



# **Lecture 2-B: Data Processing**

# Review: Data Preprocessing

---

- a. Aggregation
- b. Sampling
- c. Dimensionality Reduction
- d. Feature subset selection
- e. Feature creation
- f. Discretization and Binarization
- g. Attribute Transformation

# Today's class

---

- Similarity and Disimilarity

# Similarity and Dissimilarity

---

- Similarity

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

- Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

- Proximity refers to a similarity or dissimilarity

# Similarity/Dissimilarity for Simple Attributes

$p$  and  $q$  are the attribute values for two data objects.

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

# Metrics

---

- Euclidean Distance
- Minkowski Distance
- Mahalanobis Distance
- Metric Theories

# Euclidean Distance

---

- Euclidean Distance

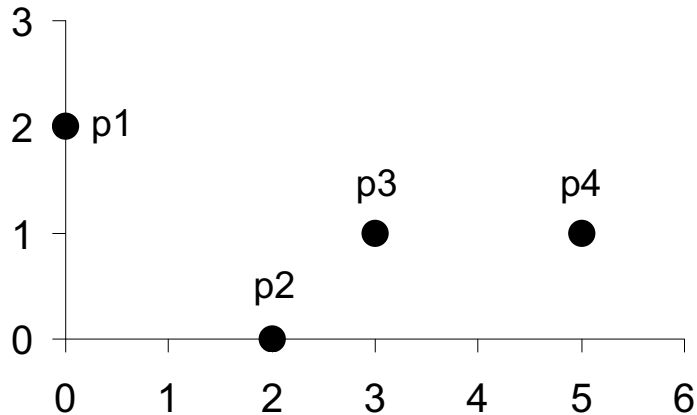
$$\mathit{dist} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Where  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k^{\text{th}}$  attributes (components) or data objects  $p$  and  $q$ .

- Standardization is necessary, if scales differ.

# Quiz: Euclidean Distance

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



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

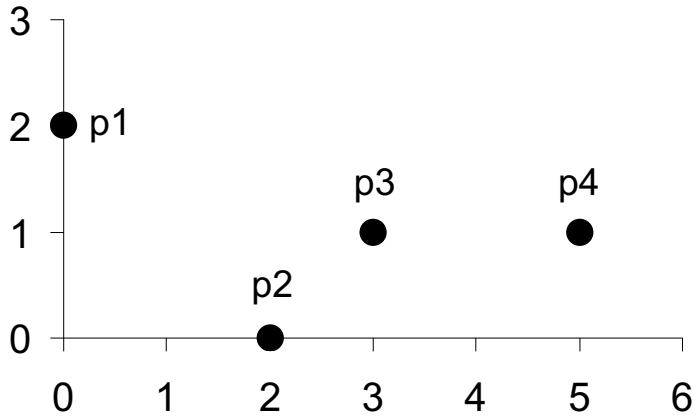
	p1	p2	p3	p4
p1				
p2				
p3				
p4				

**Distance Matrix**



# Quiz: Euclidean Distance

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



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**

# Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

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

Where  $r$  is a parameter,  $n$  is the number of dimensions (attributes) and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) or 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
- $r \rightarrow \infty$ . “supremum” ( $L_{\max}$  norm,  $L_{\infty}$  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.

# Minkowski Distance

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

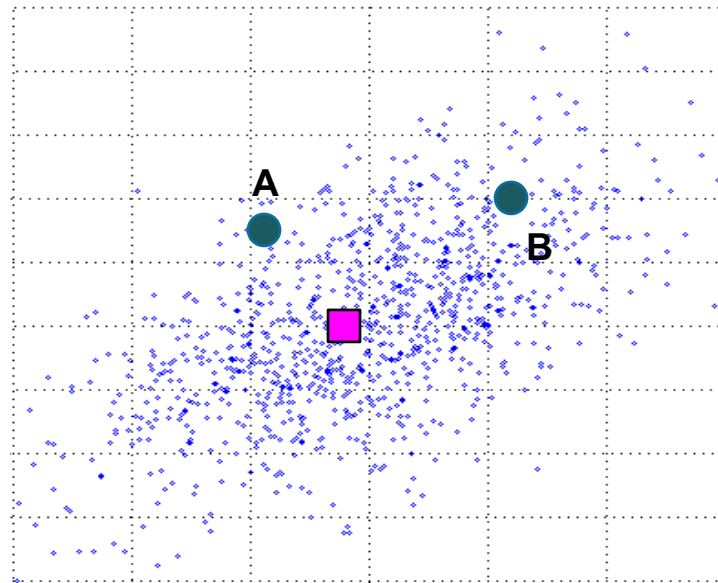
$L_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

## Distance Matrix

# Mahalanobis Distance

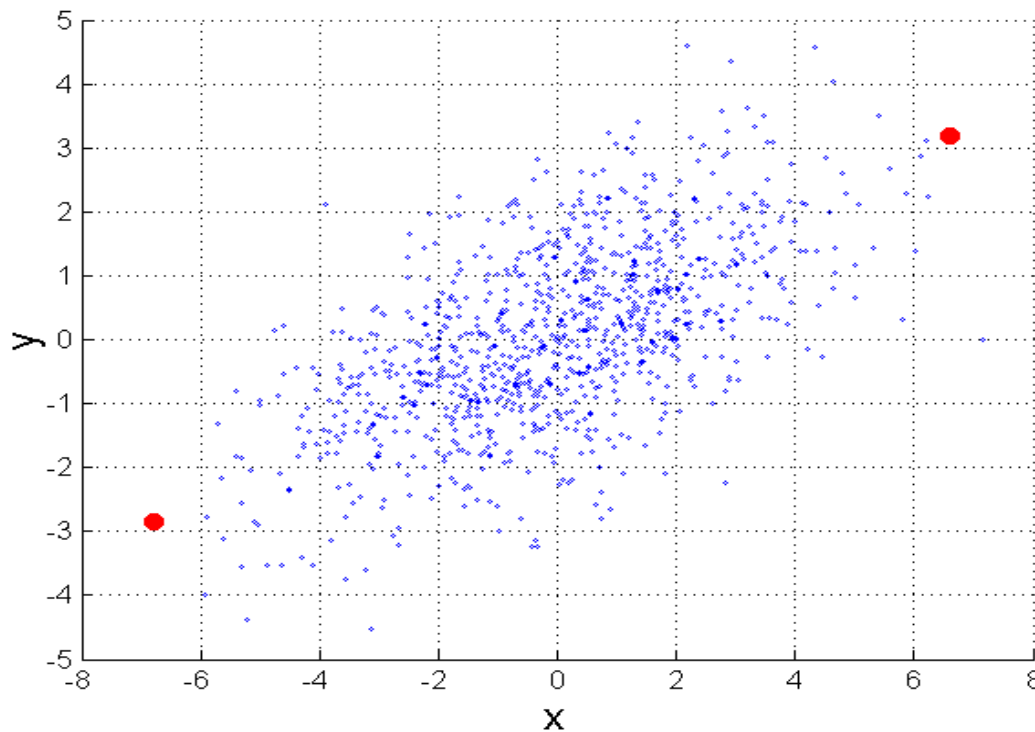
Consider the problem of estimating the probability that a test point in  $N$ -dimensional Euclidean space belongs to a set, where we are given sample points that definitely belong to that set.

Let's consider these two points A & B . What point one is least likely to be from the probability?



# Mahalanobis Distance

$$\text{mahalanobis}(p, q) = (p - q)^T \Sigma^{-1} (p - q)$$

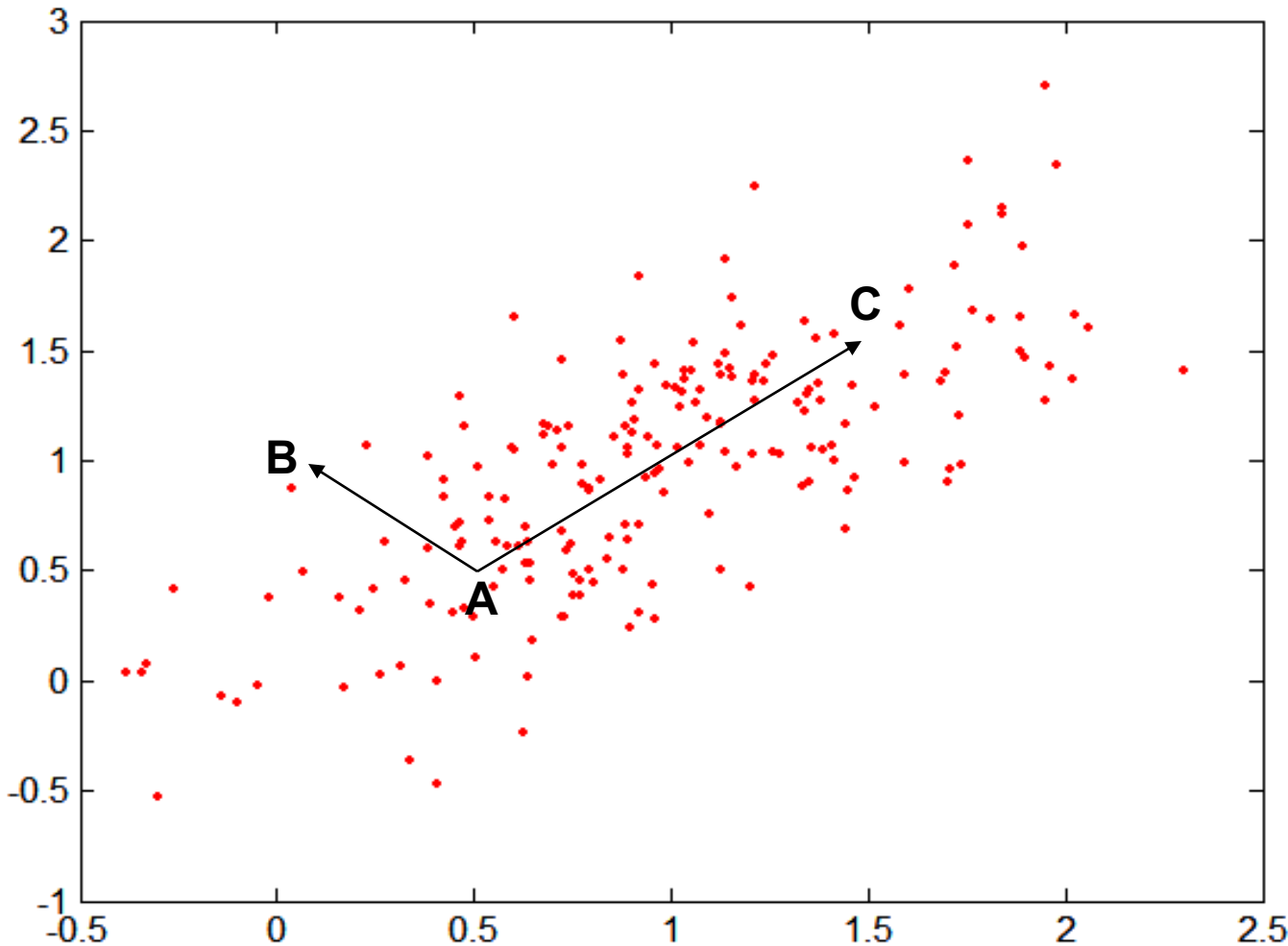


$\Sigma$  is the covariance matrix of the input data  $X$

$$\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

# Mahalanobis Distance



**Covariance Matrix:**

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

**A: (0.5, 0.5)**

**B: (0, 1)**

**C: (1.5, 1.5)**

**Mahal(A,B) = 5**

**Mahal(A,C) = 4**

# Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1.  $d(p, q) \geq 0$  for all  $p$  and  $q$  and  $d(p, q) = 0$  only if  $p = q$ . (Positive definiteness)
2.  $d(p, q) = d(q, p)$  for all  $p$  and  $q$ . (Symmetry)
3.  $d(p, r) \leq d(p, q) + d(q, r)$  for all points  $p, q$ , and  $r$ . (Triangle Inequality)

where  $d(p, q)$  is the distance (dissimilarity) between points (data objects),  $p$  and  $q$ .

- A distance that satisfies these properties is a **metric**



# Common Properties of a 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  $q$ .

# Cosine Similarity

- If  $d_1$  and  $d_2$  are two document vectors, then

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

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

- Example:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

# Correlation

---

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects,  $p$  and  $q$ , and then take their dot product

$$p'_k = (p_k - \text{mean}(p)) / \text{std}(p)$$

$$q'_k = (q_k - \text{mean}(q)) / \text{std}(q)$$

$$\text{correlation}(p, q) = p' \bullet q'$$

# General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the  $k^{th}$  attribute, compute a similarity,  $s_k$ , in the range  $[0, 1]$ .
2. Define an indicator variable,  $\delta_k$ , for the  $k^{th}$  attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

# Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights  $w_k$  which are between 0 and 1 and sum to 1.

$$\text{similarity}(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$\text{distance}(p, q) = \left( \sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$

# Summery

---

- Euclidean Distance
- Minkowski Distance
- Mahalanobis Distance
- Similarities
  - Cosine
  - Correlation
- Metric Theories
  - Distance
  - Similarity



# Density

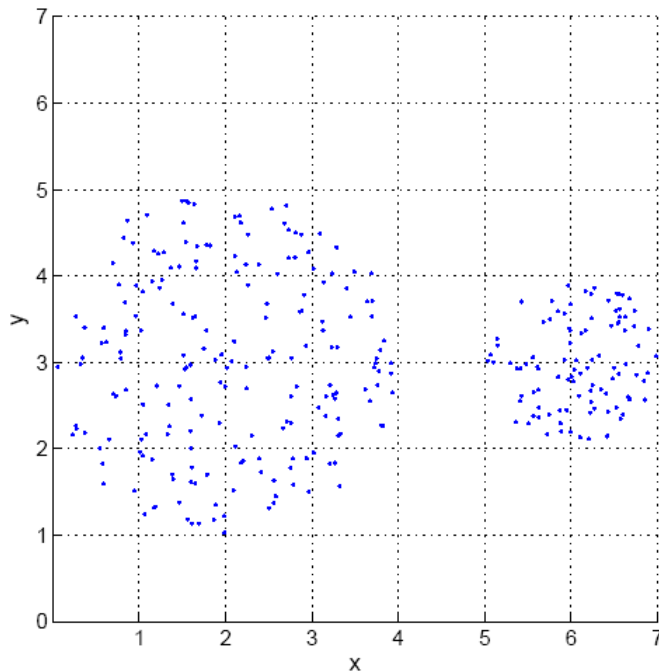
---

- Density-based clustering require a notion of density
- Examples:
  - Euclidean density
    - ◆ Euclidean density = number of points per unit volume
  - Probability density
  - Graph-based density



# Euclidean Density – Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains



**Figure 7.13.** Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

**Table 7.6.** Point counts for each grid cell.

# Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius of the point

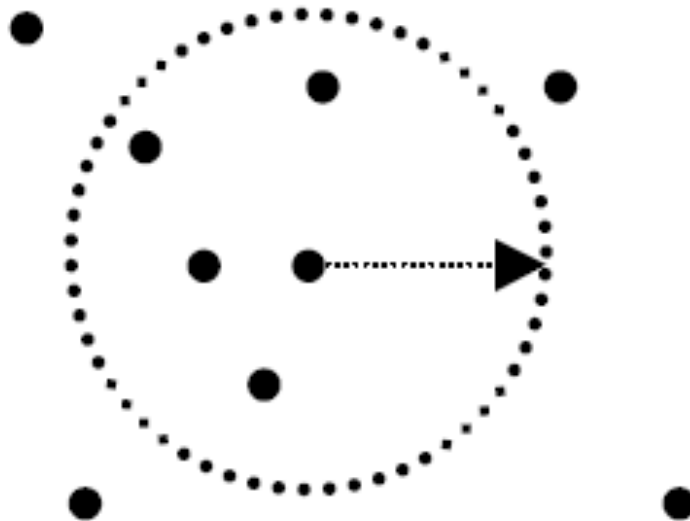


Figure 7.14. Illustration of center-based density.