SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

COURSE NAME : 19CSO302 & Introduction to Data Analytics

III YEAR/ VI SEMESTER

UNIT – I INTRODUCTION

Topic: Data & Relations

P.Poonkodi

Assistant Professor

Department of Computer Science and Engineering

Data and Relations

- Data Set
- Data Scales
- Set and Matrix Representations
- Relations
- Similarity Measures
- Dissimilarity Measures
- Sequence Relations
- Sampling and Quantization

Data Set

- It is a collection of data
- Represented in table with corresponds to row and column
- Data is collected through observations, measurements, study or analytics
- **Types**
  - Numeric dataset
  - Bivariate dataset
  - Multivariate dataset
  - Categorical dataset
  - Correction dataset

Data Scales

- A scale is a device or an object used to measure or quantify any event or another object
- Measures data
- Types
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal

- Nominal scales are used for labeling variables, without any quantitative value.
- “Nominal” scales could simply be called “labels”
- Examples

What is your gender?

- (•) M – Male
- ( ) F – Female

What is your hair color?

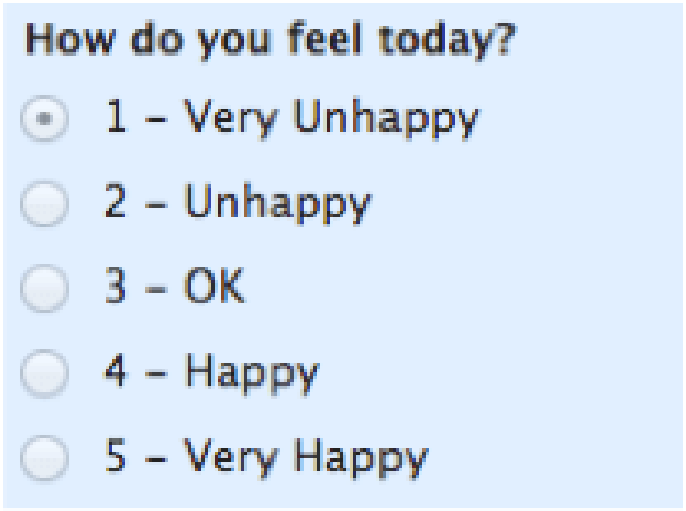
- (•) 1 – Brown
- ( ) 2 – Black
- ( ) 3 – Blonde
- ( ) 4 – Gray
- ( ) 5 – Other

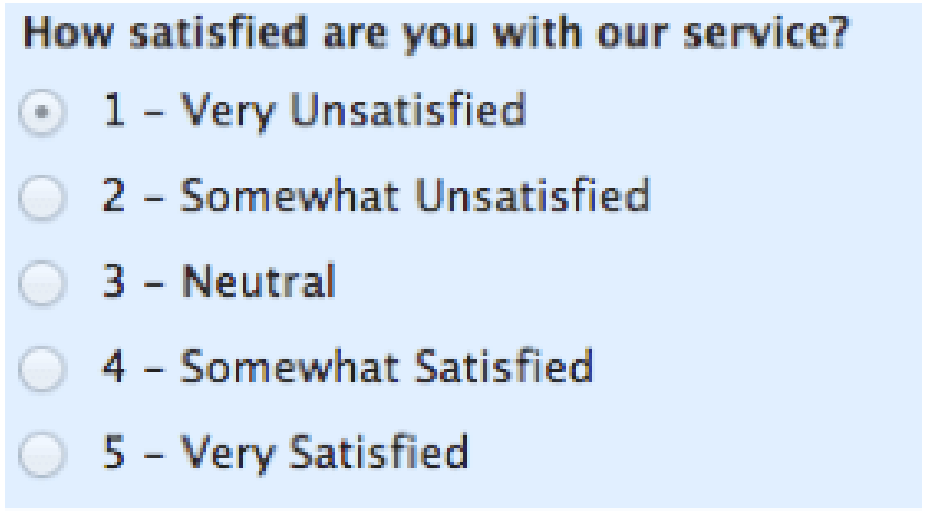
Where do you live?

- (•) A – North of the equator
- ( ) B – South of the equator
- ( ) C – Neither: In the international space station

Ordinal

- Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.
- “Ordinal” is easy to remember because is sounds like “order” and that’s the key to remember with “ordinal scales
- *Advanced note*: The best way to determine *central tendency* on a set of ordinal data is to use the **mode or median**; a purist will tell you that the mean cannot be defined from an ordinal set.





Interval

- Interval scales are numeric scales in which we know both the order and the exact differences between the values
- The classic example of an interval scale is Celsius temperature because the difference between each value is the same.
- For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

Ratio

- Ratio scales are the ultimate nirvana when it comes to data measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an absolute zero– which allows for a wide range of both descriptive and inferential statistics to be applied
- An example of a ratio scale is:

  What is your weight in Kgs?

  - Less than 55 kgs
  - 55 – 75 kgs
  - 76 – 85 kgs
  - 86 – 95 kgs
  - More than 95 kgs

The Four Scales of Measurement

Nominal Scale

Used for naming variables in no particular order

For example, eye colour

Ordinal Scale

Used for variables in ranked order, but the difference between is not determined

For example, #1 happy, #2 neutral, #3 unhappy

Interval Scale

Used for numerical variables with known equal intervals of the same distance

For example, time

Ratio Scale

Used for variables on a scale that have measurable intervals

For example, weight

Summary

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Similarity Measure

- It is a numerical measure of how alike two data objects are
- higher when objects are more alike
- often falls in the range [0,1]

Similarity might be used to identify

- duplicate data that may have differences due to typos
- equivalent instances from different data sets. E.g. names and/or addresses that are the same but have misspellings
- groups of data that are very close (clusters)

Dissimilarity Measure

- It is a numerical measure of how different two data objects are
- lower when objects are more alike
- minimum dissimilarity is often 0 while the upper limit varies depending on how much variation can be

Dissimilarity might be used to identify

- outliers
- interesting exceptions, e.g. credit card fraud
- boundaries to clusters

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	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Nominal is binary if two values are equal or not

Ordinal is the difference between two values, normalized by the maximum distance

Quantitative dissimilarity is just a distance between, similarity attempts to scale that distance to [0,1]

Distance between instances with multiple attributes.

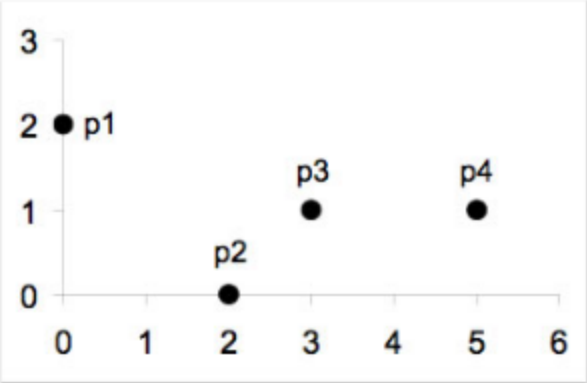
Attributes are naturally numbers or ordinal, but nominal must resort to the binary 0 or 1 if match or not

Euclidean distance

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

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the *k-th* attributes (components) or data objects **x** and **y**

Standardization/normalization may be necessary to ensure an attribute does not skew the distances due to different scales.



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

is a generalization of Euclidean Distance

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

where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the *k-th* attributes (components) or data objects **x** and **y**.

Examples

r = 1. "City block", "Manhattan", "taxicab", L_1 norm distance.

- Another 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 (L_2 norm)

r = ∞. "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 distance measures are defined for all numbers of dimensions.

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

Distance Matrix

References

- Runkler TA, “Data Analytics: Models and algorithms for intelligent data analysis”, Springer, Third Edition 2020


Thank You!