Bingbing Li
18210705

Data Mining COMP40370
Practical 8

Date 19/11/2019

Question 1: Categorical Variables

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

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

where m is the number of matches and p is the total number of variables. The goal is to calculate the dissimilarity matrix (hint: note that categorical variables can be encoded by asymmetric binary variables).

1. **How many binary variables are needed for the attribute variable T1?**

The variable T1 is categorical data, so I use the dummy coding system to make the categorical data into a series of binary variables. For all but one of the levels of the categorical variable, a new variable will be created that has a value of one for each observation at that level and zero for all others.
Because variable T1 has 3 possible values which are Code-A, Code-B, Code-C, so there are 3-1= 2 binary variables which are needed for the attribute variable T1.
In order to create these variables, we are going to take 2 of the kinds of "Code type", and create a variable corresponding to each type, which will have the value of 1 or 0 . Each instance of "Code type" would then be recoded into a value for "C1," and "C2".
If a variable were a "Code A " then "C1" would be equal to 0, "C2" would be equal to 0.
If a variable were a "Code B " then "C1" would be equal to 1, "C2" would be equal to 0.
If a variable were a "Code C " then "C1" would be equal to 0, "C2" would be equal to 1.
The following figure shows the binary variables needed for each categorical variables.

Code type	C1	C2
Code-A	0	0
Code-B	1	0
Code-C	0	1
Code-A	0	0

2. **Calculate the dissimilarity matrix, showing all the steps of your calculation.**

First, we need to compute the dissimilarity of the matrix:

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{pmatrix}$$

Then according to the formula of ratio of mismatches:

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

m is the number of matches and p is the total number of variables. In this question, because there is only one categorical variable : code-type, so the p would be set to 1. So that d(i, j) evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, the dissimilarity matrix would be as follows:

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

Question 2: Ordinal Variables

1. Explain how are the ordinal variables handled.

An ordinal variable can be discrete or continuous. For example, In this question , variable T2 is discrete, so the values of an ordinal variable can be mapped to ranks. Suppose that an ordinal variable f has Mf states, then these ordered states define the ranking 1,..., Mf.

2. Describe briefly the necessary steps for handling this type of variables.

Step 1: for the ith object xif , and f has Mf ordered states, representing the ranking 1, ⋯ , Mf. Replace each xif by its corresponding rank: $rif \in \{1...Mf\}$

Step 2: 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. So we need to replace the rank rif of the ith object in the fth variable by:

$$Z\text{if} = \frac{rif - 1}{\text{Mf - 1}}$$

Step 3: Dissimilarity can then be computed using any of the distance measures described for interval-scaled variables.

3. Assume that the Euclidan distance is used as a distance measure. Calculate the dissimilarity matrix for the attribute variable T2.

In this question, firstly, I assign 1(Fair), 2(Good),3(Excellent), then I normalize the ranking 1 to 0.0, 2 to 0.5, 3 to 1.0. According to the formula of Euclidan distance which is

$$\sqrt{(q-p)^2} = |q-p|.$$

The dissimilarity matrix for the attribute variable T2 I calculate is as follows:

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Question 3: Ratio-scaled Variables

1. Explain how can you handle the dissimilarity between objects of type ratio-scaled.
There are some methods to do this: the first one is to one is applying logarithmic transformation, $yi = \log(xi)$, and the second one is to handle them as continuous ordinal data and treat their rank as interval-scaled.

2. Give the necessary steps for calculating such dissimilarity.

Step 1: Convert the ratio-scaled variables in the table using a logarithmic transformation
$yi = \log(xi)$.

Step 2: Using the Euclidean distance to generate the dissimilarity matrix.

3. Assume the the distance measure is chosen to be the Euclidian distance. Calculate the dissimilarity matrix for the attribute variable T3.

The variables after conversion:

Identifier	T3(ratio-scaled)	log(xi)
01	445	2.65
02	22	1.34
03	164	2.22
04	1210	3.08

the dissimilarity matrix:

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.43 & 0.88 & 0 & \\ 0.43 & 1.74 & 0.86 & 0 \end{bmatrix}$$

Question 4: Mixed type Variables

Firstly , I convert the data as follows:

Identifier	T1	T2	T3
01	Code-A	1	2.65
02	Code-B	0	1.34
03	Code-C	0.5	2.22
04	Code-A	1	3.08

Then I use the dissimilarity formula for mixed type variables to compute the dissimilarity between each of these identifiers:

$$\begin{bmatrix} 0 & & & \\ 1.1 & 0 & & \\ 0.64 & 0.79 & 0 & \\ 0.215 & 1.24 & 0.79 & 0 \end{bmatrix}$$