GUC
German University in Cairo

Spring 2019

# CSEN1083: Data Mining

## *Data (3)*

Seif Eldawlatly

# Data Preprocessing: Feature Subset Selection

- Another way to reduce the dimensionality is to use only a subset of the features

- Redundant features: duplicate much or all of the information contained in one or more other attributes

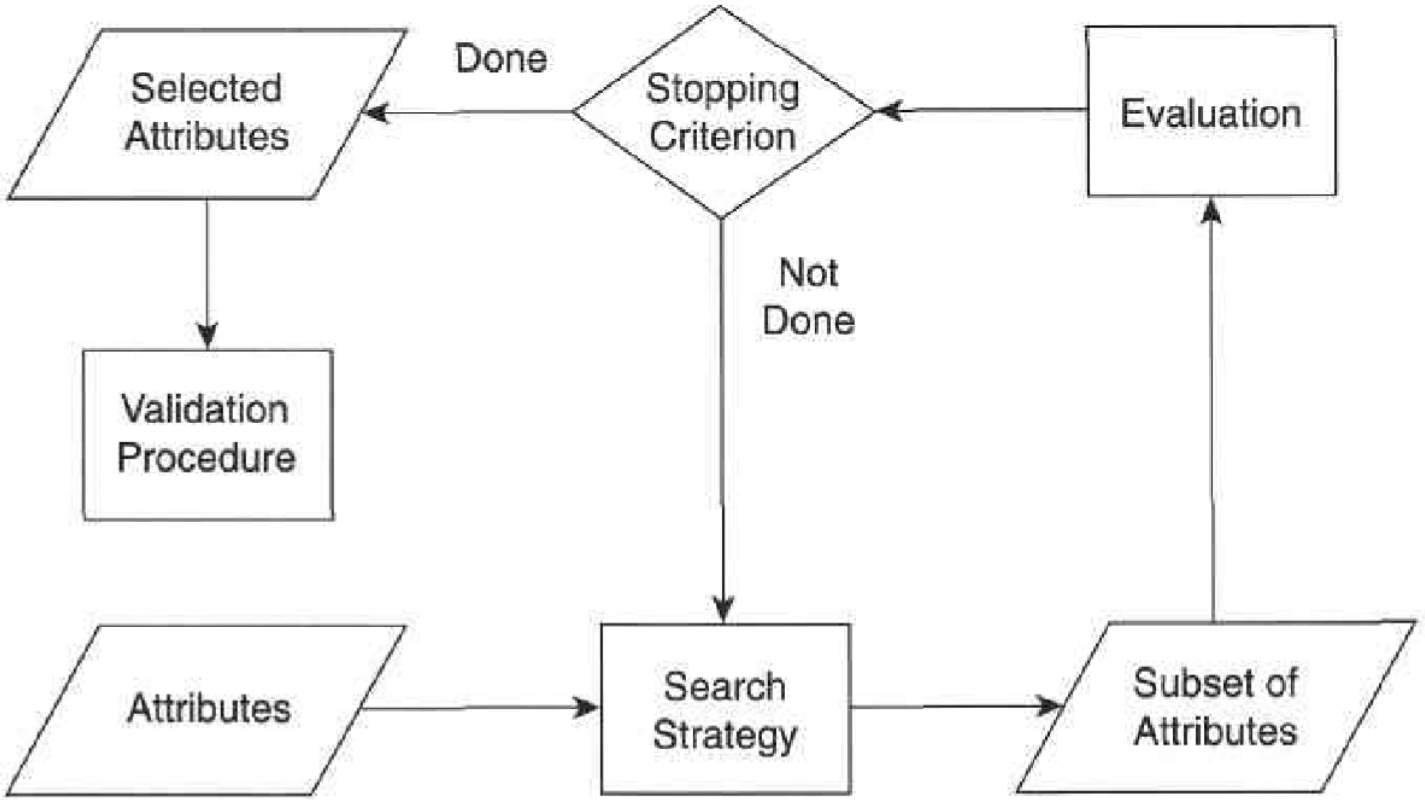
- Example: The purchase price of a product and the amount of sales tax paid contain much of the same information

- Irrelevant features: Contain almost no useful information for the data mining task at hand

- Example: Students' ID numbers are irrelevant to the task of predicting students' grade point averages

# Data Preprocessing: Feature Subset Selection

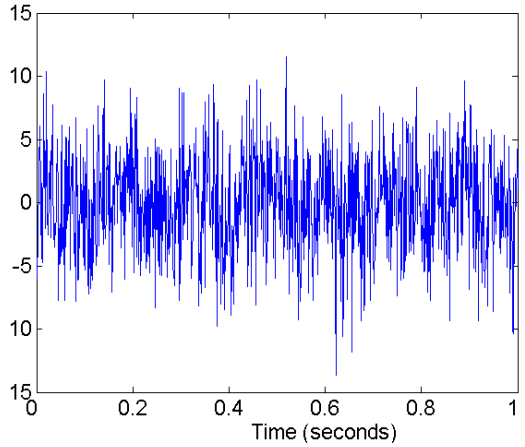
- Selecting the best subset of features frequently requires a systematic approach



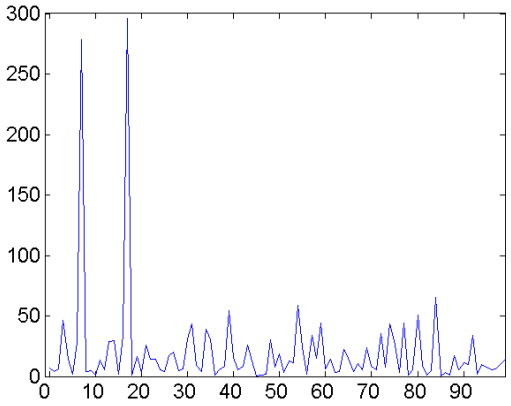
# Data Preprocessing: Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction - Example: extracting edges from images
  - Feature construction - Example: dividing mass by volume to get density
  - Mapping data to new space - Example: Fourier and wavelet analysis




**Time-domain Signal with 2 Sine Waves added to Noise**




**Frequency-domain Representation**

# Data Preprocessing: Discretization

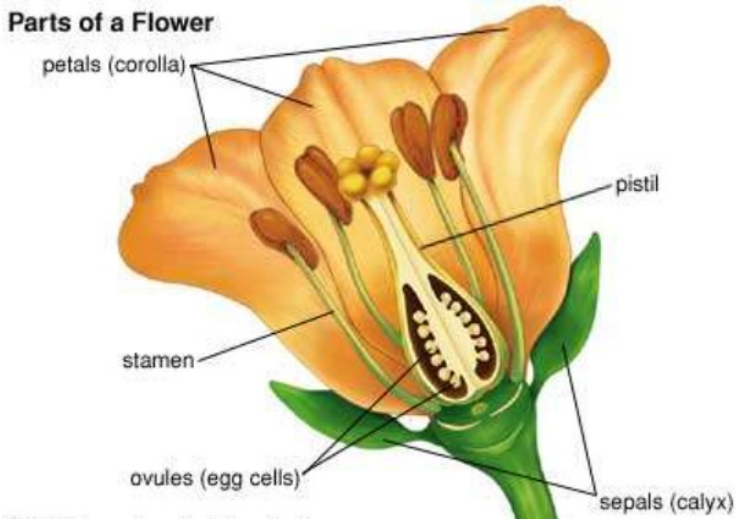
- Some data mining algorithms, especially certain classification algorithms, require that the data be in the form of categorical attributes
- Algorithms that find association patterns require that the data be in the form of binary attributes
- Discretization is the process of converting a continuous attribute into an ordinal attribute
- The best discretization approach is the one that produces the best result for the data mining algorithm that will be used to analyze the data

# Data Preprocessing: Discretization

- Example: Iris Plant data set.
- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

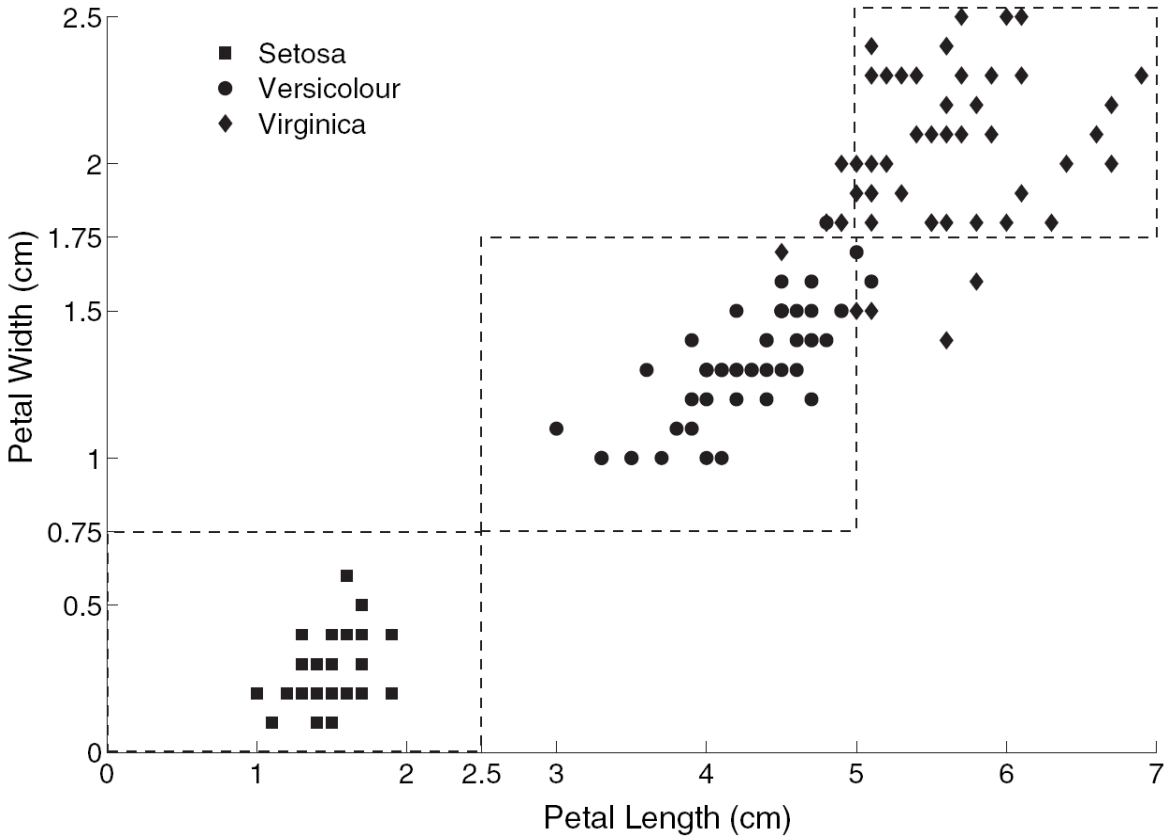
- Three flower types (classes):
  - Setosa
  - Versicolour
  - Virginica

- Four (non-class) attributes
  - Sepal width and length
  - Petal width and length




# Data Preprocessing: Discretization

- Example: Iris Plant data set.



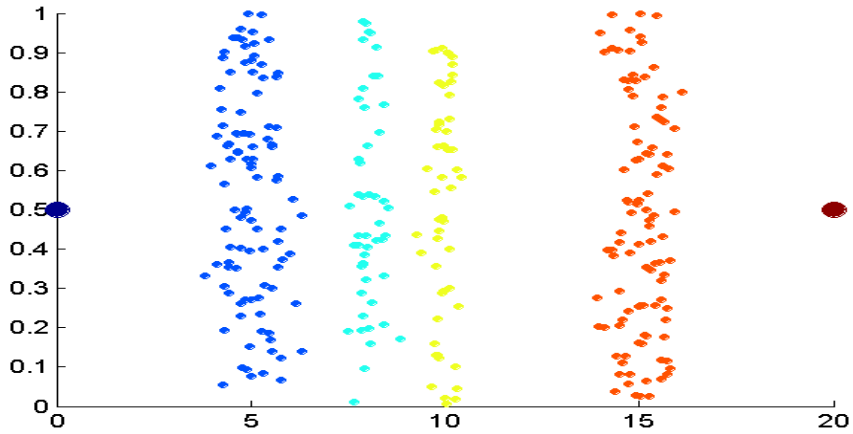

- Petal width low or petal length low implies Setosa
- Petal width medium or petal length medium implies Versicolour
- Petal width high or petal length high implies Virginica

# Data Preprocessing: Discretization

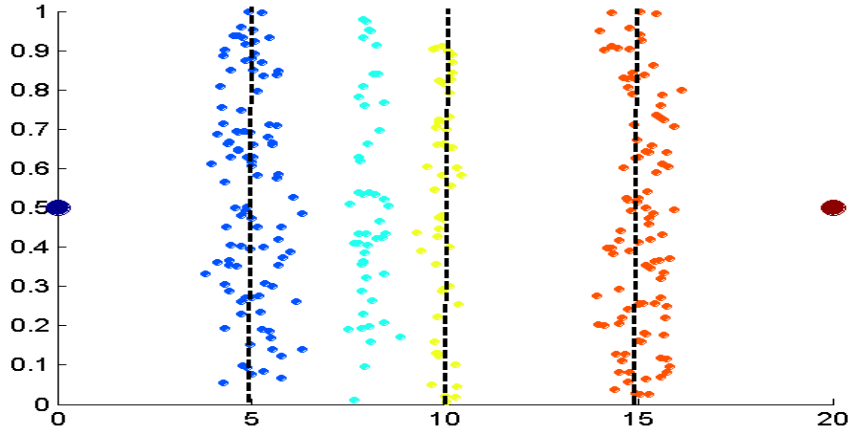
- Transformation of a continuous attribute to a categorical attribute involves two subtasks:
  - Setting the number of categories: Dividing the values into $n$ intervals by specifying $n$ – 1 split points
  - Determining how to map the values of the continuous attribute to these categories: all the values in one interval are mapped to the same categorical value
- Two Approaches:
  - Unsupervised
  - Supervised

# Data Preprocessing: Discretization

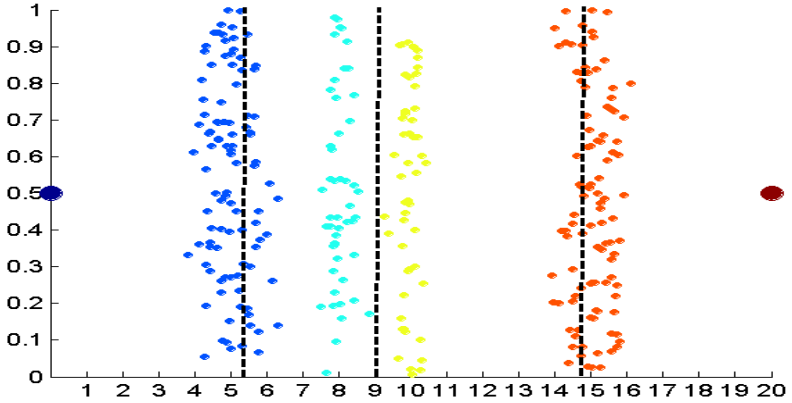
- Example:



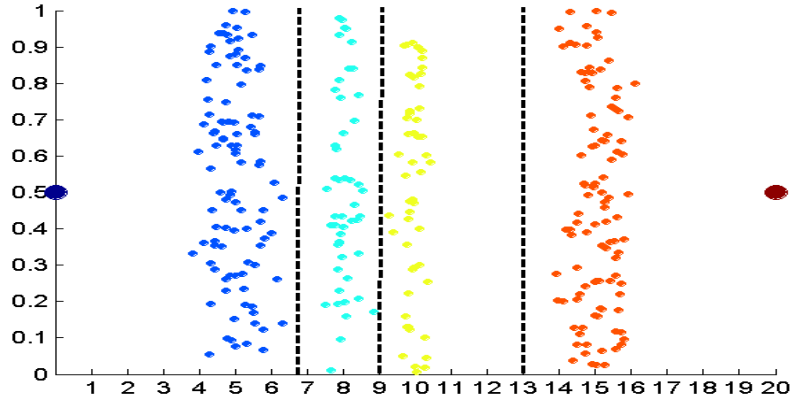

**Data consists of four groups of points and two outliers**




**Equal interval width approach used to obtain 4 values**




**Equal frequency approach used to obtain 4 values**




**K-means approach to obtain 4 values**

# Data Preprocessing: Variable Transformation

- Refers to a transformation that is applied to all the values of a variable
- A common type of variable transformation is the standardization or normalization of a variable
- A traditional example is that of "standardizing a variable" in statistics
- Any Gaussian random variable can be standardized as follows

**Standardizing a Normal Random Variable**

If $X$ is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma} \qquad (4\text{-}10)$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, $Z$ is a standard normal random variable.

# Data Preprocessing: Variable Transformation

- Proof:
  - $E(Z)$ = 0 → obvious
  - $Var(Z)$ = 1:
    - Consider a random variable $Y = aX + b$. First, we prove that $Var(Y) = a^2 Var(X)$

$$Var(aX+b) = E\left[\left\{(aX+b)-(a\mu+b)\right\}^2\right]$$

$$= E\left[a^2(X-\mu)^2\right]$$

$$= a^2 E\left[X^2 - 2X\mu + \mu^2\right]$$

$$= a^2\left(E(X^2) - \mu^2\right)$$

$$= a^2 Var(X)$$

- If $Z = \frac{X-\mu}{\sigma}$ is compared to $Z = aX + b$, then $a = \frac{1}{\sigma}$, $b = \frac{-\mu}{\sigma}$
- Therefore, $Var(Z) = a^2 Var(X) = 1$

# Data Preprocessing: Variable Transformation

- Example:

  Consider comparing people based on two variables: age and income

- For any two people, the difference in income will likely be much higher in absolute terms (hundreds or thousands of dollars) than the difference in age (less than 150)

- If the similarity or dissimilarity of two people is calculated using Euclidean distance, the income values will dominate the calculation

# Measures of Similarity and Dissimilarity

- Similarity and dissimilarity are important because they are used by a number of data mining techniques

- In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed

- The similarity between two objects is a numerical measure of the degree to which the two objects are alike. Usually non-negative and are often between 0 (no similarity) and 1 (complete similarity)

- The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different

# Measures of Similarity and Dissimilarity

- For two objects with one nominal attribute, they will either have the same value or not (similarity is 0 or 1)
- For two objects with a single ordinal attribute, it is more complex
- Example: Consider an attribute that measures the quality of a product, on the scale {poor, fair, OK, good, wonderful}, it would seem reasonable that a product rated wonderful would be closer to a product rated good than it would be to a product rated OK
- The values of the ordinal attribute are often mapped to successive integers, beginning at 0 or 1, {poor:0, fair:1, OK:2, good:3, wonderful:4}
- For interval or ratio attributes, the natural measure of dissimilarity between two objects is the absolute difference of their values

# Measures of Similarity and Dissimilarity

- For all 4 types of attributes:

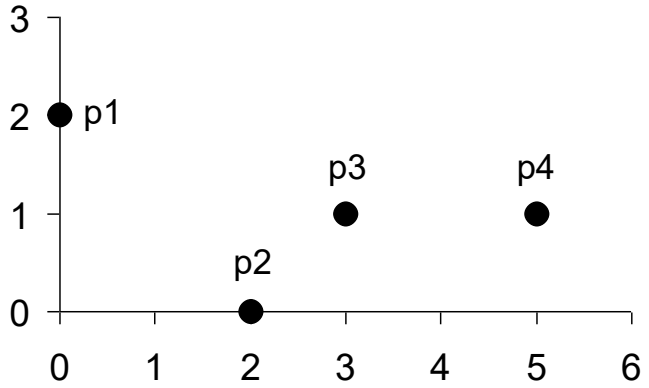
| Attribute Type | Dissimilarity | Similarity |
|---|---|---|
| Nominal | $d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$ | $s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$ |
| Ordinal | $d = \lvert x - y \rvert / (n-1)$ (values mapped to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

# Dissimilarities between Data Objects

- Euclidean Distance: For two $n$-dimensional objects, $\mathbf{x}$ and $\mathbf{y}$

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

- 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 |

# Dissimilarities between Data Objects

- Minkowski Distance: a generalization of Euclidean Distance

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

- Three most common examples of Minkowski distance:
  - $r$ = 1: City block (Manhattan, taxicab, L1 norm) distance

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

  - $r$ = 2: Euclidean distance (L2 norm)

  - $r$ = ∞: Supremum ($L_{max}$, or $L_\infty$) distance or Chebyshev Distance. This is the maximum difference between any attribute of the objects

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

# Dissimilarities between Data Objects

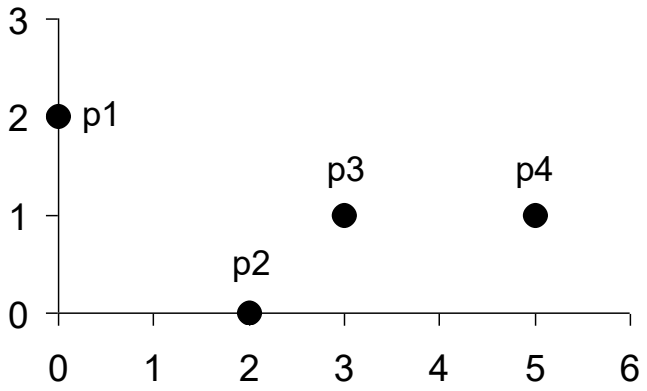
- Chebyshev Distance Proof: Suppose that $\max\{a, b\} = a$

$$\begin{aligned}
\lim_{p \to \infty} (a^p + b^p)^{1/p} &\geq \lim_{p \to \infty} (a^p)^{1/p} \\
&= \lim_{p \to \infty} a \\
&= a \\
&= \max\{a, b\}
\end{aligned}$$

$$\begin{aligned}
\lim_{p \to \infty} (a^p + b^p)^{1/p} &\leq \lim_{p \to \infty} (a^p + a^p)^{1/p} \\
&= \lim_{p \to \infty} (2a^p)^{1/p} \\
&= \lim_{p \to \infty} a \cdot 2^{1/p} \\
&= a \lim_{p \to \infty} 2^{1/p} \\
&= a \\
&= \max\{a, b\}
\end{aligned}$$

# Dissimilarities between Data Objects

- Example:




| 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 |

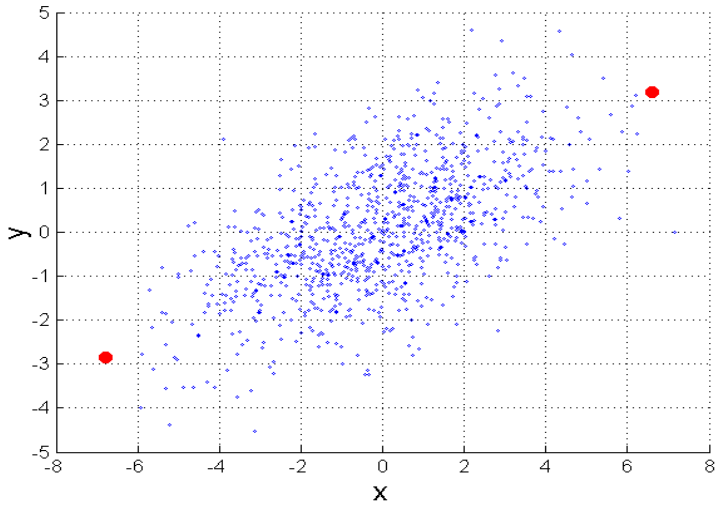
# Dissimilarities between Data Objects

- Mahalanobis Distance: useful when attributes are correlated, have different ranges of values (different variances), and the distribution of the data is approximately Gaussian

$$d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})\Sigma^{-1}(\mathbf{x}-\mathbf{y})^T$$

where $\sum$ is the covariance matrix

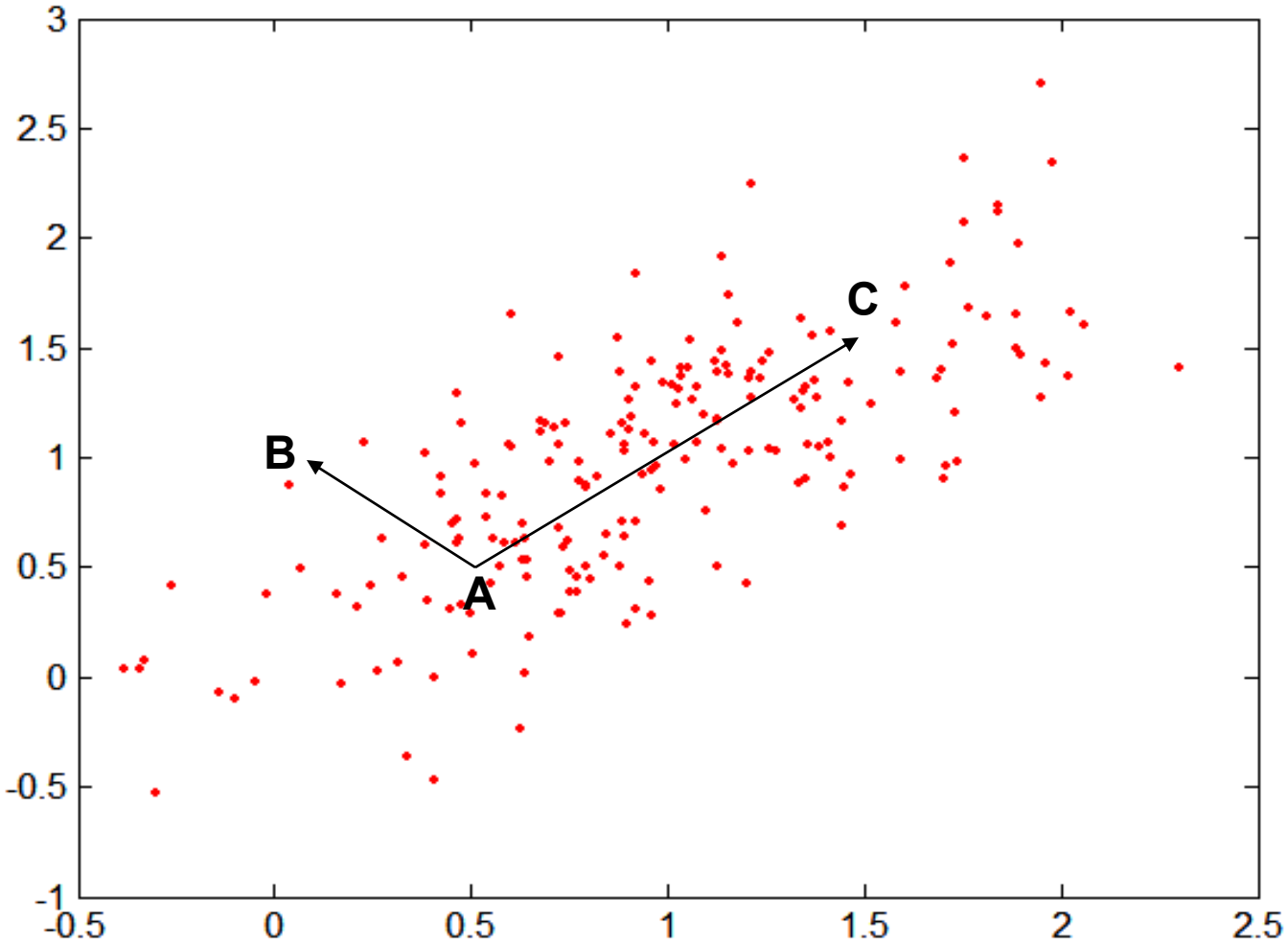
- We do not want to say that two things are very different if they differ slightly in many different aspects, all of which are similar to one another




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

# Dissimilarities between Data Objects

- Example:




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

# Dissimilarities between Data Objects

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

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$ and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)

2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

- A distance that satisfies these properties is a metric

- Non-metric Dissimilarity Example: Given two sets A and B, A - B is the set of elements of A that are not in B. If A={1,2,3,4} and B={2,3,4}, then A – B= {1} and B - A = 0, the empty set. We can define the distance *d* between two sets A and B as *d*(*A*, *B*) = size(*A* - *B*).

# Similarity Measures for Binary Data

- A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar

- Let $\mathbf{x}$ and $\mathbf{y}$ be two objects that consist of $n$ binary attributes. The comparison of two such objects leads to the following four quantities (frequencies):

  - $f_{00}$: the number of attributes where $\mathbf{x}$ is 0 and $\mathbf{y}$ is 0
  - $f_{01}$: the number of attributes where $\mathbf{x}$ is 0 and $\mathbf{y}$ is 1
  - $f_{10}$: the number of attributes where $\mathbf{x}$ is 1 and $\mathbf{y}$ is 0
  - $f_{11}$: the number of attributes where $\mathbf{x}$ is 1 and $\mathbf{y}$ is 1

# Similarity Measures for Binary Data

- Simple Matching Coefficient (SMC):

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

- Jaccard Coefficients:

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- Example:

Since the number of products not purchased by any customer outnumbers the number of products that were purchased, a similarity measure such as SMC would say that all transactions are very similar

# Similarity Measures for Binary Data

- 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)

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

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

# Similarity Measures for Non-binary Data

- Cosine Similarity: is one of the most common measures of document similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

- Example: Let two documents $\mathbf{d}_1$ and $\mathbf{d}_2$ be defined by counting the number of occurrences of some keywords

$$\mathbf{d}_1 = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$
$$\mathbf{d}_2 = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$\mathbf{d_1}.\ \mathbf{d2} = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$
$$\| \mathbf{d_1} \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$
$$\| \mathbf{d_2} \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d_1}, \mathbf{d_2}) = 0.315$$

# Similarity Measures for Non-binary Data

- Correlation: a measure of the linear relationship between the attributes of the objects

$$corr(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y}$$

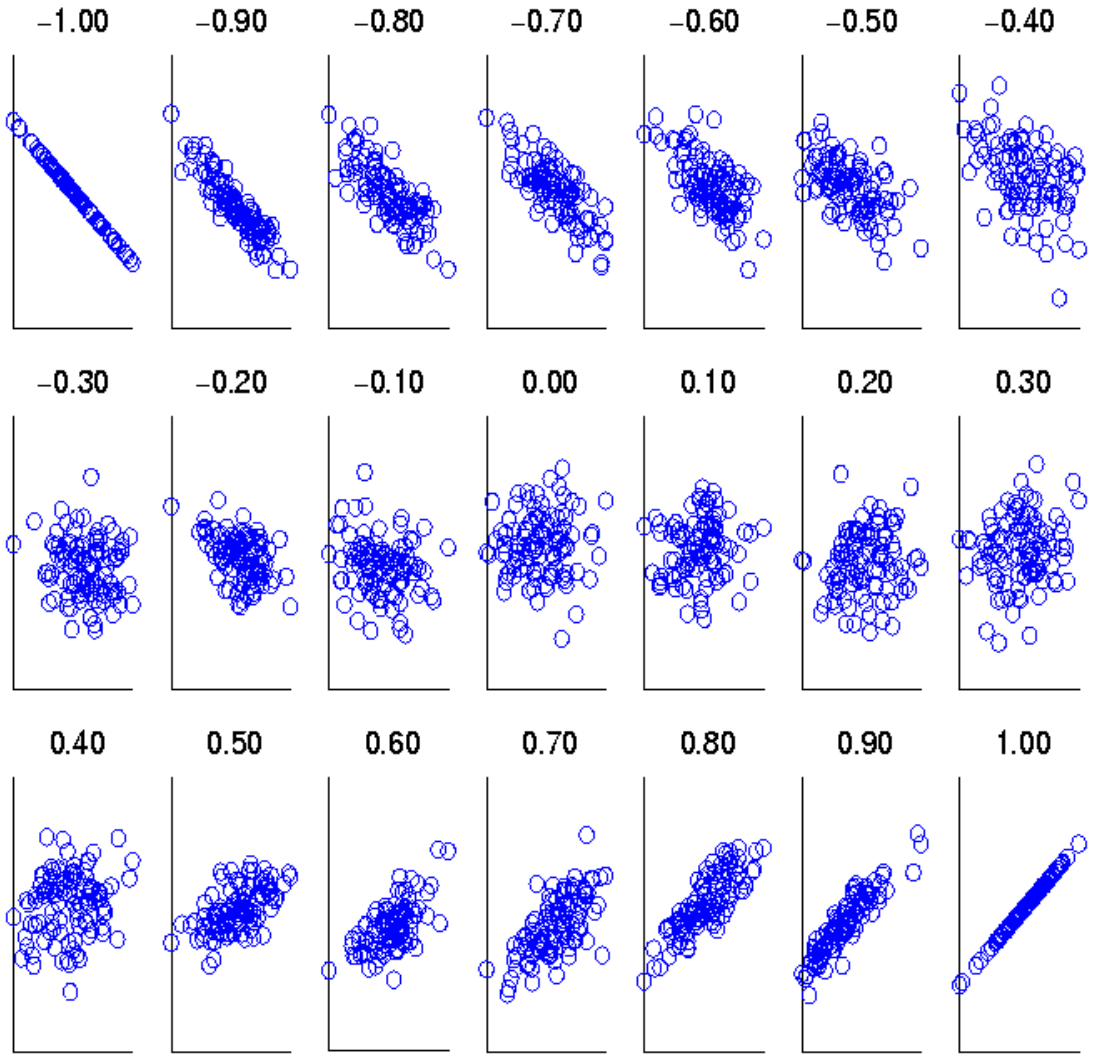
where $s_x$ and $s_y$ are the standard deviation of $\mathbf{x}$ and $\mathbf{y}$, respectively

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2} \qquad \bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2} \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

# Similarity Measures for Non-binary Data

- Example:



# Similarity Measures for Non-binary Data

- Example: Non-linear Relationship

  $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$

  $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

  $y_i = x_i^2$

  – mean(**x**) = 0, mean(**y**) = 4
  – std(**x**) = 2.16, std(**y**) = 3.74

  – corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / ( 6 * 2.16 * 3.74 )
    = 0