CENG3521 - Data Mining

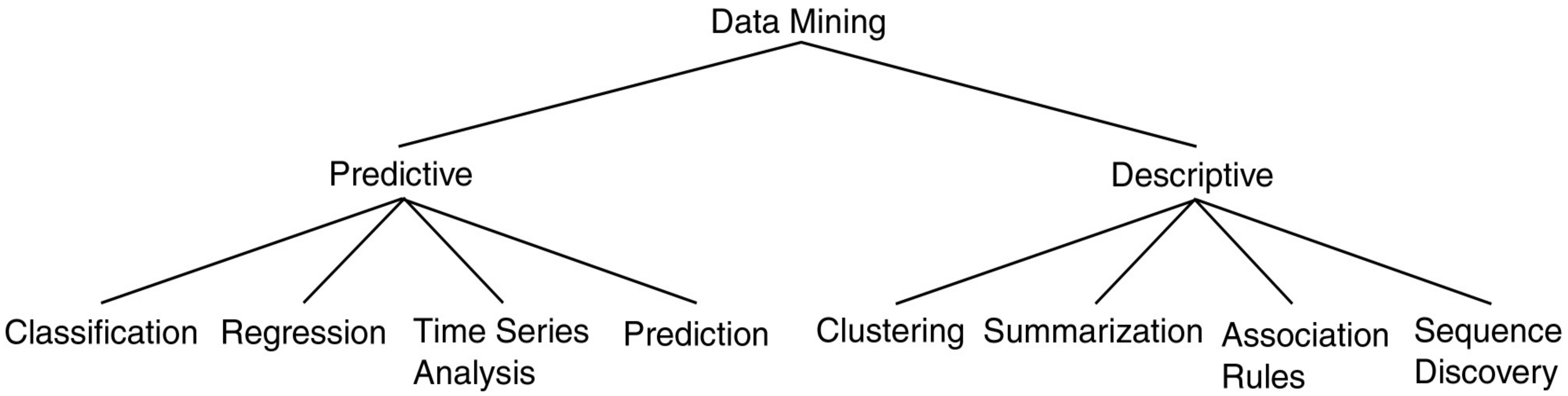
LECTURE 1

What is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.

Data Mining Definition

- Finding hidden information in a database
- Fit data to a model
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning



Data Mining Models and Tasks

Data Objects & Attributes

- A data set is a file, in which **the objects are records** (or rows) in the file and each field (or column) corresponds to an **attribute**.

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

Data Objects & Attributes

- An **attribute** is a data field, representing a characteristic or feature of a data object.

- Other names for attribute: *dimension*, *feature*, and *variable*

Data Objects & Attributes

- The **type** of an attribute is determined by the set of possible values: *nominal, binary, ordinal, or numeric*

1. Nominal(Categorical)Attributes:

i.e. *hair color* and *marital status*

- Nominal attribute values do not have any meaningful order about them and are **not quantitative**, so no mean (average) value or median (middle) value for such an attribute. Instead, **mode** is calculated, which is the attribute's **most commonly occurring value**

Data Objects & Attributes

2. Binary Attributes:

i.e. smoking

- A **binary attribute** is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as **Boolean** if the two states correspond to *true* and *false*.

  - Symmetric – gender (male, female)
  - Asymmetric – medical test result (HIV+, HIV-)

Data Objects & Attributes

3. Ordinal Attributes:

i.e. *drink size (small, medium*, and *large)*

grade (e.g., *A*+, *A*, *A*−, *B*+, and so on)

- An ordinal attribute is an attribute with possible values that have a **meaningful order** or **ranking** among them

- **Note** that nominal, binary, and ordinal attributes are ***qualitative***

Data Objects & Attributes

4. Numeric Attributes:

A numeric attribute is **quantitative**; that is, it is a **measurable quantity**, represented in integer or real values.

a. Interval-scaled attributes
b. Ratio-scaled attributes

Data Objects & Attributes

4. Numeric Attributes:

a. Interval-scaled Attributes:

- Interval-scaled attributes are measured on a scale of equal-size units
- The values of interval-scaled attributes have order and can be positive, 0, or negative
- In addition to providing a ranking of values, such attributes allow us to compare and quantify the *difference* between values.
- Because interval-scaled attributes are numeric, we can compute their **mean** value, in addition to the **median** and **mode.**

Data Objects & Attributes

4. Numeric Attributes:

a. Interval-scaled Attributes:

i.e. Temperatures in Celsius and Fahrenheit (**do not have a true zero-point**, that is, 0◦C or 0◦F **don't indicate "no temperature."**)

Without a true zero, we **cannot** say, for instance, that 10◦C is twice as warm as 5◦C.

We can quantify the difference between values. For example, a temperature of 20◦C is five degrees higher than a temperature of 15◦C.

Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.

Data Objects & Attributes

4. Numeric Attributes:

b. Ratio-scaled Attributes:

- A ratio-scaled attribute is a numeric attribute with a true zero-point.
- If a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- The values are ordered.
- We can compute the difference between values, as well as the **mean, median**, and **mode**

Data Objects & Attributes

4. **Numeric Attributes:**

b. **Ratio-scaled Attributes:**

i.e. Weight, length, counts, monetary quantities

- We can talk about ratios (order of magnitudes) - X is twice heavier than Y

Data Objects & Attributes

Discrete vs Continuous Attributes

1. **Continuous Attributes**

- A continuous attribute is one whose values are real numbers.
- Continuous attributes are typically represented as floating-point variables
- Real values can only be measured and represented with limited precision

  i.e. temperature, height, weight or time

Data Objects & Attributes

Discrete vs Continuous Attributes

2. Discrete Attributes

- A discrete attribute has a finite number of values, and so are discrete.
- Such attributes can be categorical, such as zip codes or ID numbers, or numeric, such as counts

  i.e. *hair color, smoker*

Data Objects & Attributes

Discrete vs Continuous Attributes

Discrete

- Your age in years
- Your eye color
- Cholesterol level in blood sample
- Your marital (relationship) status
- Your grade for this course

Continuous

- Your current weight
- Temperature of your room
- Time took to finish a homework
- Current time (at this moment)
- Your mom's age

Similarity & Dissimilarity

- **Similarity** between two objects is a numerical measure of the degree to which the two objects are **alike**.
- Similarities are **higher** for pairs of objects that are more alike.
- Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity)

Similarity & Dissimilarity

- The **dissimilarity** between two objects is a numerical measure of the degree to which the two objects are **different**.
- Dissimilarities are lower for more similar pairs of objects.
- The term distance is used as a synonym for dissimilarity.
- Dissimilarities sometimes fall in the interval [0,1]

Similarity & Dissimilarity

Example:

Consider objects described by **one nominal attribute**.

Since nominal attributes only convey information about the distinctness of objects, all we can say is that two objects either have the same value or they do not.

- In this case, similarity is defined as 1 if attribute values match, and as 0 otherwise.
- Dissimilarity would be defined in the opposite way: 0 if the attribute values match, and 1 if they do not

Similarity & Dissimilarity

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	$d = \|x - y\|/(n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \|x - y\|$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Similarity and dissimilarity for simple attributes

Similarity & Dissimilarity

Example:

Object with a single ordinal attribute.

- In this case, information about order should be taken into account.
- Consider an attribute that measures the quality of a candy bar, on the scale *{poor, fair, OK, good, wonderful}*.

P1=wonderful, P2=good, P3=OK, P4= fair, P5= poor

To make things quantitative, the values of the ordinal attribute are often mapped to successive integers, beginning at 0 or 1,

e.g. *{poor=0, fair=1, OK=3, good=3, wonderful=4}*

Similarity & Dissimilarity

Example:

- Dissimilarity:

d(P1,P2)=3-2=1

Or if we want the dissimilarity to fall between [0,1];

$D(P1,P2)=\frac{|x-y|}{n-1}=\frac{3-2}{4}=0.25$

- Similarity:

s=1-d

Dissimilarity between Data Objects - Distances

- *Minkowski distance*: A popular distance measure

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Dissimilarity between Data Objects - Distances

- h = 1: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- h = 2: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

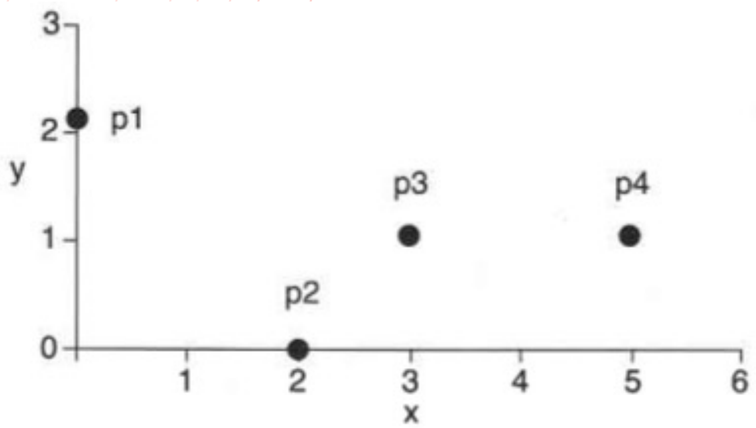
- $h \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Dissimilarity between Data Objects - Distances

HOMEWORK:

Calculate and create the Euclidean, L_1 and L_∞ distance matrixes for points p1, p2, p3 and p4 given below. (show your calculations)



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

Dissimilarity between Data Objects

Correlation & Redundancy (

- The **correlation** between two data objects that have binary or continuous variables is a **measure of the linear relationship** between the attributes of the objects.

- Correlation is **always in the range -1 to 1**. A correlation of **1** means x and y have a **perfect positive linear relationship**; **-1** means a **perfect negative relationship**.

i.e.: if *A* and *B* are *positively correlated*, meaning that the values of *A* increase as the values of *B* increase.

- The higher the value, the stronger the correlation (i.e., the more each attribute implies the other). Hence, a higher value may indicate that *A* (or *B*) may be removed as a **redundancy**.

Pearson's correlation coefficient between two data objects, x and y, is defined by the following equation:

$$\operatorname{corr}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{covariance}(\mathbf{x}, \mathbf{y})}{\operatorname{standard\_deviation}(\mathbf{x}) * \operatorname{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x\, s_y}, \quad (2.10)$$

where we are using the following standard statistical notation and definitions:

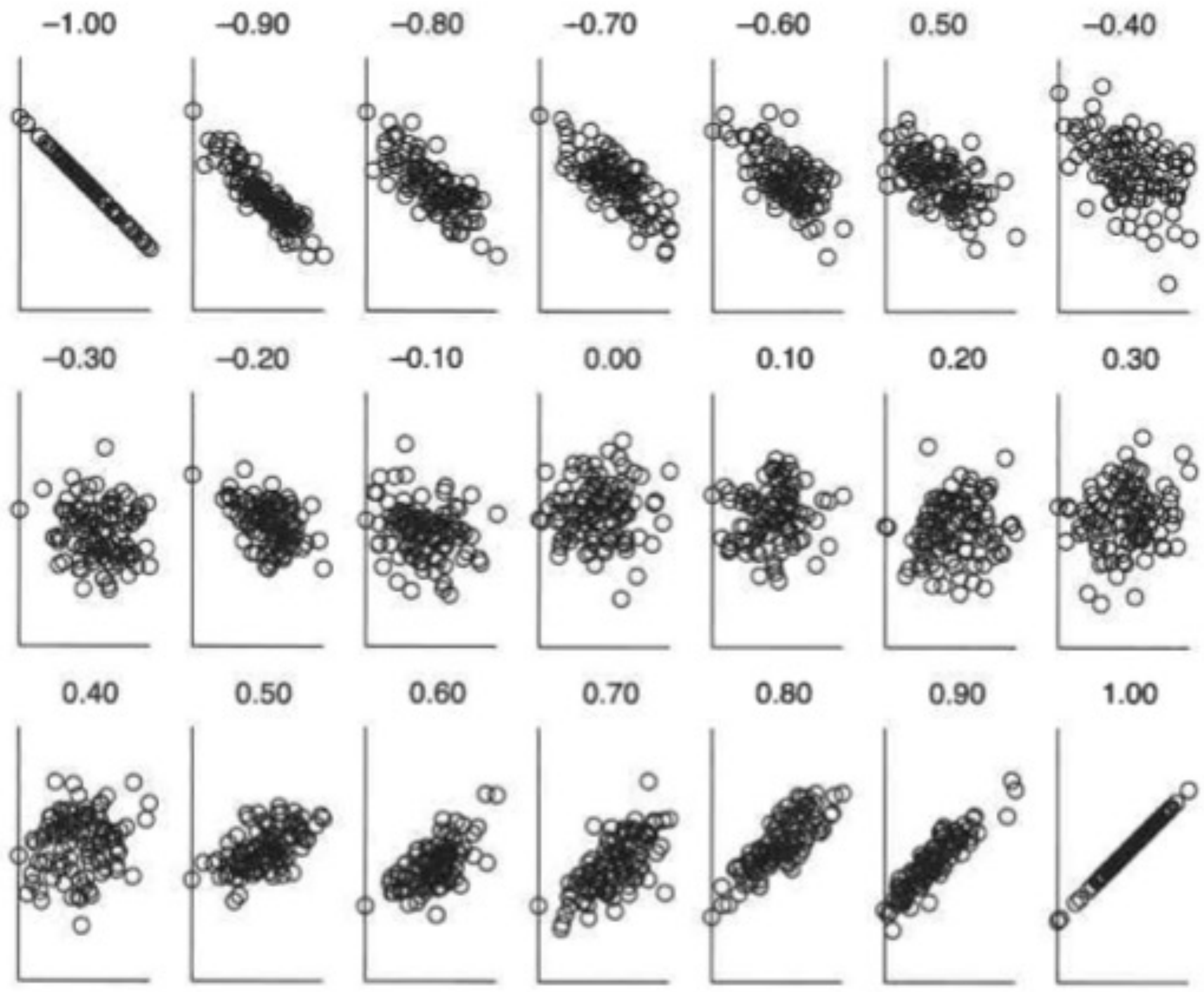
$$\operatorname{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.11)$$

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

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

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

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



Scatterplots illustrating correlations from -1 to 1