IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Cluster analysis: Introduction - I

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

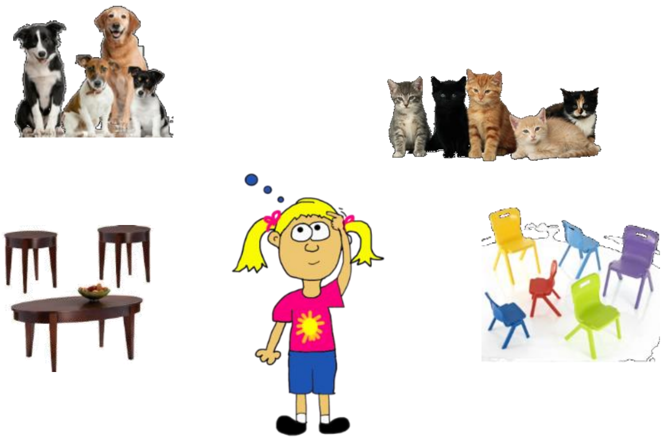
- Understanding cluster analysis and its purpose
- Introduction to types of data and how to handle them

Cluster Analysis

- Cluster analysis is the art of finding groups in data
- In cluster analysis basically, one wants to form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible

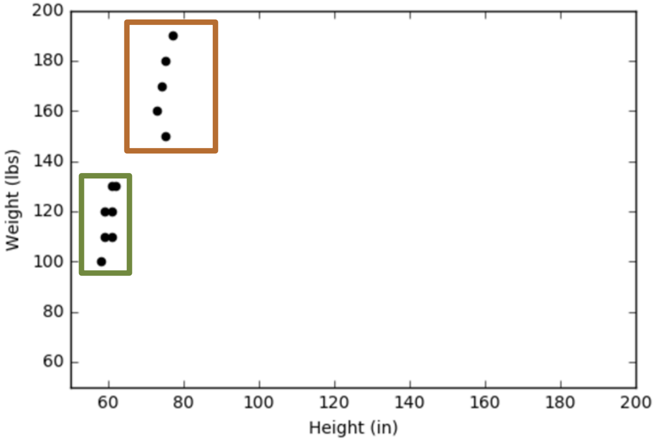
Cluster analysis

- The classification of similar objects into groups is an important human activity, this is part of the learning process
- i.e. A child learns to distinguish between cats and dogs, between tables and chairs, between men and women, by means of continuously improving subconscious classification schemes
- This explains why cluster analysis is often considered as a branch of pattern recognition and artificial intelligence



Example

- Lets illustrate with the help of an example:
- It is a plot of twelve objects, on which two variables were measured. For instance, the weight of an object might be displayed on the vertical axis and its height on the h



Example

- Because this example contains only two variables, we can investigate it by merely looking at the plot
- In this small data set there are clearly two distinct groups of objects
- Such groups are called clusters, and to discover them is the aim of cluster analysis

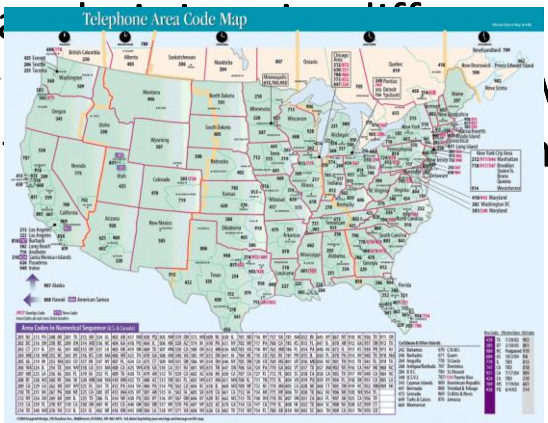
Cluster and discriminant analysis

- Cluster Analysis is an unsupervised classification technique in the sense that it is applied to a dataset where patterns want to be discovered (i.e. groups of individuals or variables want to be found)
- No prior knowledge is needed for this grouping, and it is sensitive to several decisions that have to be taken (similarity/dissimilarity measures, clustering method,...)

- Discriminant Analysis (DA) is a statistical technique used to build a prediction model that is used to classify objects from a dataset depending on the features observed on them. In this case, the dependent variable is the grouping variable, which identifies to which group and object belongs
- This grouping variable should be known at the beginning, for the function to be built up. Sometimes DA is considered as a Supervised tool, as there is a previous known classification for the elements of the dataset

Cluster analysis and discriminant analysis

- Cluster analysis can be used not only to identify a structure already present in the data, but also to impose a structure on a more or less homogeneous data set that has to be split up in a "fair" way, for instance when dividing a country into telephone areas
- Cluster a … from discriminant analysis in that it actually … whereas discriminant analysis assigns objects … ned in advance



Telephone area code for USA

Types of data and how to handle them

- Let us take an example, there are n objects to be clustered, which may be persons, flowers, words, countries, or anything
- Clustering algorithms typically operate on either of two input structures:
  - The first represents the objects by means of p measurements or attributes, such as height, weight, sex, color, and so on
  - These measurements can be arranged in an n-by-p matrix, where the rows correspond to the objects and the columns to the attributes

Example

Attributes

Objects

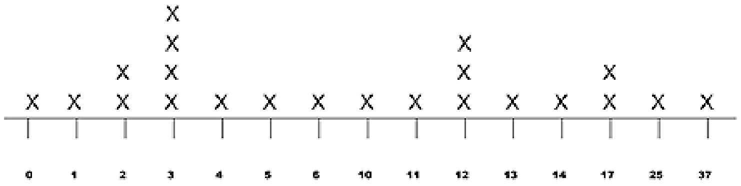
	Price	*Quality*	*Time*
Like	*A*	*B*	*B*
Intermediate	*B*	*A*	*A*
Need	*C*	*C*	*C*

Types of data and how to handle them

- The second structure is a collection of proximities that must be available for all pairs of objects
- These proximities make up an n-by-n table, which is called a one-mode matrix because the row and column entities are the same set of objects
- one shall consider two types of proximities, namely dissimilarities (which measure how far away two objects are from each other) and similarities (which measure how much they resemble each other)

Type of data

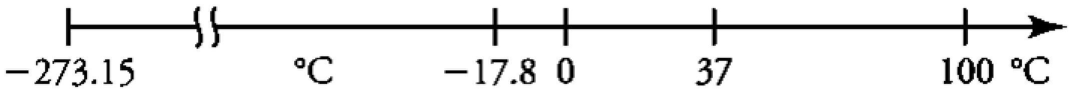
- **Interval-Scaled Variables**
- In this situation the n objects are characterized by p continuous measurements
- These values are positive or negative real numbers, such as height, weight, temperature, age, cost, ..., which follow a linear scale
- For instance, th 1910 was equal in length to that betweer



Time scale in years

Type of data

- Also, it takes the same amount of energy to heat an object of -16.4°C to -12.4°C as to increase it from 35.2°C to 39.2°C
- In general it is required that intervals keep the same importance throughout the scale



Interval-Scaled Variables

- These measurements can be organized in an n-by-p matrix, where the rows correspond to the objects (or cases) and the columns correspond to the variables.
- When the fth measurement of the ith object is denoted by x_{if} (where i = 1,. . . . , n and f = 1

$$
\begin{array}{cc}
 & p \text{ variables} \\
n \text{ objects} & \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}
\end{array}
$$

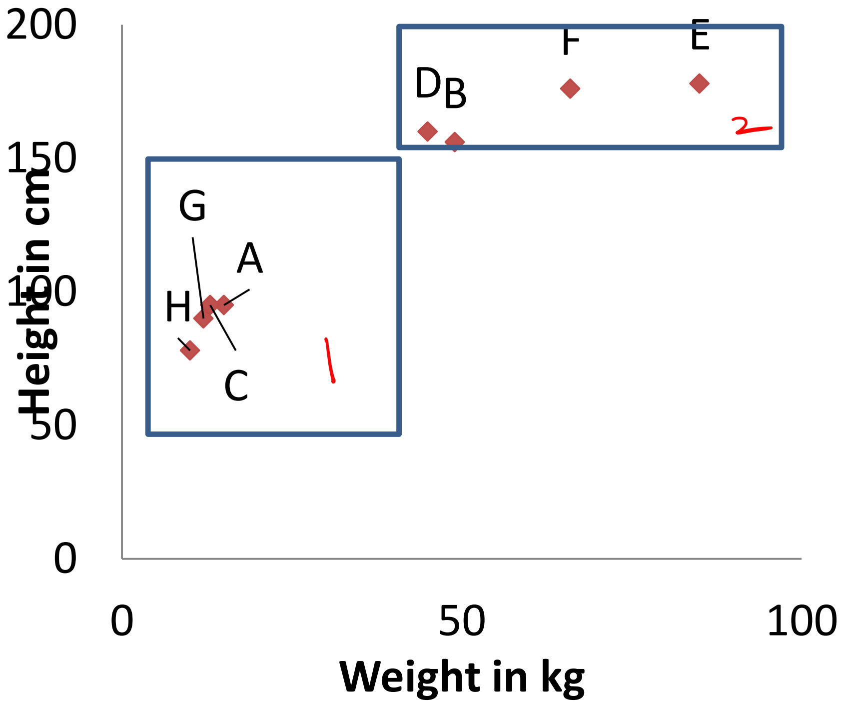
Interval-Scaled Variables

- For example :
- Take eight people, the weight (in kilograms) and the height (in centimetres
- In this situation, n = 8 and p = 2.

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

Table :1

Figure 1



Interval-Scaled Variables

- The units on the vertical axis are drawn to the same size as those on the horizontal axis, even though they represent different physical concepts
- The plot contains two obvious clusters, which can in this case be interpreted easily: the one consists of small children and the other of adults
- However, other variables might have led to completely different clustering
- For instance, measuring the concentration of certain natural hormones might have yielded a clear cut partition into different male and female persons

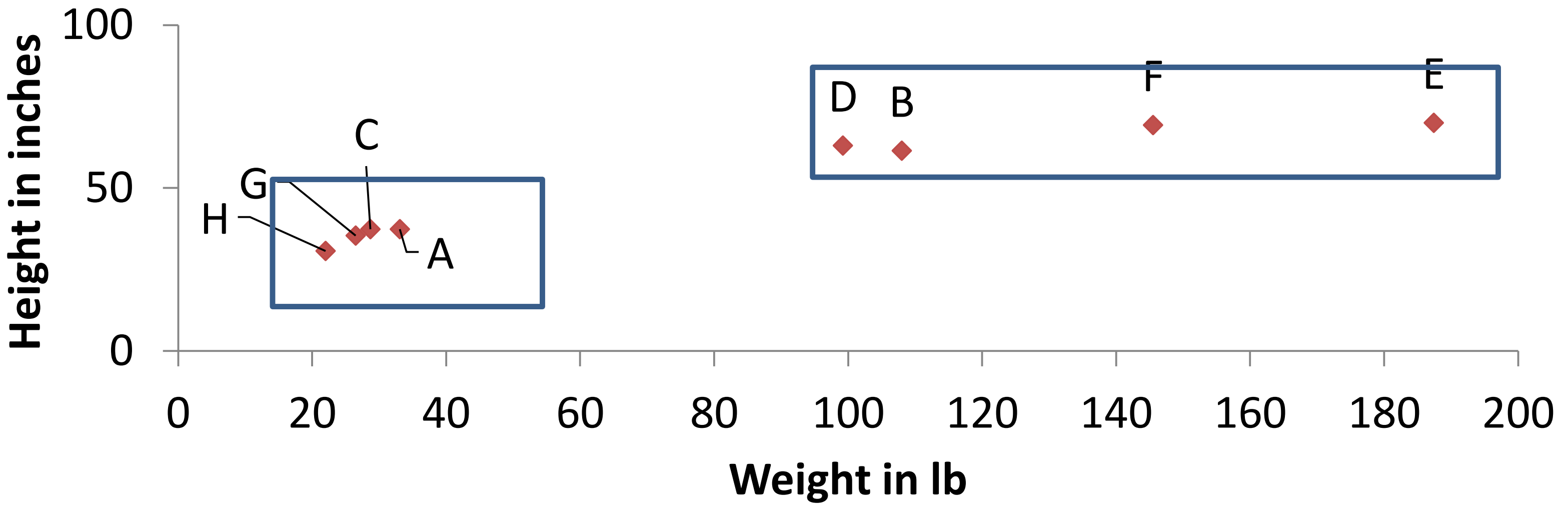
Interval-Scaled Variables

- Let us now consider the effect of changing measurement units.
- If weight and height of the subjects had been expressed in pounds and inches, the results would have looked quite different.
- A pound equals 0.4536 kg and an inch is 2.54 cm
- Therefore, Table 2 contains larger numbers in the column of weights and smaller numbers in the column of heights. Figure 2

Person	Weight(lb)	Height(in)
A	33.1	37.4
B	108	61.4
C	28.7	37.4
D	99.2	63
E	187.4	70
F	145.5	69.3
G	26.5	35.4
H	22	30.7

Table :2

Figure 2


100
50
0
Height in inches
0
20
40
60
80
100
120
140
160
180
200
Weight in lb
H
G
C
A
D
B
F
E

Interpretation

- Although plotting essentially the same data as Figure 1, Figure 2 looks much flatter
- In this figure, the relative importance of the variable “weight” is much larger than in Figure 1
- As a consequence, the two clusters are not as nicely separated as in Figure 1 because in this particular example the height of a person gives a better indication of adulthood than his or her weight. If height had been expressed in feet (1 ft = 30.48 cm), the plot would become flatter still and the variable “weight” would be rather dominant
- In some applications, changing the measurement units may even lead one to see a very different clustering structure

Standardizing the data

- To avoid this dependence on the choice of measurement units, one has the option of standardizing the data
- This converts the original measurements to unitless variables
- First one calculates the m ven by:

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$$

for each f = 1,. . . , p

Standardizing the data

- Then one computes a measure of the dispersion or "spread" of this fth variable
- Genera

$$\text{std}_f = \sqrt{\frac{1}{n-1}\left\{(x_{1f} - m_f)^2 + (x_{2f} - m_f)^2 + \cdots + (x_{nf} - m_f)^2\right\}}$$

Standardizing the data

- However, this measure is affected very much by the presence of outlying values
- For instance, suppose that one of the x_{if} has been wrongly recorded, so that it is much too large
- MAD
  In this case std_f will ~~be unduly inflated~~, because x_{if} - m_f is squared
- Hartigan (1975, p. 299) notes that one needs a dispersion measure that is not too sensitive to outliers
- Therefore, we ne contribution of bsolute value $|x_{if} - m_f|$

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

Standardizing the data

- Let us assume that s_f is nonzero (otherwise variable f is constant over all objects and must be removed)
- Then the standardized measurements are defined by and sometimes called z-scores
- They are unitless because both the numerator and the denominator are expressed in the same uni

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- By construction, the z_{if} hav and their mean absolute deviation is equal to1

(0, 1)

Standardizing the data

- When applying standardization, one forgets about the original data and uses the new data matrix in all subsequent computations

$$
\begin{array}{cc}
 & \text{variables} \\
\text{objects} & \begin{bmatrix} z_{11} & \cdots & z_{1f} & \cdots & z_{1p} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{if} & \cdots & z_{ip} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nf} & \cdots & z_{np} \end{bmatrix}
\end{array}
$$

Detecting outlier

- The advantage of using s_f rather than std_f, in the denominator of z-score formula is that s_f will not be blown up so much in the case of an outlying x_{lf}, and hence the corresponding z_{if} will still be noticeable so the i^{th} object can be recognized as an outlier by the clustering algorithm, which will typically put it in a separate cluster

Standardizing the data

- The preceding description might convey the impression that standardization would be beneficial in all situations.
- However, it is merely an option that may or may not be useful in a given application
- Sometimes the variables have an absolute meaning, and should not be standardized
- For instance, it may happen that several variables are expressed in the same units, so they should not be divided by different s_f
- Often standardization dampens a clustering structure by reducing the large effects because the variables with a big contribution are divided by a large s_f

Thank you