IIT ROORKEE

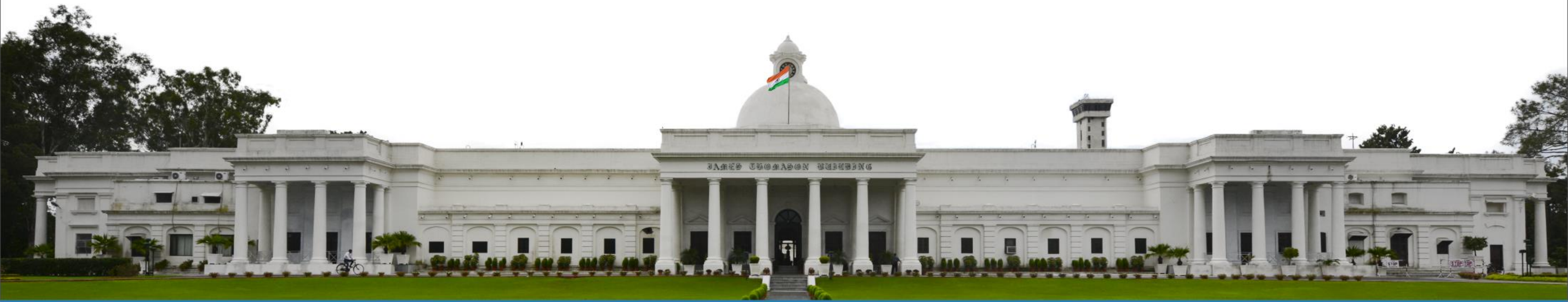
FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

NPTEL

NPTEL ONLINE
CERTIFICATION COURSE

Cluster analysis: Part - IV

Dr. A. Ramesh
DEPARTMENT OF MANAGEMENT STUDIES



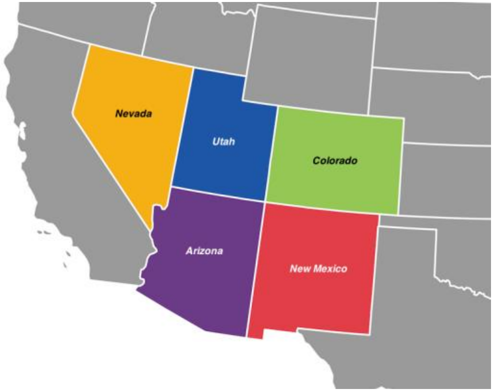
Agenda

- How to handle the following types of variables :
  - Interval scale variable
  - Binary variables
  - Categorical Variables
  - Ordinal Variables
  - Ratio-Scaled Variables
  - Variables of mixed type

Categorical Variables

Categorical Variables

- A categorical variable is a generalization of the binary variable in that it can take on more than two states or values.
- For example, map color is a categorical variable that may have, say, five states: red, yellow, green, purple, and blue



Categorical Variables

- Let the number of states of a categorical variable be M
- The states can be denoted by letters, symbols, or a set of integers, such as 1, 2,..., M
- Notice that such integers are used just for data handling and do not represent any specific ordering

Categorical Variables

- “How is dissimilarity computed between objects described by categorical variables?”

Categorical Variables

- The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i, j) = \frac{p - m}{p},$$

- 'm' represents the number of matches (i.e., the number of variables for which objects i and j are in the same state), and 'p' denotes the total number of variables.
- Weights can also be assigned to modify the influence of 'm' or to give greater importance to matches in variables with a larger number of states. This can be achieved by multiplying 'm' by a weight factor before computing the dissimilarity.
- Assigning weights allows for customization of the dissimilarity calculation, enabling researchers to emphasize certain variables or characteristics based on their importance in the analysis.

$$D(i, j) = \frac{p - w \cdot m}{p}$$

Dissimilarity between categorical variables

- Suppose that we have the sample data as shown in the table
- Let only the object-identifier and the variable (or attribute) test-1 are available, which is a categorical data

object identifier	*test-1 (categorical)*	*test-2 (ordinal)*	*test-3 (ratio-scaled)*
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Finding Groups in Data: An Introduction to Cluster Analysis
Author(s): Leonard Kaufman, Peter J. Rousseeuw
March 1990, John Wiley & Sons, Inc.

Dissimilarity matrix

$$
\begin{array}{c c}
 & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} &
\begin{bmatrix}
0 & & & \\
d(2,1) & 0 & & \\
d(3,1) & d(3,2) & 0 & \\
d(4,1) & d(4,2) & d(4,3) & 0
\end{bmatrix}
\end{array}
$$

Dissimilarity between categorical variables

- Since here we have one categorical variable, test-1, we set p = 1 in Equation $$d(i, j) = \frac{p - m}{p},$$

So that d(i, j) evaluates to '0' if objects i and j match, and '1' if the objects differ

- Thus, we get

$$\begin{bmatrix} 0 & & & \\ \boxed{1} & 0 & & \\ 1 & 1 & 0 & \\ \boxed{0} & 1 & 1 & 0 \end{bmatrix}$$

d(2,1) = (1-0)/1 = 1
d(4,1) = (1-1)/1 = 0

object identifier	*test-1 (categorical)*
1	code-A
2	code-B
3	code-C
4	code-A

Ordinal Variables

- A discrete ordinal variable resembles a categorical variable, except that the ‘M’ states of the ordinal value are ordered in a meaningful sequence
- Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively

  V. good, hard, bad
  1 2 3
- For example, professional ranks are often enumerated in a sequential order, such as Assistant, Associate, and full for Professors
- A continuous ordinal variable looks like a set of continuous data of an unknown scale; that is, the relative ordering of the values is essential but their actual magnitude is not

  94 – 1
  50 – 2

Ordinal Variables

- For example, the relative ranking in a particular sport (e.g., gold, silver, bronze) is often more essential than the actual values of a particular measure
- Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes
- The values of an ordinal variable can be mapped to ranks

Dissimilarity computation

- The treatment of ordinal variables is quite similar to that of interval-scaled variables when computing the dissimilarity between objects
- Suppose that 'f' is a variable from a set of ordinal variables describing 'n ' objects
- The dissimilarity computation with respect to 'f' involves the following steps:
- The value of 'f' for the i^{th} object is x_{if}, and 'f' has M_f ordered states, representing the ranking $1, \ldots, M_f$.
- Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$.

Dissimilarity computation

	A	B	C	D	E	F	G
B	69.8						
C	1 $\boxed{2.0}$	70.8					
D	71.6	3 $\boxed{5.7}$	72.5				
E	108.6	42.2	109.9	43.9			
F	95.7	26.3	96.8	26.4	19.1		
G	5.8	75.7	2 $\boxed{5.1}$	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

Standardization of ordinal variable

- Since each ordinal variable can have a different number of states, it is often necessary to ~~map the~~ range of each variable onto [0.0,1.0] so that each variable has equal weight.
- This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variable by:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Dissimilarity computation

- Dissimilarity can then be computed using any of the distance measures described earlier (like that for interval data)

Example

- Suppose that we have the sample data of the following Table ,
- Except that this time only the object-identifier and the continuous ordinal variable, test-2, are available
- There are three states for test-2, namely fair, good, and excellent, that is Mf = 3

object identifier	*test-1 (categorical)*		*test-2 (ordinal)*	*test-3 (ratio-scaled)*
1	code-A	3	excellent	445
2	code-B	1	fair	22
3	code-C	2	good	164
4	code-A	3	excellent	1,210

Example

- For step 1, if we replace each value for test-2 by its rank, the four objects are assigned the rank $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, respectively

- Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0
- For step 3, we can use, say, the Euclidean distance, which results in the following dissimilarity matrix:

Dissimilarity computation

Object identifier   Ordinal data   Z

$1 \rightarrow 3 \rightarrow 1$

$2 \rightarrow 1 \rightarrow 0$

$3 \rightarrow 2 \rightarrow 0.5$

$4 \rightarrow 3 \rightarrow 1$

$$
\begin{array}{c c}
 & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} &
\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ \underline{0.5} & \underline{0.5} & 0 & \\ \underline{0} & \underline{1.0} & \underline{0.5} & 0 \end{bmatrix}
\end{array}
$$

Ratio-Scaled Variables

- A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately following the formula

$$Ae^{Bt} \quad \text{or} \quad Ae^{-Bt}$$

where A and B are positive constants, and t typically represents time

- Common examples include the growth of a bacteria population or the decay of a radioactive element

Computing the dissimilarity between objects

- There are three methods to handle ratio-scaled variables for computing the dissimilarity between objects:

1. Treat ratio-scaled variables like interval-scaled variables
   - This, however, is not usually a good choice since it is likely that the scale may be distorted
2. Apply logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula $y_{if} = \log(x_{i\,f})$
   - The y_{if} values can be treated as interval valued, Notice that for some ratio-scaled variables, log-log or other transformations may be applied, depending on the variable's definition and the application

Computing the dissimilarity between objects

3. Treat x_{if} as continuous ordinal data and treat their ranks as interval-valued

- The latter two methods are the most effective, although the choice of method used may depend on the given application

Example

- This time, we have the sample data of the following Table,
- Except that only the object-identifier and the ratio-scaled variable, test-3, are available

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Example

- Let's try a logarithmic transformation 2 3 4
- Taking the log of test-3 results in the values 2.65, 1.34, 2.21, and 3.08 for the objects 1 to 4, respectively
- Using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix:

$$
\begin{array}{c c}
 & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}
\end{array}
$$

object identifier	test-3 (ratio-scaled)
1	445
2	22
3	164
4	1,210

Variables of Mixed Types

- So far we have discussed how to compute the dissimilarity between objects described by variables of the same type, where these types may be either interval-scaled, symmetric binary, asymmetric binary, categorical, ordinal, or ratio-scaled
- However, in many real databases, objects are described by a mixture of variable types

Variables of Mixed Types

- In general, a database can contain all of the six variable types listed above
- "So, how can we compute the dissimilarity between objects of mixed variable types?"
- One approach is to group each kind of variable together, performing a separate cluster analysis for each variable type
  - This is feasible if these analyses derive compatible results
  - However, in real applications, it is unlikely that a separate cluster analysis per variable type will generate compatible results

Variables of Mixed Types

- A more preferable approach is to process all variable types together, performing a single cluster analysis
- One such technique combines the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval [0.0,1.0]

Variables of Mixed Types

- Suppose that the data set contains p variables of mixed type
- The dissimilarity d(i, j) between objects i and j is defined as

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

where the indicator $\delta_{ij}^{(f)} = 0$ if either

- x_{if} or x_{jf} is missing (i.e., there is no measurement of variable f for object i or object j), or $x_{if} = x_{jf} = 0$ and variable f is asymmetric binary;
- otherwise, $\delta_{ij}^{(f)} = 1$

Variables of Mixed Types

- The contribution of variable f to the dissimilarity between i and j, that is, $d_{ij}^{(f)}$, is computed dependent on its type:
- If 'f' is interval-based: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$,

  where h runs overall non missing objects for variable f
- If 'f' is binary or categorical: $d_{ij}^{(f)} = 0$, if $x_{if} = x_{jf}$
  - otherwise $d_{ij}^{(f)} = 1$

Variables of Mixed Types

- If 'f' is ordinal: compute the ranks r_{if} and $z_{if} = r_{if}-1 / M_f-1$, and treat z_{if} as interval scaled
- If 'f' is ratio-scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat 'f' as continuous ordinal data, compute r_{if} and z_{if}, and then treat z_{if} as interval-scaled
- The above steps are identical to what we have already seen for each of the individual variable types

Variables of Mixed Types

- The only difference is for interval-based variables, where here we normalize so that the values map to the interval [0.0,1.0]
- Thus, the dissimilarity between objects can be computed even when the variables describing the objects are of different types