IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Cluster analysis: Part - III

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Clustering analysis part III

Agenda

- Handling missing data
- Calculation of similarity and dissimilarity matrix

Handling missing data

- It often happens that not all measurements are actually available, so there are some “holes” in the data matrix
- Such an absent measurement is called a missing value and it may have several causes
- The value of the measurement may have been lost or it may not have been recorded at all by oversight or lack of time

Handling missing data

- Sometimes the information is simply not available, for example the birth date of an orphan, or the patient may not remember whether he or she ever had the measles, or it may be impossible to measure the desired quantity due to the malfunctioning of some instrument
- In certain instances the question does not apply (such as the colour of hair of a bald person) or there may be more than one possible answer (when two experimenters obtain very different results)

Handling missing data

- How can we handle a data set with missing values?
- If there exists an object in the data set for which all measurements are missing, there is really no information on this object so it has to be deleted
- Analogously, a variable consisting exclusively of missing values has to be removed too

Handling missing data

- If the data are standardized, the mean value m, of the f^{th} variable is calculated by making use of the present values only
- The same goes for s_f,

$$s_f = \frac{1}{n}\{|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|\}$$

In the denominator , we must replace 'n' by the number of non missing values for that variable

- But of course only when the corresponding x_i, is not missing itself

Handling missing data

- In the computation of distances (based on either the x_i, or the z_i,) similar precautions must be taken
- When calculating the distances d(i, j), only those variables are considered in the sum for which the measurements for both objects are present subsequently the sum is multiplied by p and divided by the actual number of terms (in the case of Euclidean distances this is done before taking the square root)
- Such a procedure only makes sense when the variables are thought of as having the same weight (for instance, this can be done after standardization)

Handling missing data

- When computing these distances, one might come across a pair of objects that do not have any common measured variables, so their distance cannot be computed by means of the above mentioned approach.
- Several remedies are possible: One could remove either object or one could fill in some average distance value based on the rest of the data
- Or by replacing all missing x_{if} by the mean mf of that variable; then all distances can be computed
- Applying any of these methods, one finally possesses a “full” set of distances

Dissimilarities

- The entries of a n-by n matrix may be Euclidean or Manhattan distances
- However, there are many other possibilities, so we no longer speak of distances but of dissimilarities (or dissimilarity coefficients)
- Basically, dissimilarities are non-negative numbers d(i, j) that are small (close to zero) when i and j are "near" to each other and that become large when i and j are very different
- We shall usually assume that dissimilarities are symmetric and that the dissimilarity of an object to itself is zero, but in general the triangle inequality does not hold

Dissimilarities

- Dissimilarities can be obtained in several ways.
- Often they can be computed from variables that are binary, nominal, ordinal, interval, or a combination of these
- Also, dissimilarities can be simple subjective ratings of how much certain objects differ from each other, from the point of view of one or more observers
- This kind of data is typical in the social sciences and in marketing

Example

- Fourteen postgraduate economics students (coming from different parts of the world) were asked to indicate the subjective dissimilarities between 11 scientific disciplines.
- All of them had to fill in a matrix like Table 4, where the dissimilarities had to be given as integer numbers on a scale from 0 (identical) to 10 (very different)
- The actual entries of the Table in next slide, are the averages of the values given by the students

Example

- It appears that the smallest dissimilarity is perceived between mathematics and computer science (1.43), whereas the most remote fields were psychology and astronomy (9.36)

Astronomy	0.00										
Biology	7.86	0.00									
Chemistry	6.50	2.93	0.00								
Computer sci.	5.00	6.86	6.50	0.00							
Economics	8.00	8.14	8.21	4.79	0.00						
Geography	4.29	7.00	7.64	7.71	5.93	0.00					
History	8.07	8.14	8.71	8.57	5.86	3.86	0.00				
Mathematics	3.64	7.14	4.43	1.43	3.57	7.07	9.07	0.00			
Medicine	8.21	2.50	2.93	6.36	8.43	7.86	8.43	6.29	0.00		
Physics	2.71	5.21	4.57	4.21	8.36	7.29	8.64	2.21	5.07	0.00	
Psychology	9.36	5.57	7.29	7.21	6.86	8.29	7.64	8.71	3.79	8.64	0.00

Dissimilarities

- when conducting cluster analysis on a set of variables, various measures of dissimilarity can be employed. One common approach involves computing either the parametric Pearson product moment correlation or the non-parametric Spearman correlation between the variables of interest, such as f and g.
- The Pearson correlation is suitable when the relationship between variables is linear and both variables follow a normal distribution. On the other hand, the Spearman correlation is more robust to non-linear relationships and does not require the assumption of normality, making it suitable for ordinal or non-normally distributed data.
- Selecting the appropriate measure depends on the nature of the variables and the assumptions underlying the analysis

Dissimilarities

- Both coefficients lie between - 1 and + 1 and do not depend on the choice of measurement units
- The main distinction between the Pearson and Spearman coefficients lies in the types of relationships they assess between variables. The Pearson coefficient primarily examines linear relationships between variables, measuring the strength and direction of their linear association. On the other hand, the Spearman coefficient evaluates monotonic relationships, which can be any consistently increasing or decreasing pattern, not necessarily strictly linear.

$$\rho_{(x,y)} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

$$R(f,g) = \frac{\sum_{i=1}^{n}(x_{if} - m_f)(x_{ig} - m_g)}{\sqrt{\sum_{i=1}^{n}(x_{if} - m_f)^2}\sqrt{\sum_{i=1}^{n}(x_{ig} - m_g)^2}}$$

Dissimilarities

- Correlation coefficients are useful for clustering purposes because they measure the extent to which two variables are related
- Correlation coefficients, whether parametric or nonparametric, can be converted to dissimilarities d(f, g), for instance by setting

$$d(f, g) = (1 - R(f, g))/2$$

With this formula, variables with a high positive correlation receive a dissimilarity coefficient close to zero, whereas variables with a strongly negative correlation will be considered very dissimilar

Similarities

- The more objects i and j are alike (or close), the larger s(i, j) becomes
- Such a similarity s(i, j) typically takes on values between 0 and 1, where 0 means that i and j are not similar at all and 1 reflects maximal similarity
- Values in between 0 and 1 indicate various degrees of resemblance
- Often it is assumed that the following conditions hold:

(S1) $0 \leq s(i, j) \leq 1$

(S2) $s(i, i) = 1$

(S3) $s(i, j) = s(j, i)$

Similarities

- For all objects i and j , the numbers s(i, j) can be arranged in an n-by-n matrix ,which is then called a similarity matrix
- Both similarity and dissimilarity matrices are generally referred to as proximity matrices, or sometimes as resemblance matrices.
- In order to define similarities between variables, we can again resort to the Pearson or the Spearman correlation coefficient
- However, neither correlation measure can be used directly as a similarity coefficient because they also take on negative values

Similarities

- Some transformation is in order to bring the coefficients into the zero-one range
- There are essentially two ways to do this, depending on the meaning of the data and the purpose of the application
- If variables with a strong negative correlation are considered to be very different because they are oriented in the opposite direction (like mileage and weight of a set of cars), then it is best to take something like the following:

$$s(f, g) = (1 + R(f, g))/2$$

which yields s(f, g) = 0 whenever R(f, g) = - 1.

Similarities

- There are situations in which variables with a strong negative correlation should be grouped, because they measure essentially the same thing
- For instance, this happens if one wants to reduce the number of variables in a regression data set by selecting one variable from each cluster
- In that case it is better to use a formula like

$$s(f, g) = |R(f, g)|$$

which yields s(f, g) = 1 when R(f, g) = -1

Similarities

- Suppose the data consist of a similarity matrix but one wants to apply a clustering algorithm designed for dissimilarities
- Then it is necessary to transform the similarities into dissimilarities
- The larger the similarity s(i, j) between i and j, the smaller their dissimilarity d(i, j) should be
- Therefore, we need a decreasing transformation, such as

$$d(i, j) = 1 - s(i, j)$$

Binary Variables

- A contingency table for binary variables.

		object j		
		1	0	sum
	1	q	$\boxed{r}$	$q+r$
object i	0	$\boxed{s}$	t	$s+t$
	sum	$q+s$	$r+t$	$\boxed{p}$

Dissimilarity between two binary variables

- q → is the number of variables that equal 1 for both objects *i* and *j*,
- r → is the number of variables that equal 1 for object *i* but that are 0 for object *j*,
- S → is the number of variables that equal 0 for object *i* but equal 1 for object *j*, and
- t → is the number of variables that equal 0 for both objects *i* and *j*.
- The total number of variables is p, where $p = q+r+s+t$.

Symmetric Binary Dissimilarity

$$d(i, j) = \frac{r + s}{q + r + s + t}.$$

Gender

0 - male

1 - female

Asymmetric binary variable

- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the *positive* and *negative* outcomes of a disease *test*.
- By convention, we shall code the most important outcome, which is usually the rarest one, by 1 (e.g., *HIV positive*) and the other by 0 (e.g., *HIV negative*).
- Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match).
- Therefore, such binary variables are often considered “monary” (as if having one state).

asymmetric binary dissimilarity

		object j		
		1	0	sum
	1	q	r	$q+r$
object i	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

$$d(i, j) = \frac{r+s}{q+r+s}.$$

Jaccard coefficient

$$sim(i, j) = \frac{q}{q + r + s} = 1 - d(i, j).$$

Dissimilarity between binary variables

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Dissimilarity between Jack and Marry

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Jack

Marry

	1	0
1	2	1
0	0	3

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

Dissimilarity between Jack and Jim

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Jim

Jack

	1	0
1	1	1
0	1	3

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

Dissimilarity between Jim and Marry

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Jim

Marry

	1	0
1	1	2
0	1	2

$$d(Mary, Jim) = \frac{1+2}{1+1+2} = 0.75$$

Thank you