CHAPTER 3: DATA EXPLORATION

Based on CSC603.3:

Students should be able to preprocess the raw data and make it ready for the various data mining tasks

OUTLINE:

- **Data Exploration**
  - Types of Attributes
  - Statistical Description of Data
  - Data Visualization
  - Measuring similarity and dissimilarity

DATA EXPLORATION

- A preliminary exploration of the data to better understand its characteristics.
- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns (People can recognize patterns not captured by data analysis tools)

DATA OBJECTS AND TYPES OF ATTRIBUTES

May 2017

TYPES OF DATA SETS

1. **Record**
   - Relational records
   - Data matrix, e.g., numerical matrix, crosstabs
   - Document data: text documents: term-frequency vector
   - Transaction data
2. **Graph and network**
   - World Wide Web
   - Social or information networks
   - Molecular Structures
3. **Ordered**
   - Video data: sequence of images
   - Temporal data: time-series
   - Sequential Data: transaction sequences
   - Genetic sequence data
4. **Spatial, image and multimedia:**
   - Spatial data: maps
   - Image data:
   - Video data:

TYPES OF DATA SETS

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, cross tabulations.
  - Document data: text documents: term-frequency vector
  - Transaction data

Relational records

Login	First name	Last name
koala	John	Clemens
lion	Mary	Stevens

record

Login	phone
koala	039689852639

Document data

	team	ball	lost	timeout
Document1	3	5	2	2
Document2	0	0	3	0
Document3	0	1	0	0

record

Transactional data

TID	Items Books
1	Bred, Cake, Milk
2	Beer, Bred

record

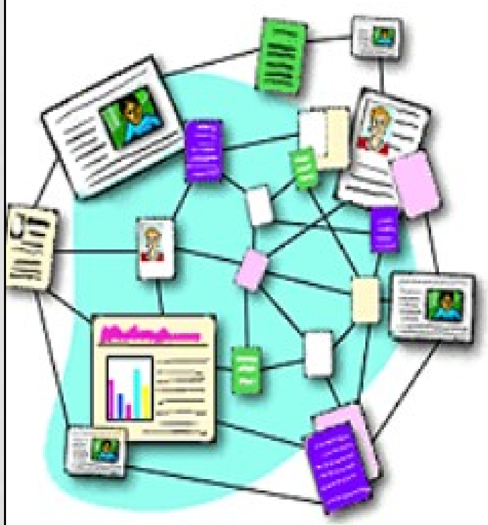
Cross tabulation

	Books	Multimedia devices
Big spenders	30%	70%
Budget spenders	60%	25%
Very Tight spenders	10%	5%

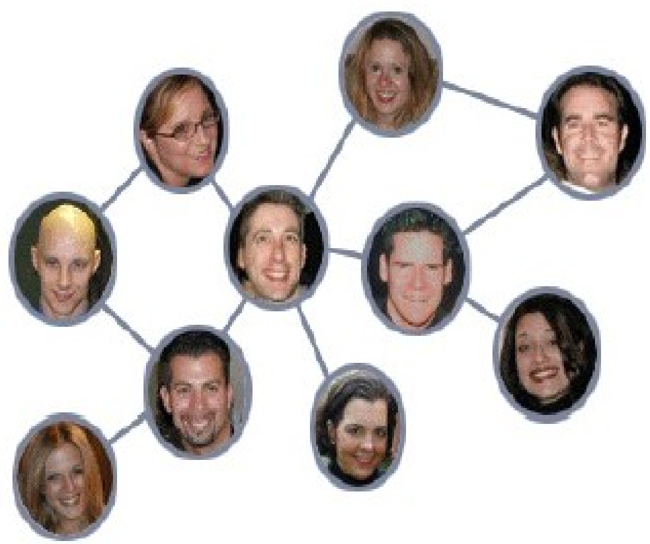
record

TYPES OF DATA SETS

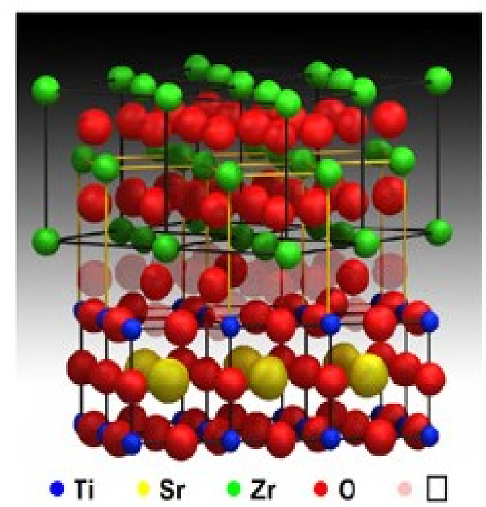
- **Graph and Network**
  - → World Wide Web
  - → Social or information networks
  - → Molecular structures networks



World Wide Web



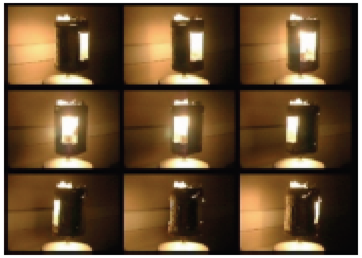
Social Networks



Molecular Structures Network

TYPES OF DATA SETS

- **Ordered**
  - → Videos
  - → Temporal data
  - → Sequential data
  - → Genetic sequence data

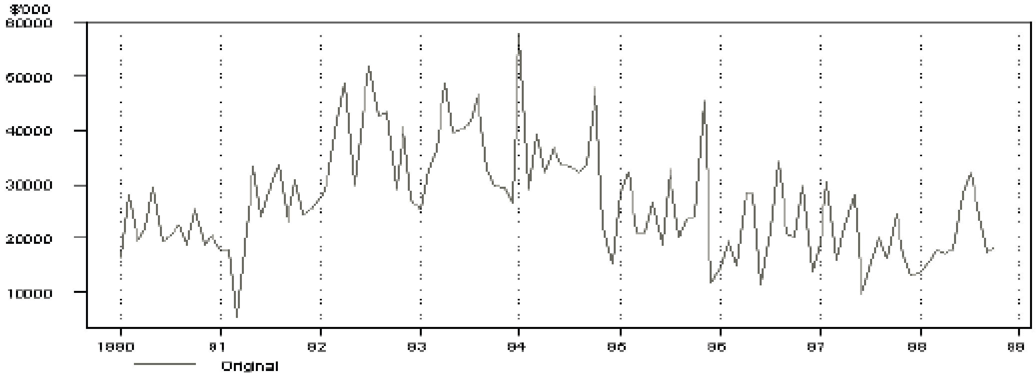


Video: sequence of mages

Transactional sequence

Computer-> Web cam ->USB key

Generic Sequence:
DNA-code



Temporal data: Time-series
monthly Value of Building Approvals

TYPES OF DATA SETS

- **Spatial, image and multimedia**
  - → Spatial data
  - → Image data
  - → Video data
  - → Audio Data

Spatial data: maps

Images

Videos

Audios

DATA

- Data sets are made up of data objects.
- A **data object** represents an entity.
- **Examples**
  - → Sales database: customers, store items, sales
  - → Medical database: patients, treatments
  - → University database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes.**
- Database rows -> data objects; columns ->attributes.

Patient_ID	Age	Height	Weight	Gender
1569	30	1,76m	70 kg	male
2596	26	1,65m	58kg	female

Data Object

Attributes

DATA ATTRIBUTES (MAY 17)

- **Attribute (**or **dimensions**, **features**, **variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- **Types:**
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

TYPES OF DATA ATTRIBUTE VALUES

1. **Nominal:** categories, states, or "names of things"
    - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
    - marital status, occupation, ID numbers, zip codes
2. **Binary**
    - Nominal attribute with only 2 states (0 and 1)
    - **Symmetric binary**: both outcomes equally important
        - e.g., gender
    - **Asymmetric binary:** outcomes not equally important.
        - e.g., medical test (positive vs. negative)
        - Convention: assign 1 to most important outcome (e.g., HIV positive)
3. **Ordinal**
    - Values have a meaningful order (ranking) but magnitude between successive values is not known.
    - *Size = {small, medium, large}*, grades, army rankings

TYPES OF DATA ATTRIBUTE VALUES

4.Numeric: Quantity (integer or real-valued)

a. **Interval**
   - Measured on a scale of **equal-sized units**
   - Values have order
   - E.g.. Weight, height, latitude, longitude, temperature

b. **Ratio**
   - Makes a positive measurement on a non-linear scale
   - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
     - e.g., *temperature in Kelvin, length, counts, monetary quantities*

PROPERTIES OF ATTRIBUTE VALUES

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \neq$
  - Order: $< >$
  - Addition: $+ -$
  - Multiplication: $* /$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

COMPARISON OF TYPES OF DATA ATTRIBUTES

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects (< >).	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

TYPES OF ATTRIBUTE

Discrete Data with example:

Discrete data have finite value. It can be in numerical form and can also be in categorical form.

Example:

Attribute	Value
Profession	Teacher, Bussiness Man, Peon etc
Postal Code	42200, 42300 etc

Continuous data with example:

Continuous data technically have an infinite number of steps.

Continuous data is in float type. There can be many numbers in between 1 and 2

Example:

Attribute	Value
Height	5.4…, 6.5….. etc
Weight	50.09…. etc

Statistical Description of Data

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data

DESCRIPTIVE DATA SUMMARIZATION

- **Motivation**
  - For data preprocessing, it is essential to have an overall picture of your data
  - Data summarization techniques can be used to
    - Define the typical properties of the data
    - Highlight which data should be treated as noise or outliers
- **Data dispersion characteristics**
  - Median, max, min, quantiles, outliers, variance, etc.
- **From the data mining point of view it is important to**
  - Examine how these measures are computed efficiently
  - Introduce the notions of **distributive** measure, **algebraic** measure and **holistic** measure

THE MEASURES OF CENTRAL TENDENCY

- 3 measures of central tendency are commonly used in statistical analysis - MEAN, MEDIAN, and MODE.
- Each measure is designed to represent a “typical” value in the distribution.
- The choice of which measure to use depends on the shape of the distribution (whether normal or skewed).

MEAN - AVERAGE

- Most common measure of central tendency.
- Is sensitive to the influence of a few extreme values (outliers), thus it is not always the most appropriate measure of central tendency.
- Best used for making predictions when a distribution is more or less normal (or symmetrical).
- Symbolized as:
  - $\bar{x}$ for the mean of a sample
  - μ for the mean of a population

FINDING THE MEAN

- Formula for Mean: $\overline{X} = \dfrac{(\Sigma\, x)}{N}$
- Given the data set: {3, 5, 10, 4, 3}

$$\overline{X} = \frac{(3 + 5 + 10 + 4 + 3)}{5} = \frac{25}{5}$$

$$\overline{X} = 5$$

FIND THE MEAN

Q: 85, 87, 89, 91, 98, 100

A: 91.67

Median: 90

Q: 5, 87, 89, 91, 98, 100

A: 78.3

Median: 90

MEDIAN

- Used to find middle value (center) of a distribution.
- Used when one must determine whether the data values fall into either the upper 50% or lower 50% of a distribution.
- Used when one needs to report the typical value of a data set, ignoring the outliers (few extreme values in a data set).
    - Example: median salary, median home prices in a market
- Is a better indicator of central tendency than mean when one has a skewed distribution.

TO COMPUTE THE MEDIAN

- first you order the values of X from low to high: ➔ 85, 90, 94, 94, 95, 97, 97, 97, 97, 98
- then count number of observations = 10.
- When the number of observations are even, average the two middle numbers to calculate the median.
- This example, 96 is the median
  (middle) score.

MEDIAN

- Find the Median

  4 5 6 6 7 8 9 10 12

- Find the Median

  5 6 6 7 8 9 10 12

- Find the Median

  5 6 6 7 8 9 10 100,000

MODE

- Used when the most typical (common) value is desired.
- Often used with categorical data.
- The mode is not always unique. A distribution can have no mode, one mode, or more than one mode. When there are two modes, we say the distribution is *bimodal*.

EXAMPLES:

a) {1,0,5,9,12,8} - No mode
b) {4,5,5,5,9,20,30} – mode = 5
c) {2,2,5,9,9,15} - bimodal, mode 2 and 9

MEASURES OF VARIABILITY

- Central Tendency doesn't tell us everything Dispersion/Deviation/Spread tells us a lot about how the data values are distributed.

- We are most interested in:
  - Standard Deviation (σ) and
  - Variance (σ^2)

WHY CAN'T THE MEAN TELL US EVERYTHING?

- Mean describes the average outcome.
- The question becomes how good a representation of the distribution is the mean? *How good is the mean as a description of central tendency -- or how accurate is the mean as a predictor?*
- ANSWER -- it depends on the shape of the distribution. Is the distribution normal or skewed?

DISPERSION

- Once you determine that the data of interest is normally distributed, ideally by producing a histogram of the values, the next question to ask is: ***How spread out are the values about the mean?***
- **Dispersion** is a key concept in statistical thinking.
- The basic question being asked is how much do the values deviate from the Mean? **The more "bunched up" around the mean the better your ability to make accurate predictions.**

MEANS

- Consider these means for hours worked day each day:

$X = \{7, 8, 6, 7, 7, 6, 8, 7\}$

$\bar{X} = (7+8+6+7+7+6+8+7)/8$

$\bar{X} = 7$

Notice that all the data values are bunched near the mean.

Thus, 7 would be a pretty good prediction of the average hrs. worked each day.

$X = \{12, 2, 0, 14, 10, 9, 5, 4\}$

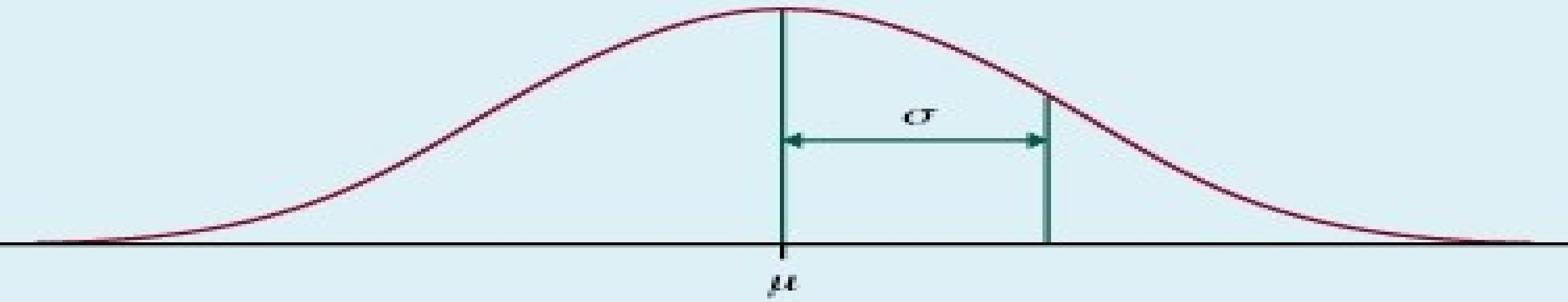
$\bar{X} = (12+2+0+14+10+9+5+4)/8$

$\bar{X} = 7$

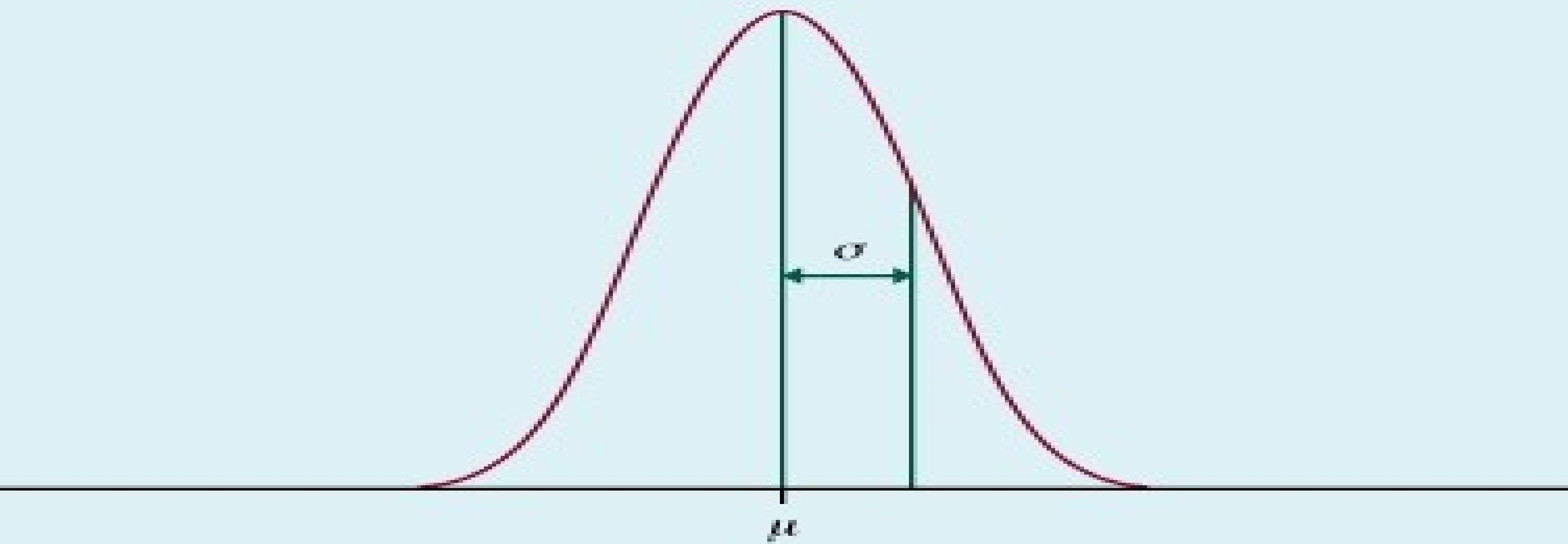
The mean is the same for this data set, but the data values are more spread out.

So, 7 is not a good prediction of hrs. worked on average each day.

Data is more spread out, meaning it has greater variability.



Below, the data is grouped closer to the center, less spread out, or smaller variability.



- How well does the mean represent the values in a distribution?
- The logic here is to determine **how much spread** is in the values. How much do the values "deviate" from the mean? Think of the mean as the **true value**, or as your **best guess**. If every X were very close to the Mean, the Mean would be a very good predictor.
- If the distribution is very sharply peaked then the mean is a good measure of central tendency and if you were to use the Mean to make predictions you would be correct or very close much of the time.

MEAN ABSOLUTE DEVIATION

The key concept for describing normal distributions and making predictions from them is called **deviation from the mean**.

We could just calculate the average distance between each observation and the mean.

- We must take the absolute value of the distance, otherwise they would just cancel out to zero!

Formula:

$$\sum \frac{|\overline{X} - X_i|}{n}$$

MEAN ABSOLUTE DEVIATION: AN EXAMPLE

Data: $X = \{6, 10, 5, 4, 9, 8\}$ $\bar{X} = 42 / 6 = 7$

$\bar{X} - X_i$	Abs. Dev.
7 – 6	1
7 – 10	3
7 – 5	2
7 – 4	3
7 – 9	2
7 – 8	1

Total: 12

1. Compute $\bar{X}$ (Average)
2. Compute $\bar{X} - X$ and take the Absolute Value to get Absolute Deviations
3. Sum the Absolute Deviations
4. Divide the sum of the absolute deviations by N

$12 / 6 = 2$

VARIANCE AND STANDARD DEVIATION

- Instead of taking the absolute value, we square the deviations from the mean. This yields a positive value.
- This will result in measures we call the Variance and the Standard Deviation

s Standard Deviation σ Standard Deviation

s^2 Variance σ^2 Variance

CALCULATING THE VARIANCE AND/OR STANDARD DEVIATION

Formulae:

Variance:

$$s^2 = \frac{\sum(\overline{X} - X_i)^2}{N}$$

Standard Deviation:

$$s = \sqrt{\frac{\sum(\overline{X} - X_i)^2}{N}}$$

Examples Follow . . .

EXAMPLE:

Data: X = {6, 10, 5, 4, 9, 8}; N = 6

X	$X-\bar{X}$	$(X-\bar{X})^2$
6	-1	1
10	3	9
5	-2	4
4	-3	9
9	2	4
8	1	1
Total: 42		Total: 28

Mean:

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

Variance:

$$s^2 = \frac{\sum(\bar{X}-X)^2}{N} = \frac{28}{6} = 4.67$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

DATA VISUALIZATION

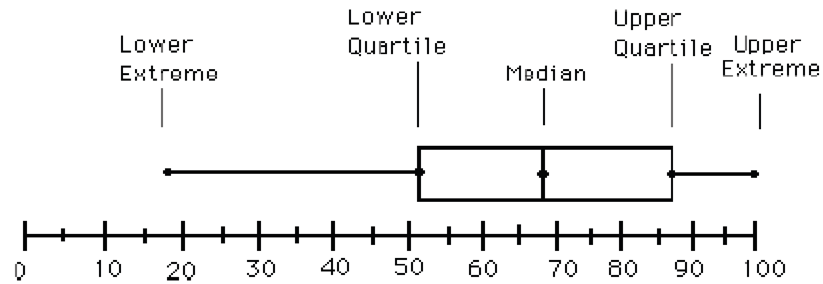
Dec 17, May 17

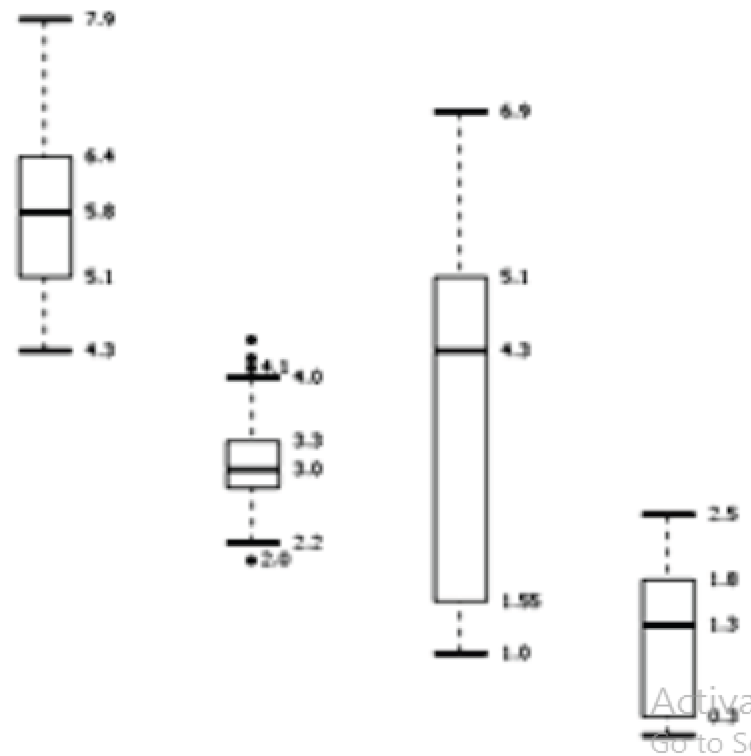
Data Visualization

- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane

1. BOX PLOT

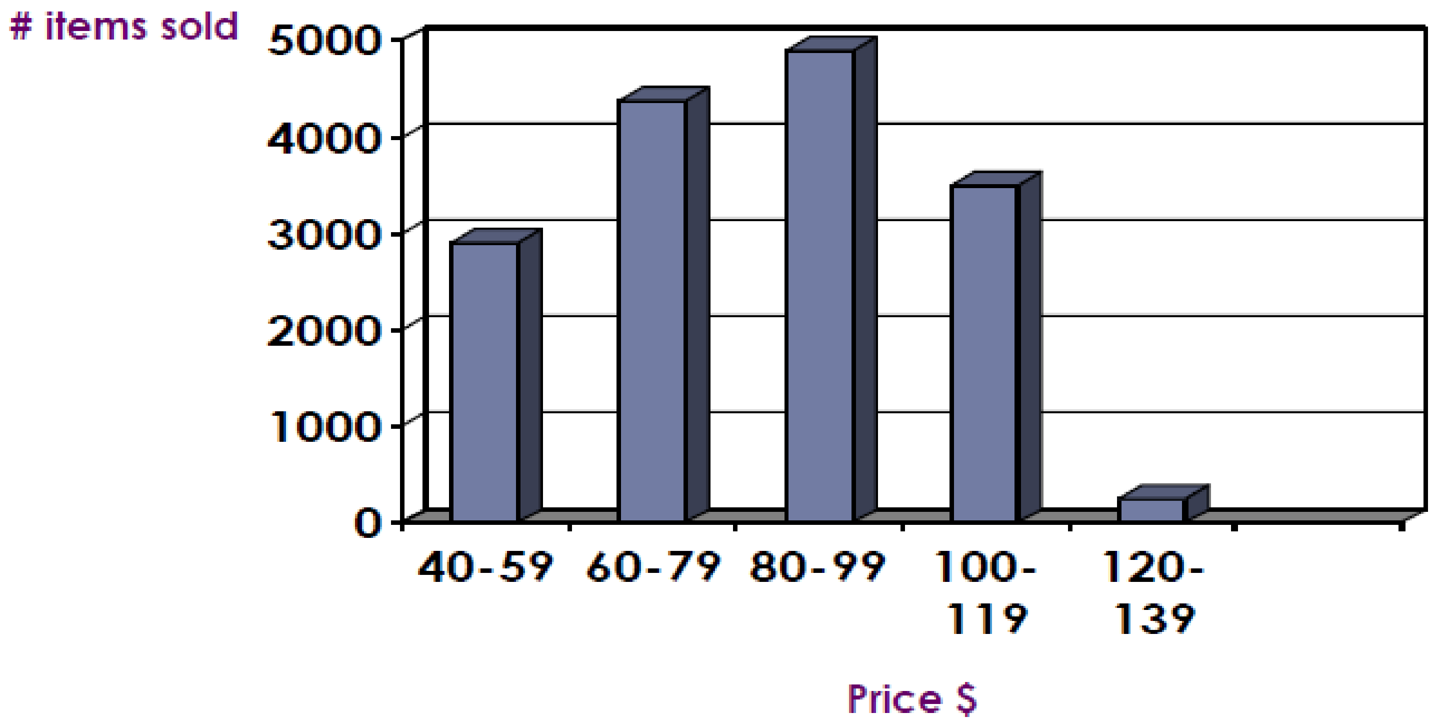
- **Five-number summary** of a distribution
  - → Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - → Data is represented with a box
  - → The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - → The median is marked by a line within the box
  - → Whiskers: two lines outside the box extended to Minimum and Maximum
  - → Outliers: points beyond a specified outlier threshold, plotted individually





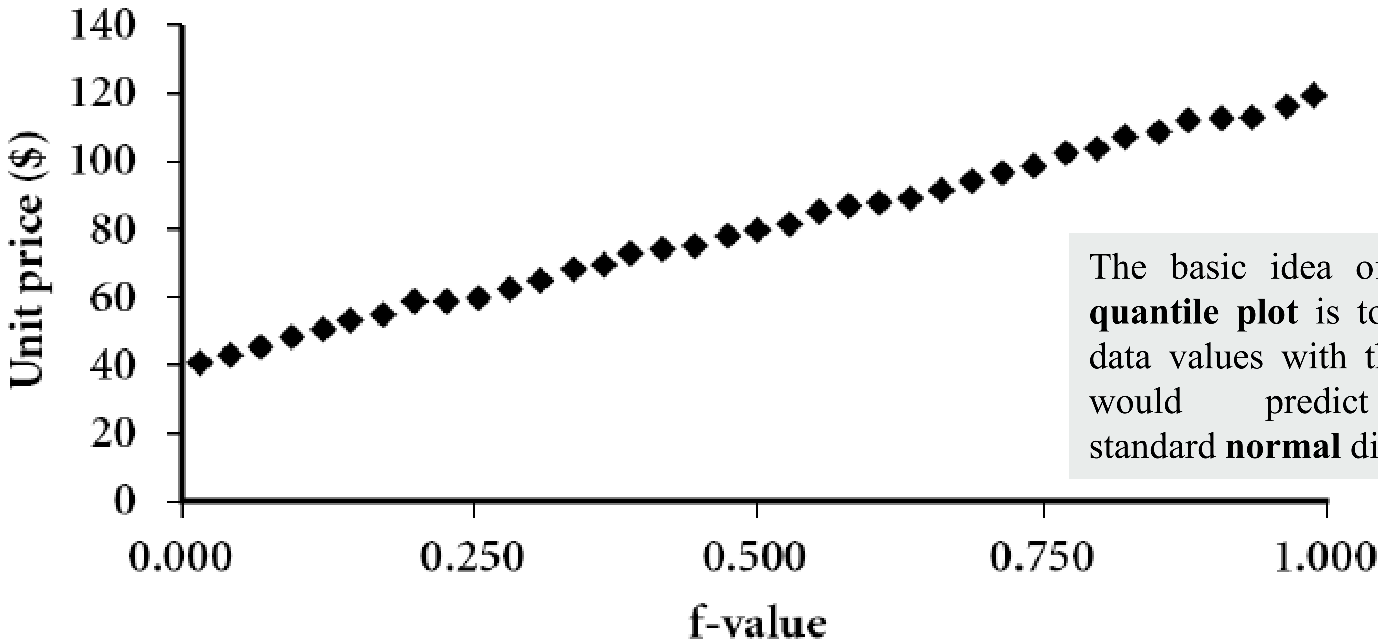
2. HISTOGRAM ANALYSIS

- **Histogram**: summarizes the distribution of a given attribute
- Partition the data distribution into disjoint subsets, or buckets
- If the attribute is **nominal** → **bar chart**
- If the attribute is **numeric** → **histogram**



3. QUANTILE PLOT

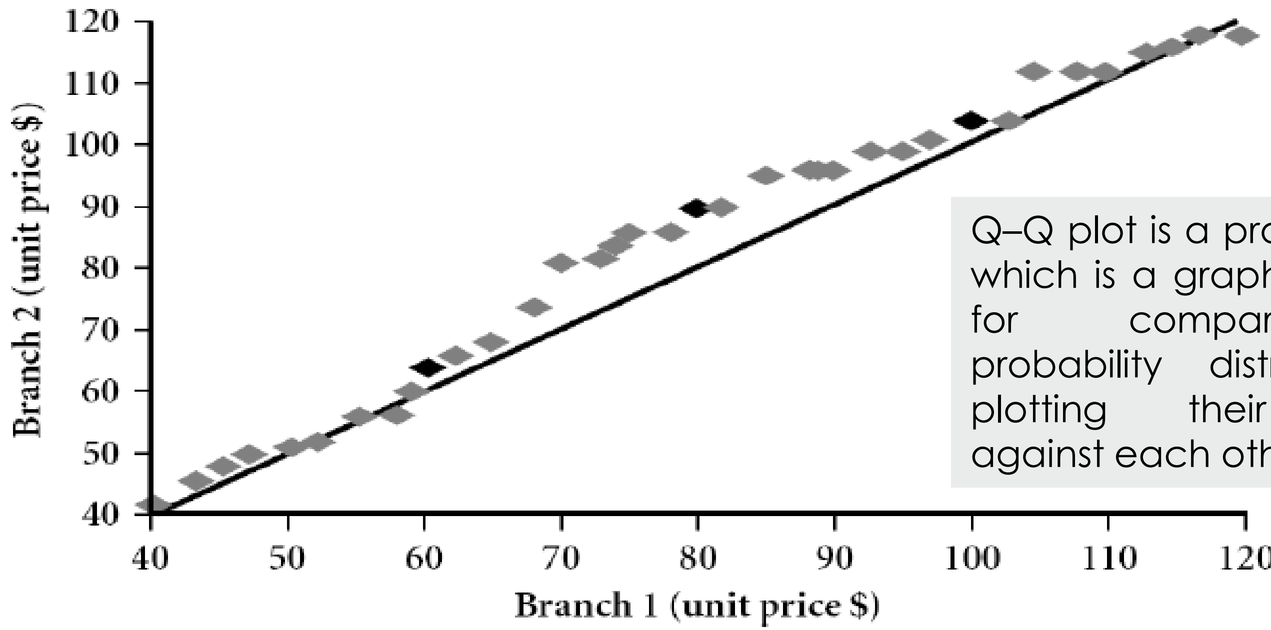
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - → For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



The basic idea of the **normal quantile plot** is to compare the data values with the values one would predict for a standard **normal** distribution.

4. QUANTILE-QUANTILE (Q-Q) PLOT

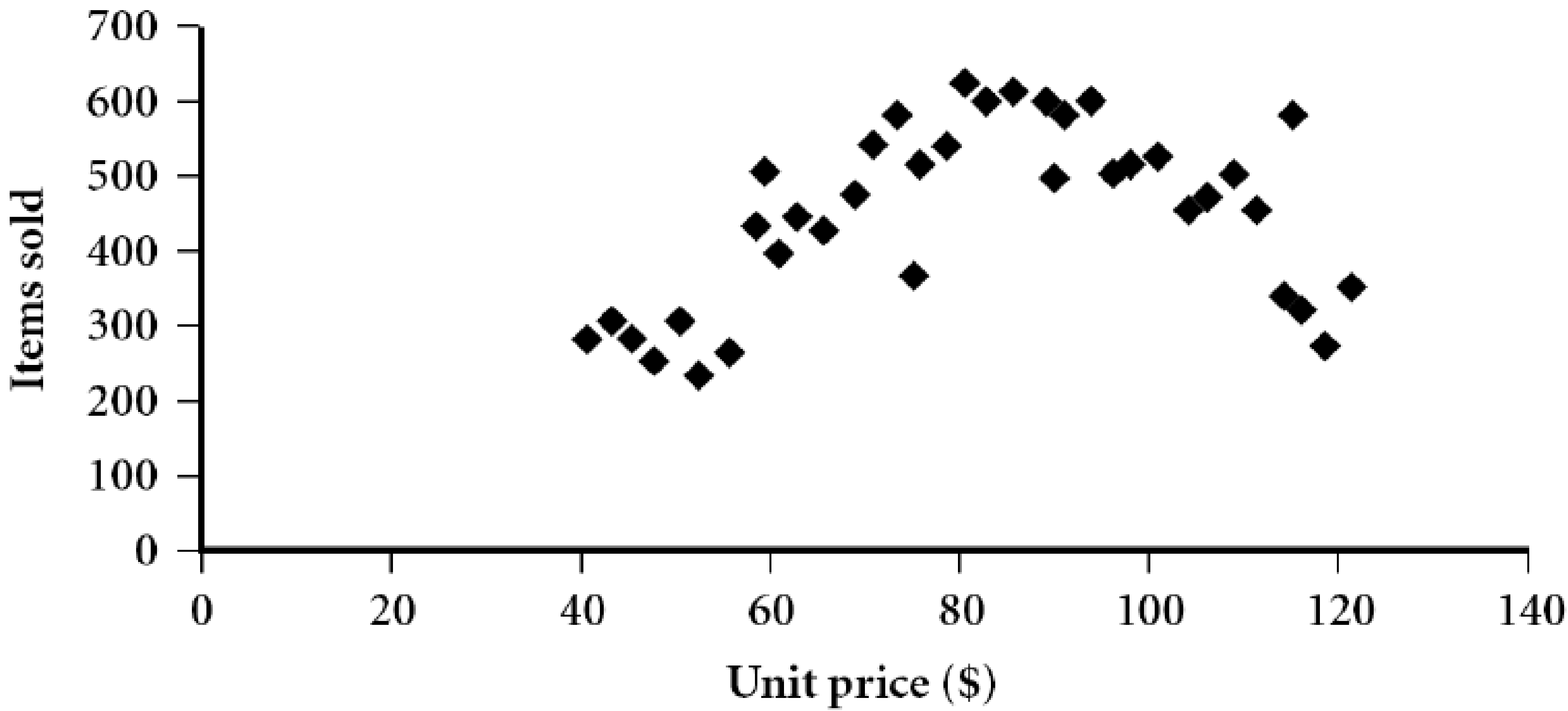
- Graph the quantiles of one univariate distribution against the corresponding quantiles of another
- **View:** Is there a shift in going from one distribution to another?
- **Example** shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

5. SCATTER PLOT

- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



SOLVE THE NUMERICAL

- Find mean, median, mode, mid range,Q1,Q3, Interquartile range(IQR), mid Quartile range, Semi quartile range for the data given below: 45,47,52,52,53,55,56,58,62,80.

Mean= (560)/10=56
Median=(53+55)/2=54
Mode= 52
Mid range= (High value + Low value)/2=(80+45)/2=62.5
Q1=median of lower half=52
Q3=median of upper half=58
IQR=Q3-Q1=58-52=6
Outlier=1.5*IQR=1.5*6=9
Semi inter Quartile range= (Q3-Q1)/2=3
Mid quartile range=(Q3+Q1)/2=55

Suppose that the data for analysis includes the attribute salary. We have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. [10]

(i) What are the *mean, median, mode* and *midrange* of the data?
(ii) Find the *first quartile* (Q1) and the *third quartile* (Q3) of the data.
(iii) Show a *boxplot* of the data.

- Mean= 58
- Median= (52+56)/2= 54
- Mode= 52,70
- Max value= 110
- Min value= 30
- Mid range= (110+30)/2=70
- Q1= median of lower part of data= (47+50)/2= 48.5 [30,36,47,50,52,52]
- Q3= median of upper part of data= (63+70)/2= 66.5 [56,60,63,70,70,110]

UNIVERSITY ASKED QUESTIONS

1. Write short note on Data Visualization. **(5 marks) Dec 2017**
2. Explain types of attributes and data visualization for data exploration **(10 marks) May 2017**

Dec 2019

Suppose that the data for analysis includes the attribute salary. We have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. [10]

(i) What are the *mean, median, mode* and *midrange* of the data?

(ii) Find the *first quartile* (Q1) and the *third quartile* (Q3) of the data.

(iii)Show a *boxplot* of the data.

MEASURING SIMILARITY AND DISSIMILARITY

May 2019

DATA SIMILARITY AND DISSIMILARITY

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

PROXIMITY MEASURES: SINGLE-ATTRIBUTE

- Consider an object, which is defined by a single attribute *A* (e.g., length) and the attribute *A* has *n*-distinct values $a_1, a_2, \ldots\ldots, a_n$.
- A data structure called "Dissimilarity matrix" is used to store a collection of proximities that are available for all pair of *n* attribute values.
  - In other words, the Dissimilarity matrix for an attribute *A* with *n* values is represented by an $n \times n$ matrix as shown below.

$$\begin{bmatrix} 0 & & & \\ p_{(2,1)} & 0 & & \\ p_{(3,1)} & p_{(3,2)} & 0 & \\ \vdots & \vdots & \vdots & \\ p_{(n,1)} & p_{(n,2)} & \ldots\ldots & 0 \end{bmatrix}_{n\times n}$$

  - Here, $p_{(i,j)}$ denotes the proximity measure between two objects with attribute values a_i and a_j.
- Note: The proximity measure is symmetric, that is, $p_{(i,j)} = p_{(j,i)}$

PROXIMITY CALCULATION

- Proximity calculation to compute $p_{(i,j)}$ is different for different types of attributes according to NOIR topology.

Proximity calculation for Nominal attributes:

- For example, binary attribute, Gender = {Male, female} where Male is equivalent to binary 1 and female is equivalent to binary 0.
- Similarity value is 1 if the two objects contains the same attribute value, while similarity value is 0 implies objects are not at all similar.

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

- Here, Similarity value let it be denoted by p, among different objects are as follows.

$$p(Ram, sita) = 0$$

$$p(Ram, Laxman) = 1$$

Note : In this case, if q denotes the dissimilarity between two objects $i \text{ and } j$ with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

PROXIMITY CALCULATION

- Now, let us focus on how to calculate proximity measures between objects which are defined by two or more binary attributes.
- Suppose, the number of attributes be b. We can define the contingency table summarizing the different matches and mismatches between any two objects x and y, which are as follows.

Table 12.3: Contingency table with binary attributes

		Object y	
		1	0
Object x	1	f_{11}	f_{10}
	0	f_{01}	f_{00}

Here, f_{11}= the number of attributes where x=1 and y=1.

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

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

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

Note : $f_{00} + f_{01} + f_{10} + f_{11} = b$, the total number of binary attributes.

Now, two cases may arise: symmetric and asymmetric binary attributes.

SIMILARITY MEASURE WITH SYMMETRIC BINARY

- To measure the similarity between two objects defined by symmetric binary attributes using a measure called symmetric binary coefficient and denoted as $\mathcal{S}$ and defined below

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

The dissimilarity measure, likewise can be denoted as $\mathcal{D}$ and defined as

$$\mathcal{D} = \frac{Number\ of\ mismatched\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Note that, $\mathcal{D} = 1 - \mathcal{S}$

SIMILARITY MEASURE WITH SYMMETRIC BINARY

Example 12.2: Proximity measures with symmetric binary attributes

Consider the following two dataset, where objects are defined with symmetric binary attributes.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I}, Hobby = {T, C}, Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{S}(\text{Hari, Ram}) = \frac{1+2}{1+2+1+2} = 0.5$$

PROXIMITY MEASURE WITH ASYMMETRIC BINARY

- Such a similarity measure between two objects defined by asymmetric binary attributes is done by Jaccard Coefficient and which is often symbolized by $\mathcal{J}$ is given by the following equation

$$\mathcal{J} = \frac{Number\ of\ matching\ presence}{Number\ of\ attributes\ not\ involved\ in\ 00\ matching}$$

or

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity Measure with Asymmetric Binary

Example 12.3: Jaccard Coefficient

Consider the following two dataset.
Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I}, Hobby = {T, C}, Job = {Y, N}

Calculate the Jaccard coefficient between Ram and Hari assuming that all binary attributes are asymmetric and for each pair values for an attribute, first one is more frequent than the second.

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

Note: $\mathcal{J}$(Ram, Tomi) = 0 and $\mathcal{J}$(Hari, Ram) = $\mathcal{J}$(Ram, Hari), etc.

PROXIMITY MEASURE WITH CATEGORICAL ATTRIBUTE

- Binary attribute is a special kind of nominal attribute where the attribute has values with two states only.
- On the other hand, categorical attribute is another kind of nominal attribute where it has values with three or more states (e.g. color = {Red, Green, Blue}).
- If $s(x, y)$ denotes the similarity between two objects x and y, then

$$s(x, y) = \frac{Number\ of\ matches}{Total\ number\ of\ attributes}$$

and the dissimilarity $d(x, y)$ is

$$d(x, y) = \frac{Number\ of\ mismatches}{Total\ number\ of\ attributes}$$

- If m = number of matches and a = number of categorical attributes with which objects are defined as

$$s(x, y) = \frac{m}{a} \quad \text{and} \quad d(x, y) = \frac{a-m}{a}$$

PROXIMITY MEASURE WITH ORDINAL ATTRIBUTE

- Ordinal attribute is a special kind of categorical attribute, where the values of attribute follows a sequence (ordering) e.g. Grade = {Ex, A, B, C} where Ex > A >B >C.
- Suppose, A is an attribute of type ordinal and the set of values of $A = \{a_1, a_2, \ldots\ldots, a_n\}$. Let n values of A are ordered in ascending order as $a_1 < a_2 < \ldots < a_n$. Let *i-th* attribute value a_i be ranked as *i, i=1,2,..n*.
- The normalized value of a_i can be expressed as

$$\hat{a}_i = \frac{i-1}{n-1}$$

- Thus, normalized values lie in the range $[0..1]$.
- As a_i is a numerical value, the similarity measure, then can be calculated using any similarity measurement method for numerical attribute.
- For example, the similarity measure between two objects $x \text{ and } y$ with attribute values a_i and a_j, then can be expressed as

$$s(x, y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

where $\hat{a}_i$ and $\hat{a}_i$ are the normalized values of $\hat{a}_i$ and $\hat{a}_i$, respectively.

PROXIMITY MEASURE WITH ORDINAL ATTRIBUTE

Example 12.5:

Consider the following set of records, where each record is defined by two ordinal attributes size={S, M, L} and Quality = {Ex, A, B, C} such that S<M<L and Ex>A>B>C.

Object	Size	Quality
A	S (0.0)	A (0.66)
B	L (1.0)	Ex (1.0)
C	L (1.0)	C (0.0)
D	M (0.5)	B (0.33)

- Normalized values are shown in brackets.
- Their similarity measures are shown in the similarity matrix below.

$$\begin{array}{cc} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ ? & & 0 & 0 \\ & & & 0 \end{bmatrix} \end{array}$$

Find the dissimilarity matrix, when each object is defined by only one ordinal attribute say size (or quality).

Proximity Measure with Interval Scale

- The measure called distance is usually referred to estimate the similarity between two objects defined with interval-scaled attributes.
- We first present a generic formula to express distance d between two objects $x \text{ and } y$ in n-dimensional space. Suppose, $x_i \text{ and } y_i$ denote the values of i^{th} attribute of the objects $x \text{ and } y$ respectively.

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{\frac{1}{r}}$$

- Here, r is any integer value.
- This distance metric most popularly known as Minkowski metric.

Proximity Measure with Interval Scale

Depending on the value of r, the distance measure is renamed accordingly.

1. **Manhattan distance (L_1 Norm: $r = 1$)**

The Manhattan distance is expressed as

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

where $|\ldots|$ denotes the absolute value. This metric is also alternatively termed as **Taxicals metric, city-block metric**.

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance is $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

- As a special instance of Manhattan distance, when attribute values $\in [0, 1]$ is called Hamming distance.
- Alternatively, Hamming distance is the number of bits that are different between two objects that have only binary values (i.e. between two binary vectors).

Proximity Measure with Interval Scale

2. Euclidean Distance (L_2 Norm: r = 2)

This metric is same as Euclidean distance between any two points x *and* y *in* $\mathcal{R}^n$.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Euclidean distance between x *and* y is

$$d(x, y) = \sqrt{(7-3)^2 + (3-2)^2 + (5-6)^2} = \sqrt{18} \approx 2.426$$

Proximity Measure with Interval Scale

3. Chebychev Distance (L_∞ Norm: $r \in \mathcal{R}$)

This metric is defined as

$$d(x, y) = \max_{\forall i}\{|x_i - y_i|\}$$

- We may clearly note the difference between Chebychev metric and Manhattan distance. That is, instead of summing up the absolute difference (in Manhattan distance), we simply take the maximum of the absolute differences (in Chebychev distance). Hence, $\mathbf{L_\infty < L_1}$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance = $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

The chebychev distance = $\text{Max}\ \{|7 - 3|, |3 - 2|, |5 - 6|\} = 4.$

PROXIMITY MEASURE FOR RATIO-SCALE

The proximity between the objects with ratio-scaled variable can be carried with the following steps:

1. Apply the appropriate transformation to the data to bring it into a linear scale. (e.g. logarithmic transformation to data of the form $X = Ae^{B}$.

2. The transformed values can be treated as interval-scaled values. Any distance measure discussed for interval-scaled variable can be applied to measure the similarity.

UNIVERSITY ASKED QUESTIONS

May 2019

Briefly outline with example, how to compute the dissimilarity between objects described by the following: [10]

i. Nominal attributes

ii. Asymmetric binary attributes

THANK YOU