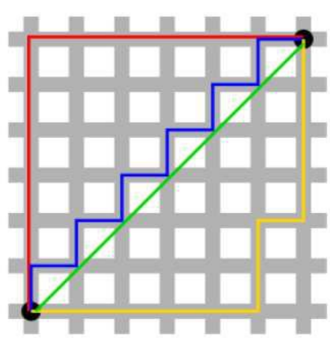
$$d(A,C) = \sqrt{(A.x - C.x)^2 + (A.y - C.y)^2} \quad \text{(ED4)}$$

$$d(A,C) = \sqrt{(0-1)^2 + (0-1)^2} \quad \text{(ED5)}$$

$$d(A,C) = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} \quad \text{(ED6)}$$

Manhattan Distance

- Absolute distance of the coordinates
- Also known as Taxicab Distance, City Block Distance...

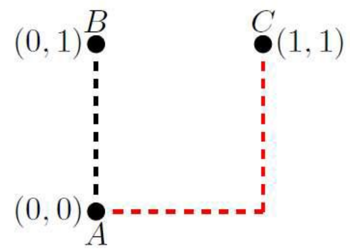


- Given two points p and q:

$$d(p,q) = |p.x - q.x| + |p.y - q.y|$$

Manhattan Distance (A,B)

$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$



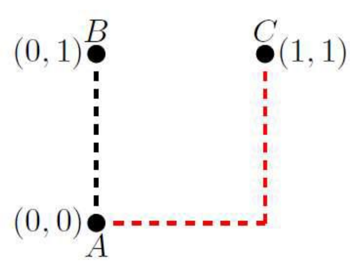
$$d(A, B) = |A.x - B.x| + |A.y - B.y| \quad \text{(MD1)}$$

$$d(A, B) = |0 - 0| + |0 - 1| \quad \text{(MD2)}$$

$$d(A, B) = 0 + 1 = 1 \quad \text{(MD3)}$$

Manhattan Distance (A,C)

$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$



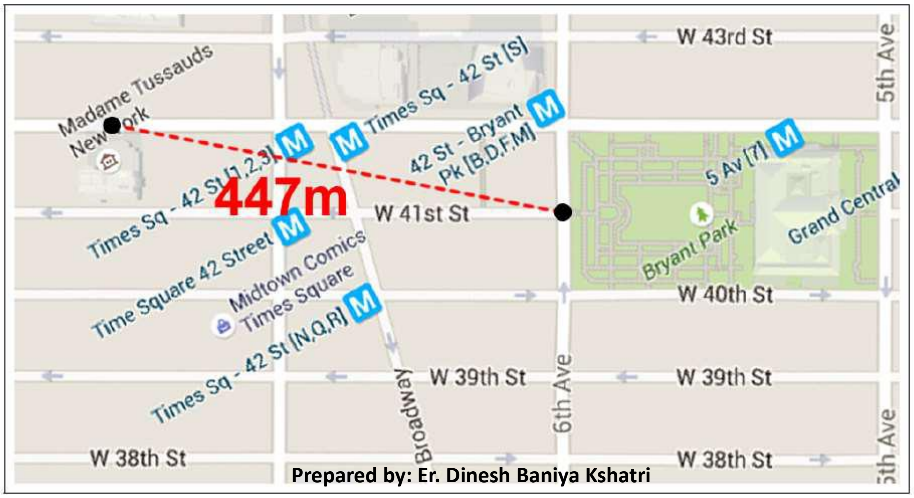
$$d(A, C) = |A.x - C.x| + |A.y - C.y| \quad \text{(MD4)}$$

$$d(A, C) = |0 - 1| + |0 - 1| \quad \text{(MD5)}$$

$$d(A, C) = 1 + 1 = 2 \quad \text{(MD6)}$$

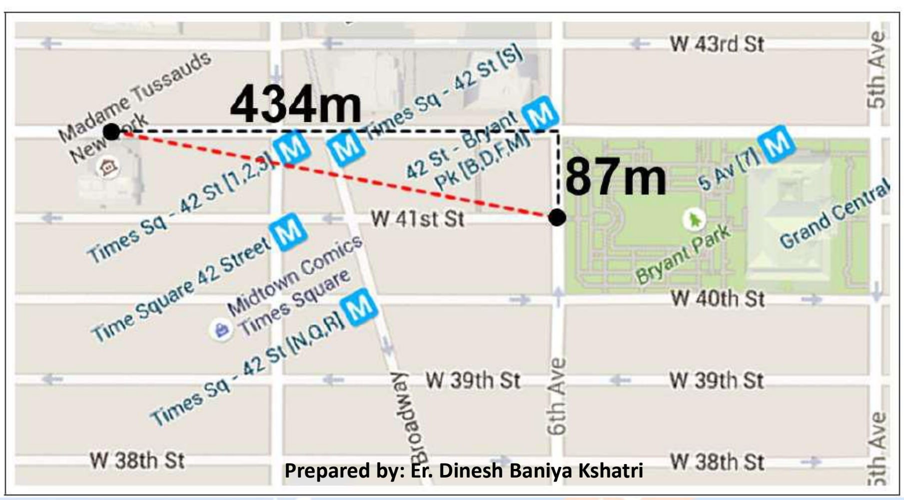
Euclidean Distance in Manhattan

- Euclidean: $d(BryantPark, MadameTussaud) = 447m$



Manhattan Distance in Manhattan

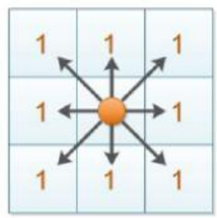
- Manhattan:
  $d(BryantPark, MadameTussaud) = 434m + 87m = 521m$



Chebyshev Distance

- Maximum difference in any coordinate
- Also know as Chessboard Distance
- Given two points p and q:

$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|)$$



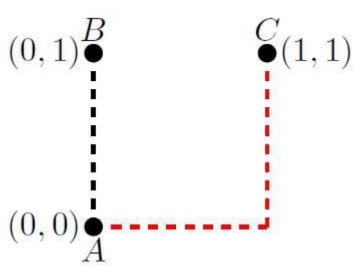
$\max(|x_1 - x_2|, |y_1 - y_2|)$

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

Prepared by: Er. Dinesh Baniya Kshatri 21

Chebyshev Distance (A,C)

$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|)$$



$$d(A,C) = max(|A.x - C.x|, |A.y - C.y|) \quad \text{(CD1)}$$

$$d(A,C) = max(|0-1|, |0-1|) \quad \text{(CD2)}$$

$$d(A,C) = max(1,1) = 1 \quad \text{(CD3)}$$

Prepared by: Er. Dinesh Baniya Kshatri 22

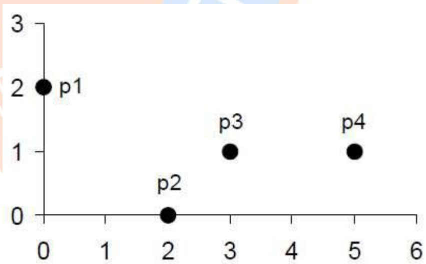
Euclidean Distance in "n" Dimensions

- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects p and q.

Euclidean Distance – Example



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

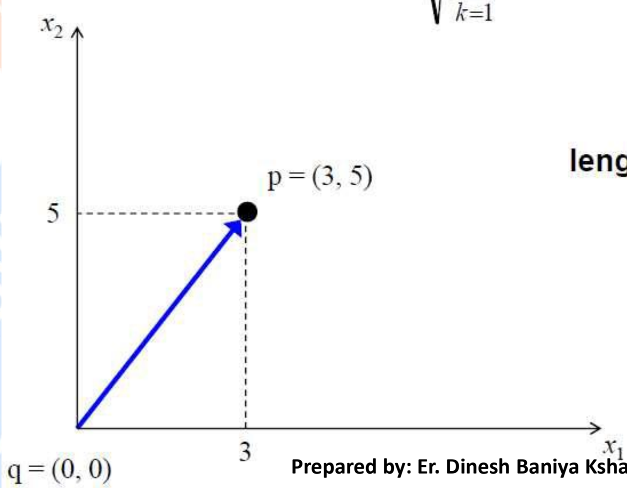
	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

More about Euclidean Distance

$$dist(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} = \sqrt{\sum_{k=1}^{n} p_k^2} = \|p\|$$

length of vector p



Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where *r* is a parameter, *n* is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects *p* and *q*.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

Prepared by: Er. Dinesh Baniya Kshatri 27

Distances for Vectors

- Vectors $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$
- L_p-norms or Minkowski distance:

$$L_p(x, y) = \left[|x_1 - y_1|^p + \dots + |x_d - y_d|^p\right]^{1/p}$$

- L_2-norm: Euclidean distance:

$$L_2(x, y) = \sqrt{|x_1 - y_1|^2 + \dots + |x_d - y_d|^2}$$

- L_1-norm: Manhattan distance:

$$L_1(x, y) = |x_1 - y_1| + \dots + |x_d - y_d|$$

- L_∞-norm:

L_p norms are known to be distance metrics

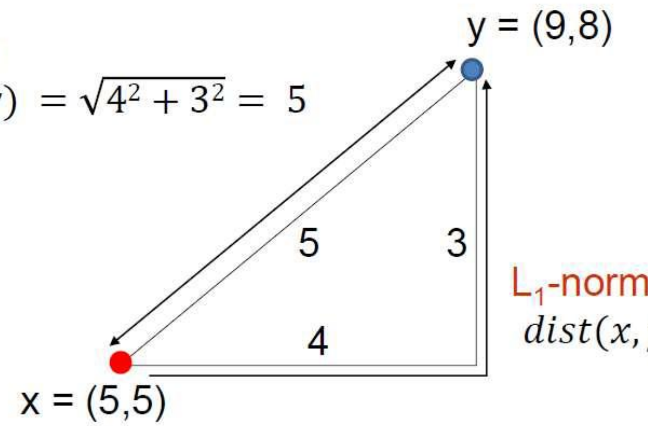
$$L_\infty(x, y) = \max\{|x_1 - y_1|, \dots, |x_d - y_d|\}$$

  - The limit of L_p as p goes to infinity.

Prepared by: Er. Dinesh Baniya Kshatri 28

Examples of Distances

L$_2$-norm:
$dist(x, y) = \sqrt{4^2 + 3^2} = 5$

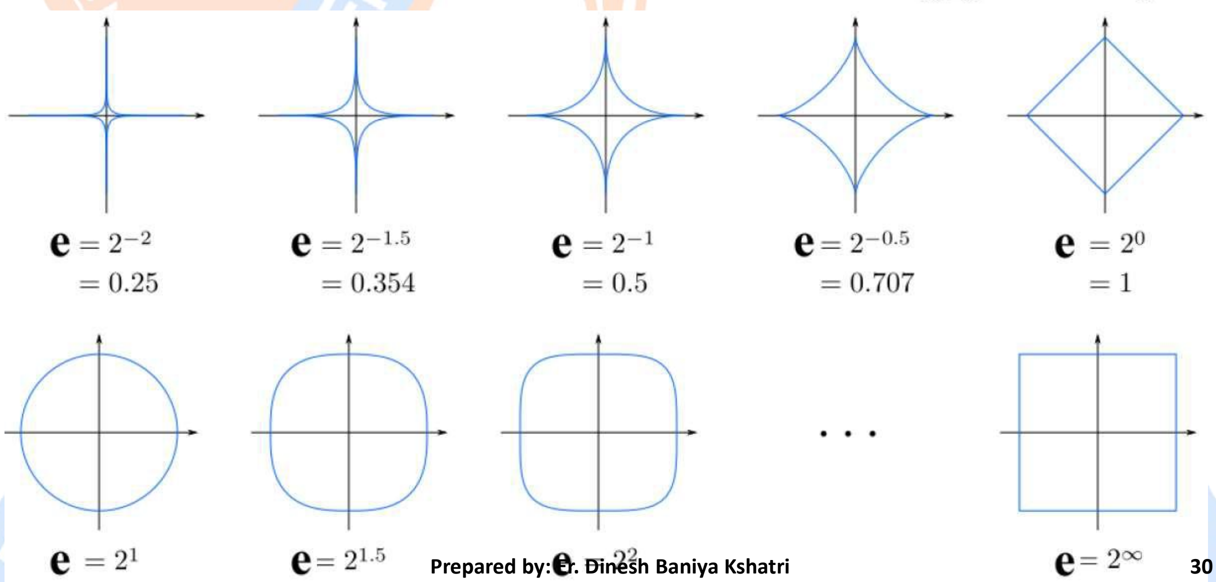


L$_1$-norm:
$dist(x, y) = 4 + 3 = 7$

L$_\infty$-norm:
$dist(x, y) = \max\{3,4\} = 4$

Minkowski Distance

$$d(p, q) = \left(\sum_{i=1}^{n} |p_i - q_i|^e\right)^{1/e}$$



Minkowski Distance – Example

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

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

Hamming Distance

- Hamming distance is the number of positions in which bit-vectors differ.
  - Example: p_1 = 10101
    p_2 = 10011.
    - $d(p_1, p_2) = 2$ because the bit-vectors differ in the 3rd and 4th positions.
    - The L_1 norm for the binary vectors
- Hamming distance between two vectors of categorical attributes is the number of positions in which they differ.
  - Example: x = (married, low income, cheat),
    y = (single, low income, not cheat)
  - d(x,y) = 2

Hamming Distance – Example

- Example : The Hamming distance between:
  - "karolin" and "kathrin" is 3.
  - "karolin" and "kerstin" is 3.
  - 1011101 and 1001001 is 2.
  - 2173896 and 2233796 is 3.
- It is used in telecommunication to count the number of flipped bits in a fixed-length binary word as an estimate of error, and therefore is sometimes called the **signal distance**.

Edit Distance for strings

- The edit distance of two strings is the number of inserts and deletes of characters needed to turn one into the other.
- Example: x = abcde ; y = bcduve.
  - Turn *x* into *y* by deleting a, then inserting u and v after d.
  - Edit distance = 3.
- Common distance measure for comparing DNA sequences

Levenshtein Distance

- Also known as Edit Distance
  - Example $d(Avaí, ?)$:
    - Avaí
    - **H**avaí
    - **Haw**aii
  - Results for $d(Avaí, ?)$:
    - $d(Avaí, Avaí) = 0$
    - $d(Avaí, Havaí) = 2$
    - $d(Avaí, Hawaii) = 5$
  - Transform Avaí into Hawaii:
    - 1 - Add H in the beggining (**H**Avaí)
    - 2 - Replace A for a (**Ha**vaí)
    - 3 - Replace v for w (**Haw**aí)
    - 4 - Replace $í$ for i (**Hawai**)
    - 5 - Add i in the end (**Hawaii**)
  - Result for $d(Avaí, Hawaii) = 5$

Similarity

- Similarities, also have some well known properties.

  1. *s(p, q) = 1* (or maximum similarity) only if *p* = *q*.
  2. *s(p, q)* = *s(q, p)* for all *p* and *q*. (Symmetry)

  where *s(p, q)* is the similarity between points (data objects), *p* and *q*.

- Often falls in the range [0,1], sometimes in [-1,1]

Similarity between Sets

- Consider the following documents

apple releases new ipod	apple releases new ipad	new apple pie recipe

- Which ones are more similar?
- How would you quantify their similarity?

Similarity : Intersection

- Number of words in common

apple releases new ipod	apple releases new ipad	new apple pie recipe

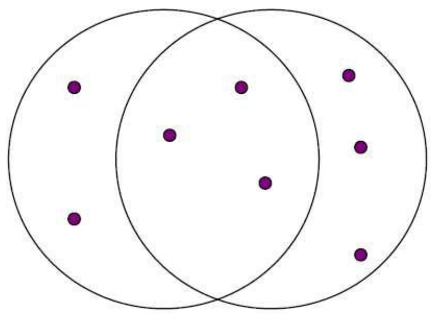
- Sim(DY,DR) = 3, Sim(DR,DB) = Sim(DY,DB) =2
- What about this document?

Vefa releases new book with apple pie recipes

- Sim(DR,DG) = Sim(DY,DG) = 3

Jaccard Similarity

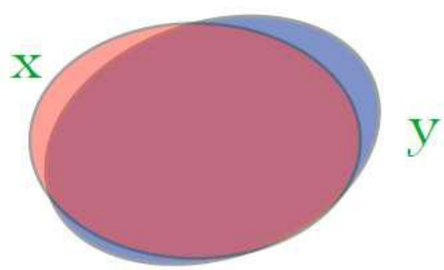
- The Jaccard similarity (Jaccard coefficient) of two sets S_1, S_2 is the size of their intersection divided by the size of their union.
  - JSim $(S_1, S_2) = |S_1 \cap S_2| / |S_1 \cup S_2|$.



3 in intersection.
8 in union.
Jaccard similarity
= 3/8

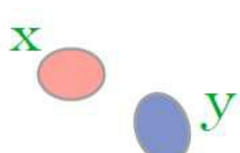
Jaccard Similarity – Extreme Cases

- Case 1 (very large almost identical documents)



$J(x, y)$ almost 1

- Case 2 (small disjoint documents)



$J(x, y) = 0$

Jaccard Similarity between sets

apple releases new ipod	apple releases new ipad	new apple pie recipe	Vefa releases new book with apple pie recipes

- JSim(DY,DR) = 3/5
- JSim(DR,DB) = JSim(DY,DB) = 2/6
- JSim(DR,DG) = JSim(DY,DG) = 3/9

Prepared by: Er. Dinesh Baniya Kshatri 41

Jaccard Similarity – Example

- $sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$
- Example Sets
  - $A = \{Giraffe, Monkey, Elephant, Bird\}$
  - $B = \{Monkey, Crocodile\}$
  - $C = \{Horse, Dog, Parrot\}$
  - $D = \{Monkey\}$
- Results for $sim(A,?)$:
  - $sim(A,A) = \frac{4}{4} = 1$
  - $sim(A,B) = \frac{1}{5} = 0.2$
  - $sim(A,C) = \frac{0}{7} = 0$
  - $sim(A,D) = \frac{1}{4} = 0.25$

Prepared by: Er. Dinesh Baniya Kshatri 42

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

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

  J = number of 11 matches / number of not-both-zero attributes values

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

Prepared by: Er. Dinesh Baniya Kshatri 43

SMC vs. Jaccard Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

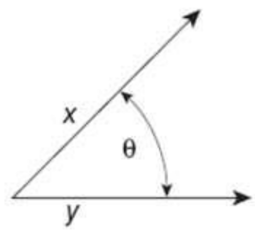
$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

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

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

Prepared by: Er. Dinesh Baniya Kshatri 44

Cosine Similarity

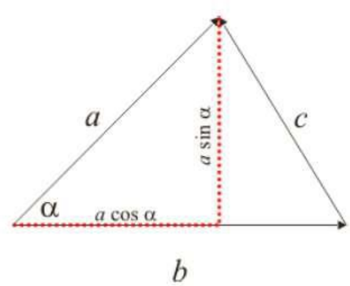


- Sim(X,Y) = cos(X,Y)
  - The cosine of the angle between X and Y
- If the vectors are aligned (correlated) angle is zero degrees and cos(X,Y)=1
- If the vectors are orthogonal (no common coordinates) angle is 90 degrees and cos(X,Y) = 0

Derivation of Cosine Similarity

Proof:

$$
\begin{aligned}
c^2 &= (b - a\cos\alpha)^2 + (a\sin\alpha)^2 \\
&= b^2 - 2ab\cos\alpha + a^2\cos^2\alpha + a^2\sin^2\alpha \\
&= a^2 + b^2 - 2ab\cos\alpha
\end{aligned}
$$



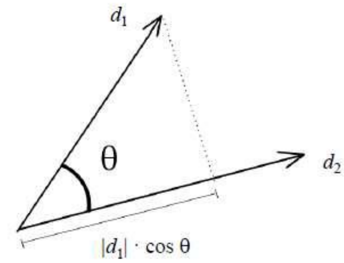
$$
\begin{aligned}
c^2 &= \vec{c}\cdot\vec{c} \\
&= (\vec{a}-\vec{b})\cdot(\vec{a}-\vec{b}) \\
&= \vec{a}\cdot\vec{a} - 2\vec{a}\cdot\vec{b} + \vec{b}\cdot\vec{b} \\
&= a^2 - 2\vec{a}\cdot\vec{b} + b^2
\end{aligned}
$$

$$\cos\alpha = \frac{\vec{a}\cdot\vec{b}}{ab}$$

By combining previous equations we get:

$$a^2 + b^2 - 2ab\cos\alpha = a^2 - 2\vec{a}\cdot\vec{b} + b^2$$

Cosine Simularity



- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|}$$

where • indicates vector dot product and $\|d\|$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42)$^{0.5}$ = 6.481
$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)$^{0.5}$ = (6)$^{0.5}$ = 2.245

cos(d_1, d_2) = .3150

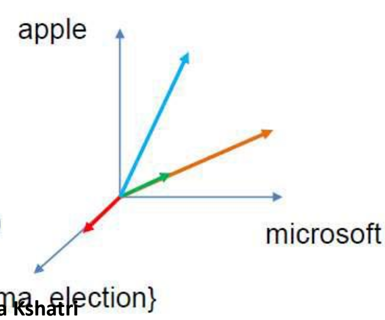
Example – Cosine Similarity

document	Apple	Microsoft	Obama	Election
D1	10	20	0	0
D2	30	60	0	0
D3	60	30	0	0
D4	0	0	10	20

Cos(D1,D2) = 1

Cos (D3,D1) = Cos(D3,D2) = 4/5

Cos(D4,D1) = Cos(D4,D2) = Cos(D4,D3) = 0



Correlation Coefficient

- The correlation coefficient measures correlation between two random variables.
- If we have observations (vectors) $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$

$$CorrCoeff(X,Y) = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_i (x_i - \mu_X)^2}\sqrt{\sum_i (y_i - \mu_Y)^2}}$$

- This is essentially the cosine similarity between the normalized vectors (where from each entry we remove the mean value of the vector.
- The correlation coefficient takes values in [-1,1]
  - -1 negative correlation, +1 positive correlation, 0 no correlation.

Correlation Coefficient Example

Normalized vectors

document	Apple	Microsoft	Obama	Election
D1	-5	+5	0	0
D2	-15	+15	0	0
D3	+15	-15	0	0
D4	0	0	-5	+5

$$CorrCoeff(X,Y) = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_i (x_i - \mu_X)^2}\sqrt{\sum_i (y_i - \mu_Y)^2}}$$

CorrCoeff(D1,D2) = 1

CorrCoeff(D1,D3) = CorrCoeff(D2,D3) = -1

CorrCoeff(D1,D4) = CorrCoeff(D2,D4) = CorrCoeff(D3,D4) = 0

Correlation and Covariance

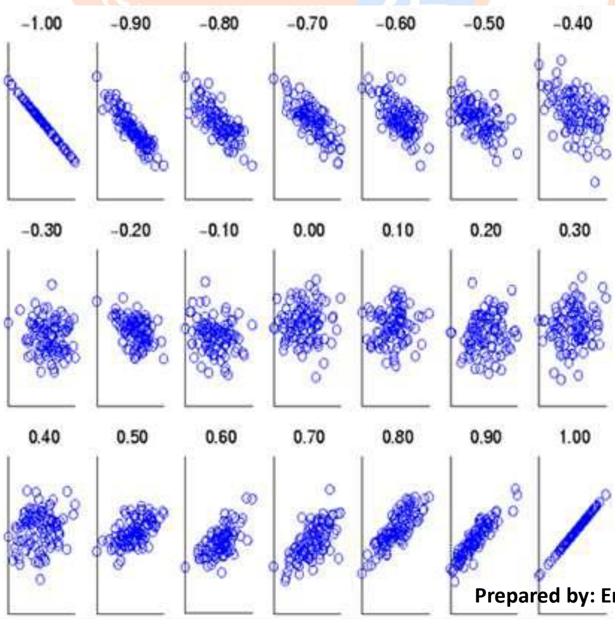
$$\text{corr}(\mathbf{x},\mathbf{y}) = \frac{\text{covariance}(\mathbf{x},\mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}$$

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

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

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

Visually Evaluating Correlation



- **Covariance – A measure of how much two variables change together**
  - Positive covariance = Variables tend to move in same direction
  - Negative covariance = Variables tend to move in opposite direction
- **Correlation – Scaled version of covariance**

Scatter plots showing the similarity from –1 to 1.