IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Cluster analysis: Part - II

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Explain effect of standardization(with help of an example)
- Different types of distances computation between the objects

Example

- Lets take four persons A, B,C, D with following age and height:

Person	Age (yr)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160

TABLE: 1

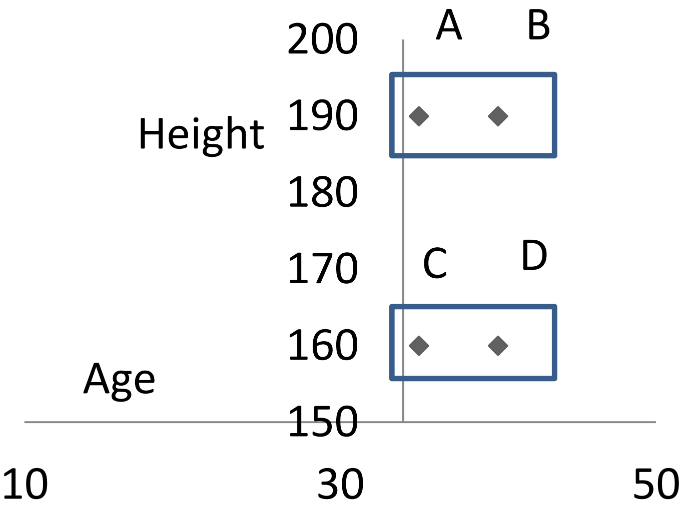


FIGURE: 1

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

Example

- In Figure 1 we can see to distinct clusters
- Let us standardize the data of Table 1
- The mean age equals m_1 = 37.5 and the mean absolute deviation of the first variable works out to be s_1 = (2.5 + 2.5 + 2.5 + 2.5)/4 = 2.5
- Therefore, standardization converts age 40 to + 1 ((40-37.5)/2.5 = 1)and age 35 ((35 - 37.5)/2.5 = -1) to – 1
- Analogously, m_2 = 175 cm and s_2 = (15 + 15 + 15 + 15)/4 = 15 cm, so 190 cm is standardized to +1 and 160 cm to - 1

Example

- The resulting data matrix, which is unitless, is given in Table 2
- Note that the new averages are zero and that the mean deviations equal 1

- Table 2

Person	Variable 1	Variable 2
A	1	1
B	-1	1
C	1	-1
D	-1	-1

- Even when the data are converted to very strange units standardization will always yield the same numbers

Example

- Plotting the values of Table 2 in Figure 2 does not give a very exciting result
- Figure 2 shows no clustering structure because the four points lie at the vertices of a square
- One could say that there are four clusters, each consisting of a single point, or that there is only one big cluster containing four points
- Here standardizing is no solution

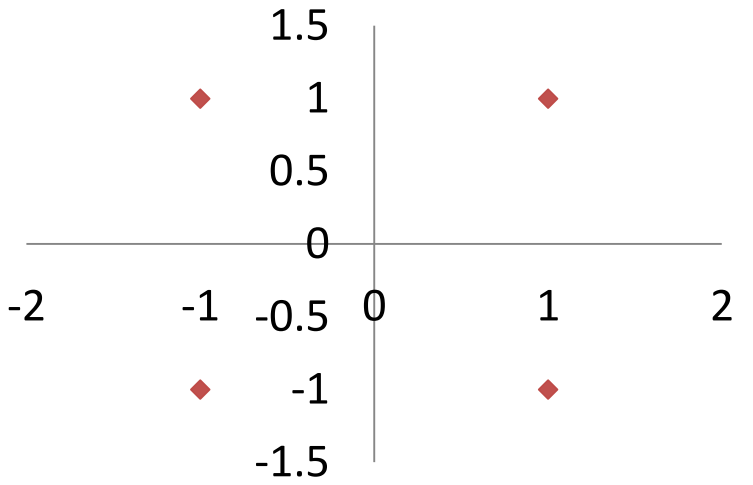


FIGURE: 2

Choice of measurement (Units)- Merits and demerits

- The choice of measurement units gives rise to relative weights of the variables
- Expressing a variable in smaller units will lead to a larger range for that variable, which will then have a large effect on the resulting structure
- On the other hand, by standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity
- As such, it may be used by a practitioner who possesses no prior knowledge

Choice of measurement- Merits and demerits

- However, it may well be that some variables are intrinsically more important than others in a particular application, and then the assignment of weights should be based on subject-matter knowledge
- On the other hand, there have been attempts to devise clustering techniques that are independent of the scale of the variables

Distances computation between the objects

- The next step is to compute distances between the objects, in order to quantify their degree of dissimilarity
- It is necessary to have a distance for each pair of objects i and j.
- The most popular choice is the Euclidean distance:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- When the data are being standardized, one has to replace all x by z in this expression
- This Formula corresponds to the true geometrical distance between the points with coordinates (x_{i1},. .., x_{ip}) and (x_{j1} ,..., x_{jp})

Example

- let us consider the special case with p = 2 (Figure 3)
- Figure shows two points with coordinates (x_{i1} , x_{i2}) and (x_{j1}, x_{j2})
- It is clear that the actual distance between objects i and j is given by the length of the hypotenuse of the triangle, yielding expression in previous slide by virtue of Pythagoras' theorem

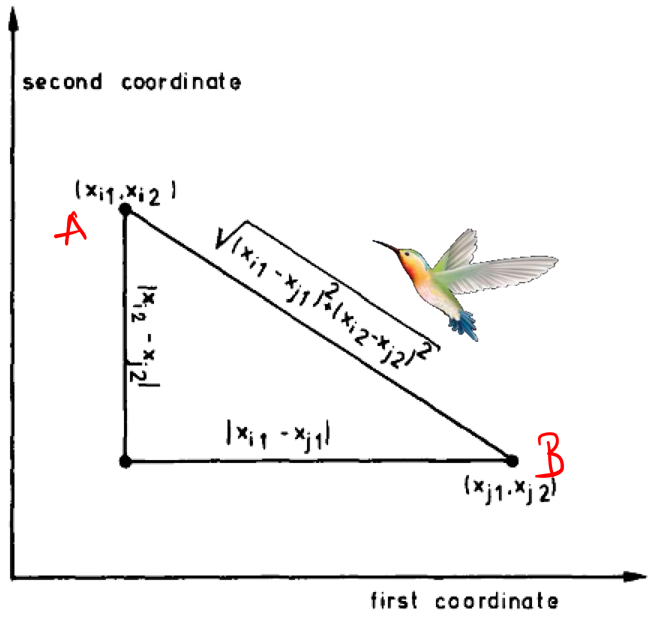
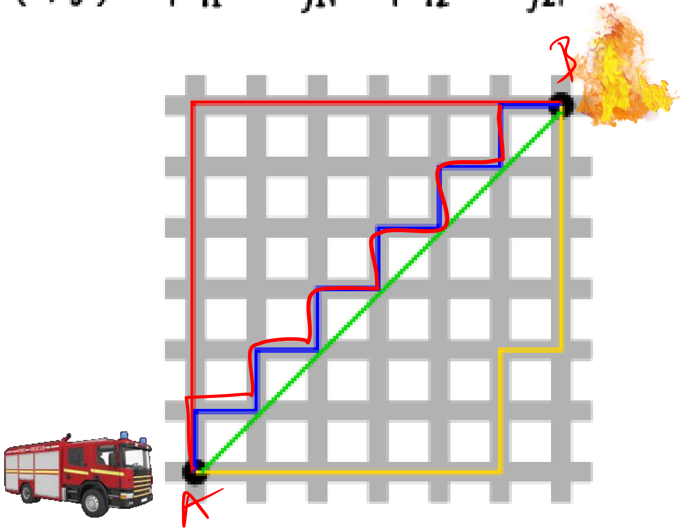


Figure 3: Illustration of the Euclidean distance formula

Distances computation between the objects

- Another well-known metric is the city block or Manhattan distance, defined by:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$



Interpretation

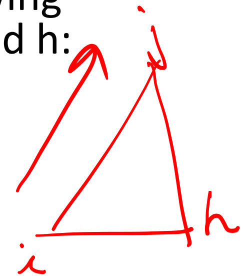
- Suppose you live in a city where the streets are all north-south or east-west, and hence perpendicular to each other
- Let Figure 3 be part of a street map of such a city, where the streets are portrayed as vertical and horizontal lines

Interpretation

- Then the actual distance you would have to travel by car to get from location i to location j would total $|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$
- This would be the shortest length among all possible paths from i to j
- Only a bird could fly straight from point i to point j, thereby covering the Euclidean distance between these points

Mathematical Requirements of a Distance Function

- Both the Euclidean metric and the Manhattan metric satisfy the following mathematical requirements of a distance function, for all objects i, j, and h:
- (D1) $d(i, j) \geq 0$
- (D2) $d(i, i) = 0$
- (D3) $d(i, j) = d(j, i)$
- (D4) $d(i, j) \leq d(i, h) + d(h, j)$
- Condition (D1) merely states that distances are nonnegative numbers and (D2) says that the distance of an object to itself is zero
- Axiom (D3) is the symmetry of the distance function
- The triangle inequality (D4) looks a little bit more complicated, but is necessary to allow a geometrical interpretation
- It says essentially that going directly from i to j is shorter than making a detour over object h



Distances computation between the objects

- If $d(i, j) = 0$ does not necessarily imply that $i = j$, because it can very well happen that two different objects have the same measurements for the variables under study
- However, the triangle inequality implies that i and j will then have the same distance to any other object h, because $d(i, h) \leq d(i, j) + d(j, h) = d(j, h)$ and at the same time $d(j, h) \leq d(j, i) + d(i, h) = d(i, h)$, which together imply that $d(i, h) = d(j, h)$

Minkowski distance

- A generalization of both the Euclidean and the Manhattan metric is the Minkowski distance given by:

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$$

Where p is any real number larger than or equal to 1

- This is also called the Lp metric, with the Euclidean (p = 2) and the Manhattan (p = 1) as special cases

Example for Calculation of Euclidean and Manhattan Distance

- Let x1 = (1, 2) and x2 = (3, 5) represent two objects as in the given Figure The Euclidean distance between the two is $\sqrt{(2^2 + 3^2)} = 3.61$. The Manhattan distance between the two is 2 + 3 = 5.

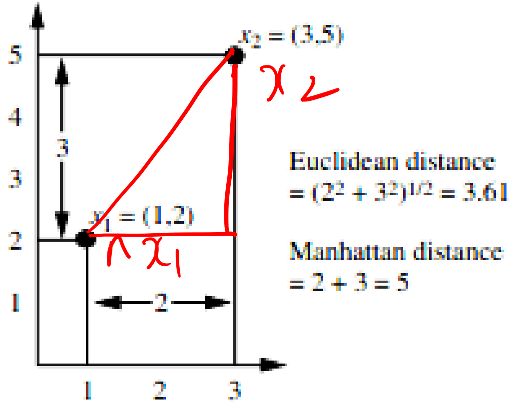


Figure: 4

n- by- n Matrix

- For example, when computing Euclidean distances between the objects of the following Table can be obtain as next slide:

Person	Weight(Kg)	Height(cm)
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78

- Euclidean distances between B and E:
- $((49 - 85)^2 + (156-178)^2)^{½} = 42.2$

n- by- n Matrix

	A	B	C	D	E	F	G	H
A	0	69.8	2.0	71.6	108.6	95.7	5.8	17.7
B	69.8	0	70.8	5.7	42.2	26.3	75.7	87.2
C	2.0	70.8	0	72.5	109.9	96.8	5.1	17.3
D	71.6	5.7	72.5	0	43.9	26.4	77.4	89.2
E	108.6	42.2	109.9	43.9	0	19.1	114.3	125.0
F	95.7	26.3	96.8	26.4	19.1	0	101.6	112.9
G	5.8	75.7	5.1	77.4	114.3	101.6	0	12.2
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2	0

Interpretation

- The distance between object B and object E can be located at the intersection of the fifth row and the second column, yielding 42.2
- The same number can also be found at the intersection of the second row and the fifth column, because the distance between B and E is equal to the distance between E and B
- Therefore, a distance matrix is always symmetric
- Moreover, note that the entries on the main diagonal are always zero, because the distance of an object to itself has to be zero

Distance matrix

- It would suffice to write down only the lower triangular half of the distance matrix

	A	B	C	D	E	F	G
B	69.8						
C	2.0	70.8					
D	71.6	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	5.1	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

Selection of variables

- It should be noted that a variable not containing any relevant information (say, the telephone number of each person) is worse than useless, because it will make the clustering less apparent.
- The Occurrence of several such “trash variables” will kill the whole clustering because they yield a lot of random terms in the distances, thereby hiding the useful information provided by the other variables.
- Therefore, such non informative variables must be given a zero weight in the analysis, which amounts to deleting them

Selection of variables

- The selection of "good" variables is a nontrivial task and may involve quite some trial and error (in addition to subject-matter knowledge and common sense)
- In this respect, cluster analysis may be considered an exploratory technique