Application of Computer in Chemistry SSC 3533

PATTERN RECOGNITION

Prof. Mohamed Noor Hasan
Dr. Hasmerya Maarof
Department of Chemistry

Outline

- Concept of Classification
- Overview of Pattern Recognition Methods
- Discriminant Analysis
- K-nearest Neighbour (KNN)
- Principal Component Analysis (PCA)
- Hierarchical Clustering

Introduction

- We often encounter problems in which some sort of classification has to be made in order to answer questions about the samples:
- IR, NMR or MS spectra – What kind of compounds?
- Blood samples from a patient – What kind of disease?
- Petrol fuel samples – Are they being mixed with lower grade products?
- Samples of polluted air – What is the source of pollution?
- Samples of meteorite – Are they from our solar system?

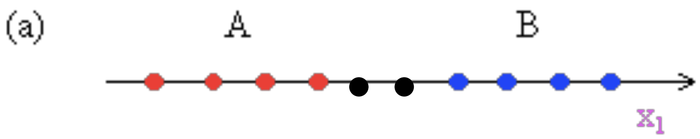
Classification Task

- Human being are used to characterize or classify objects into certain classes or groups:
  - Just by looking at it we are able to differentiate letter 'a' from letter 'b'.
  - By listening to someone's voice over the phone we can tell whether it is someone that we know?
- Although the task is quite trivial to us, the process might involve complicated steps:
  - For example, to identify a letter, we have to know which direction is the curve, which side has a straight line.

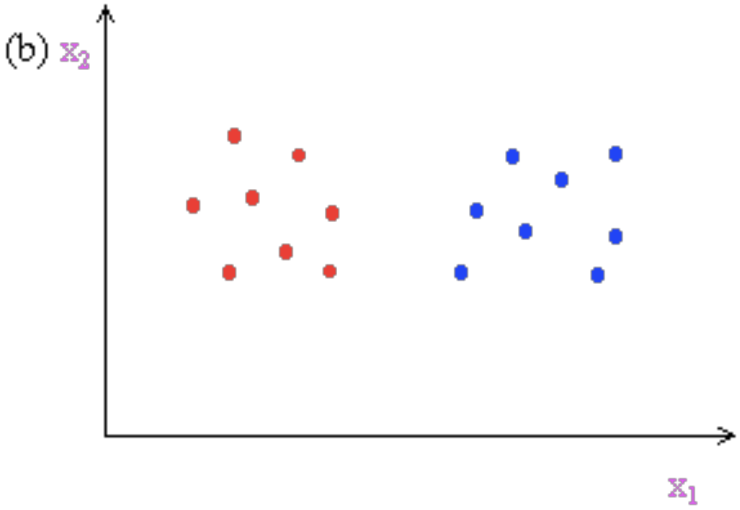
Classification with a computer

- The classification process can be ‘taught’ to the computer
- Today, scanned texts can be read by a computer and converted into word documents (Optical character reader software)
- In chemistry, samples being analyzed can produce a lot of data that must be processed and to identify patterns in the data.
- Unfortunately, the process can no longer be done using naked eyes alone, due to the multidimensional nature of the data.

Simple Classification

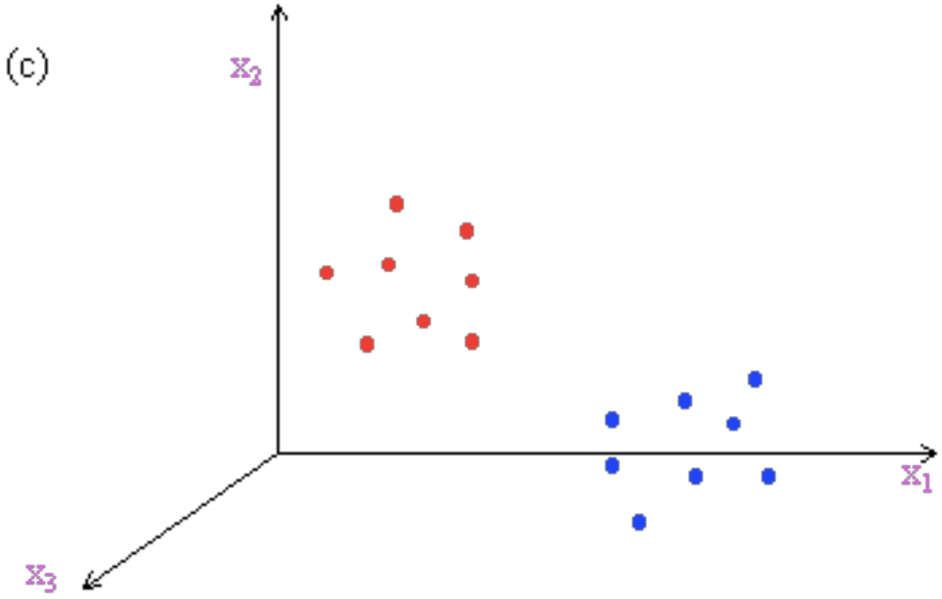


Using 1 variable



Using 2 variables

Multivariate classification



Using 3 variables

Pattern

- In differentiating letter 'a' from letter 'b', our eyes and mind use certain criteria or features to differentiate them
- The list of features form a pattern
- In chemistry, results of analyses or tests on the samples can form the features
- Mathematically, the features can be expressed in the form of a vector.
- $x_i = (x_1, x_2, x_3.........x_n)$

Class, objects and features

- A sample is an object – e.g. an air pollution sample, a water sample, an IR spectrum of food sample.
- A sample can be represented by features – e.g. air pollution: conc. of compounds present in the sample, absorbance at certain wavelength
- Based on similarity of features, samples can be grouped into classes – e.g. polluted vs. unpolluted, toxic vs. non-toxic

Concept of distance

- An object can be represented by a point in ***n*** dimensions, where *n* is number of features.
- Hopefully objects from the same class will cluster (group together) in one area.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Distance between two points i and j

$x_i = (x_{i1}, x_{i2}, \ldots\ldots\ldots x_{in})$
$x_j = (x_{j1}, x_{j2}, \ldots\ldots\ldots x_{jn})$

Degree of Similarity

- Degree of similarity between two objects:

  $s_{ij} = 1 - d_{ij}/d_{max}$

  s_{ij} – similarity between object *i* and *j*
  d_{ij} – distance between *i* and *j*
  d_{max} – maximum distance
- Shorter distance means higher similarity

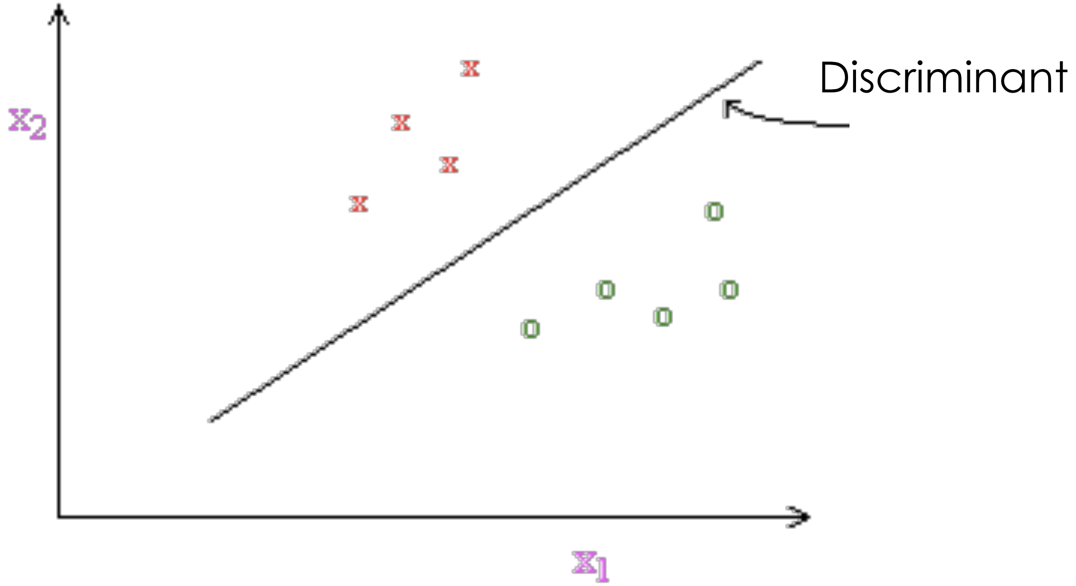
Pattern Recognition Methods

Two types of pattern recognition methods:

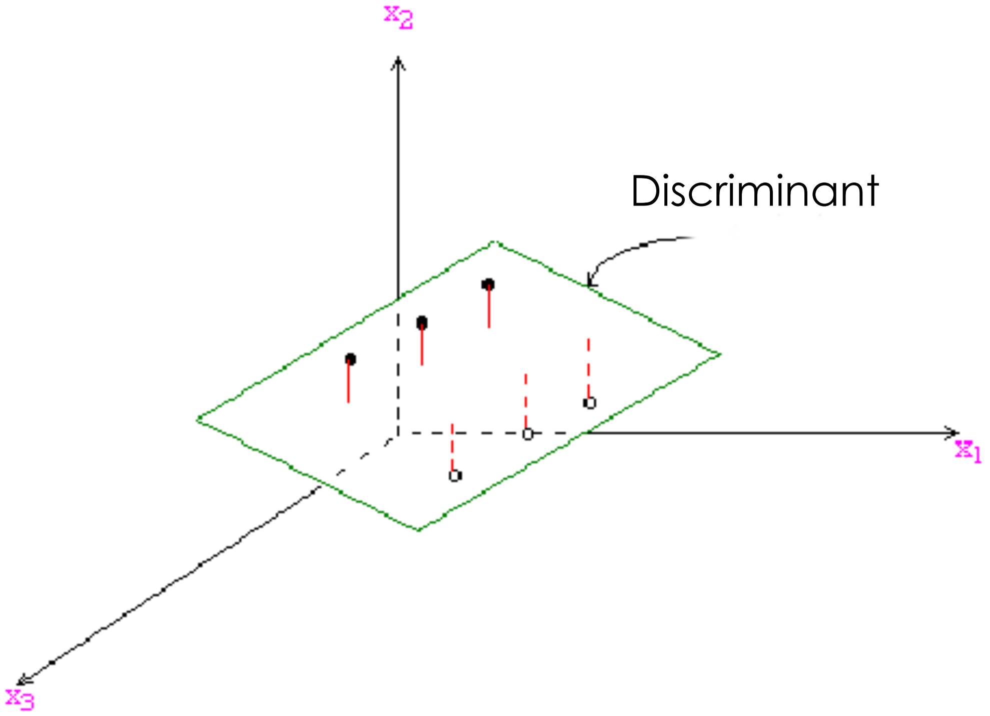
- Supervised
- Unsupervised
- In supervised methods, training set (consists of object with known classes) is used to build a discriminant (classifier) or model. The discriminant is then used to classify the unknown.
- In unsupervised methods, class of samples are not known in advance. We have to use data available to us to form clusters (groups) in the data – hese methods are also called exploratory methods.

Classification using a discriminant

- Main objective is to divide the points into classes.
- One way of doing this is by building a separator (discriminant) through the data space



Discriminant in three dimensional space



Discriminant Function

- The discriminant function f(x) divides data space into areas with certain features (class)
- For all objects *x* in class *k*, exist a function $f_k(x)$ such that:

  $f_k(x) > f_l(x)$ for all k, l

  The surface that separate *k* and *l*

  $f_k(x) - f_l(x) = 0$

  For two classes: $f(x) = f_1(x) - f_2(x)$

  if f(x) positive – class A;

  if f(x) negative – class B

Learning Machine

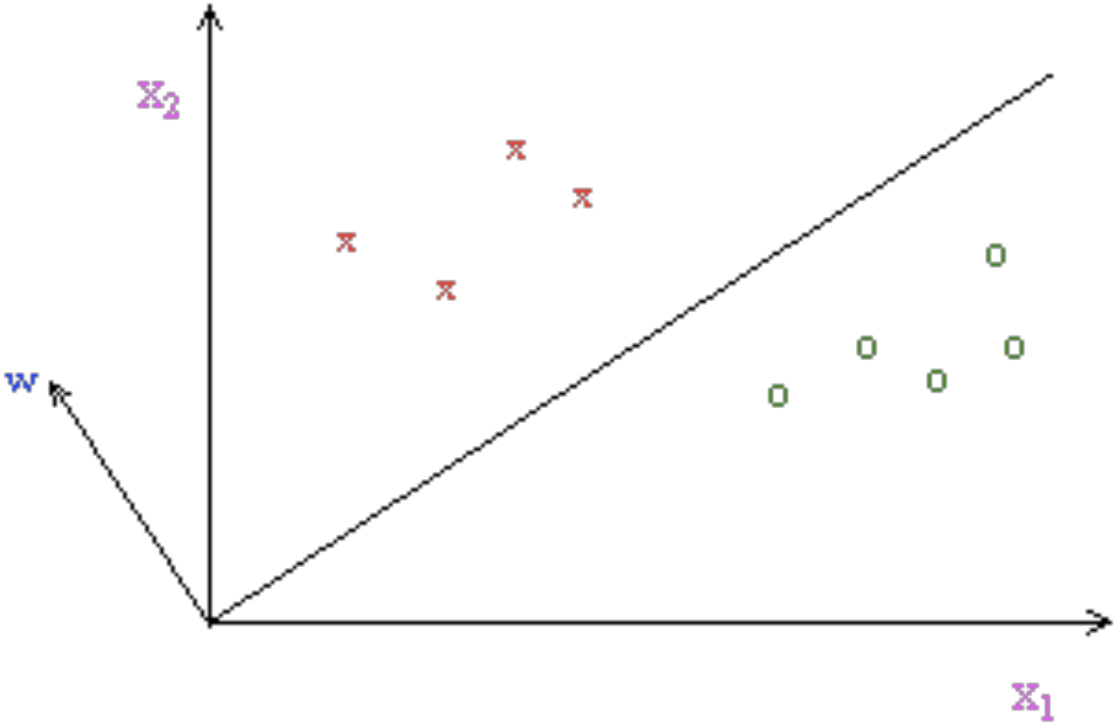
- A discriminant function can be regarded as a surface that divides the data in *n*-dimensional space.
- If the surface is a plane, it can be represented by a vector perpendicular to the discriminant surface.
- The vector is called weighting vector.

W.X $= [W]\ [X] \cos \theta$

W.X $= w_1\, x_1 + w_2\, x_2 + \ldots\ldots\ldots\ w_d\, x_d + w_d{+}1$

- If the product of the two vectors is positive, the sample is in class A; if the product is negative, class B.

Weight vector



Finding the discriminant

- Take one discriminant
- Classify members of training set, one by one.
- If a discriminant function is able to classify an object, let it be.
- When wrong classification occurs, correct the discriminant function until correct classification is obtained.
- The learning machine continues to classify all members of the training set.

Correction of discriminant

$WX = S$

When wrong classification occurs:

$WX_i = S$ S – wrong sign

Find W' such that $W'X_i = S'$ S' opposite sign

W' calculated from previous W:

$W' = W + C\,X_i$

$S' = W'X_i = (W + C\,X_i)\,X_i$

$S' = W\,X_i + C\,X_i\,X_i$

$C = (S'-S)/(X_i\,X_i)$

Choose $S' = -S$

$C = -2S\,/(X_i\,X_i)$

W' can be calculated.

Finding the discriminant (Cont.)

- The process is called ‘learning’ because the discriminant corrects its performance during classification.
- The method will find a solution if it exists. Easy to implement for simple calculations

K- Nearest Neighbour (KNN)

- A very popular classification method because the concept is very simple and easy to implement.
- An object is classified according to the class of majority of its nearest neighbours.
- Nearest neighbours are determined by computing Euclidean distances.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

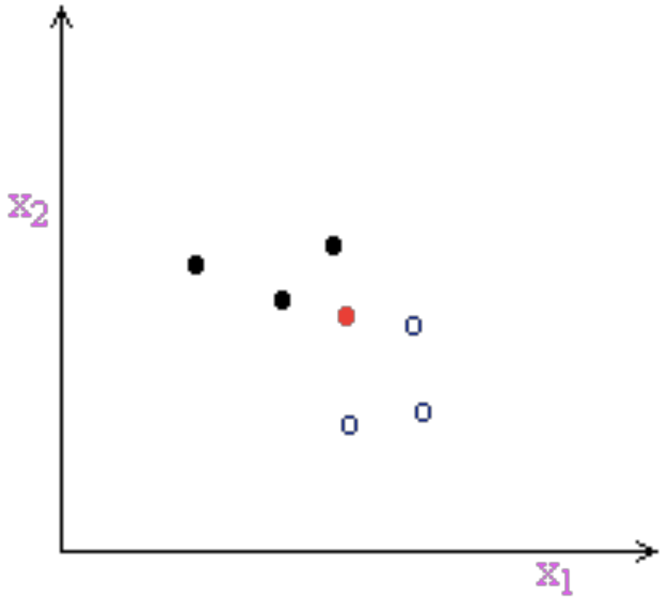
i and j – objects

k - feature

KNN Method

- Determine distance between an object and other objects with known classes.
- Arrange the distances. Choose K nearest neighbours, with K is an odd number.
- Estimate number of neighbours in class *1, 2, ...q*.
- Assign an object into class q_k if majority of its neighbours are in class q_k.

Example



In this example, the object is assigned into class 1 because 2/3 of its neghbours are from class 1.

The method is conceptually simple but requires a lot of calculations because for each object that needs to be classified all the distances have to be calculated.

The choice of *K* is also crucial because it can also influence the outcome.

KNN Exercise

- Calculate the distance between the unknown X (5.5, 5) to all objects
- Determine 3 and 5 nearest neighbors of object X
- Determine the class of object X

KNN Data

Class	X_1	X_2
A	5.77	8.86
A	10.54	5.21
A	7.16	4.89
A	10.53	5.05
A	8.96	3.23
B	3.11	6.04
B	4.22	6.89
B	6.33	8.99
B	4.36	3.88
B	3.54	8.28
unknown	***5.5***	***5***

Advantages of KNN

- KNN is a very simple approach and can be easily understood and programmed
- KNN makes very few assumptions about the data, eg normality of noise distribution.

Limitations of KNN

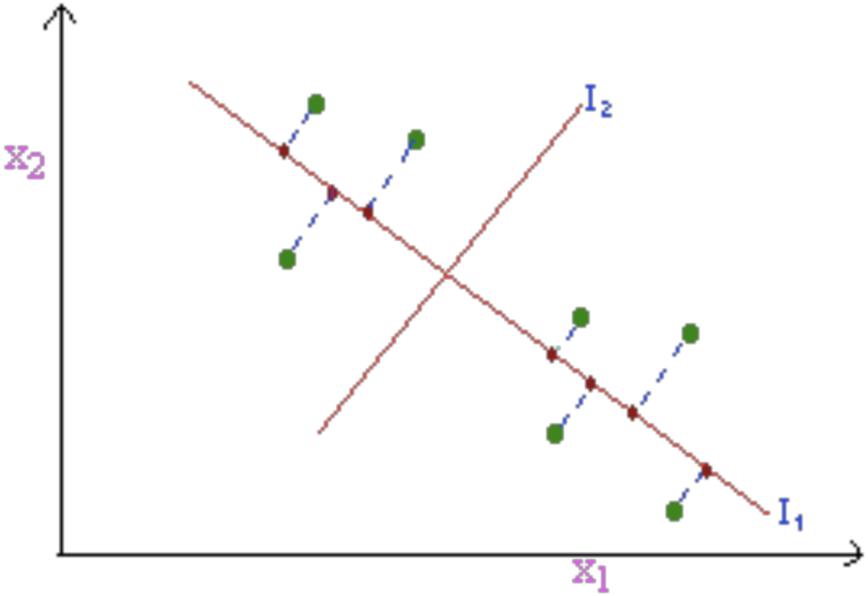
- Number of samples in each class of the training set should be approximately equal, otherwise the 'votes' will be biased towards the larger class
- Each variable assumes equal significance. In reality some of the variables are highly correlated with each other and some do not contribute to the classification. Must perform variable selection or use different distance measure

Limitations of KNN (Cont.)

- Ambiguous or outlying samples in the training set can cause problems in the classification
- The method do not take into account the spread or variance in each class

Principal Component Analysis (PCA)

- Principal component analysis extracts information from *n* dimensions into lower dimensions.
- For example, data from *n* dimensions projected into 2 dimensions.
- New variables called principal components resulted.



Principal Components

The new variable is a linear combination of the original variables:

$u_1 = ax_1 + bx_2 + cx_3$ ....

Original variables:

x_1	x_2	x_3	x_4	x_n
--	-	-	-	-
--	-	-	-	-
--	-	-	-	-
--	-	-	-	-

New variables:

u_1	u_2	u_3	u_4	u_n
-	-	-	-	-
--	-	-		-
--	-	-	-	-
--	-	-	-	-

Principal Components (cont.)

The new variable is obtained by multiplying original variables with a coefficient.

$u_1 = v_{11}x_1 + v_{12}x_2 + v_{13}x_3 + \ldots v_{1n}x_n$

$u_2 = v_{21}x_1 + v_{22}x_2 + v_{23}x_3 + \ldots v_{2n}x_n$

.

$u_n = v_{n1}x_1 + v_{n2}x_2 + v_{n3}x_3 + \ldots v_{nn}x_n$

This can be written in matrix form as:

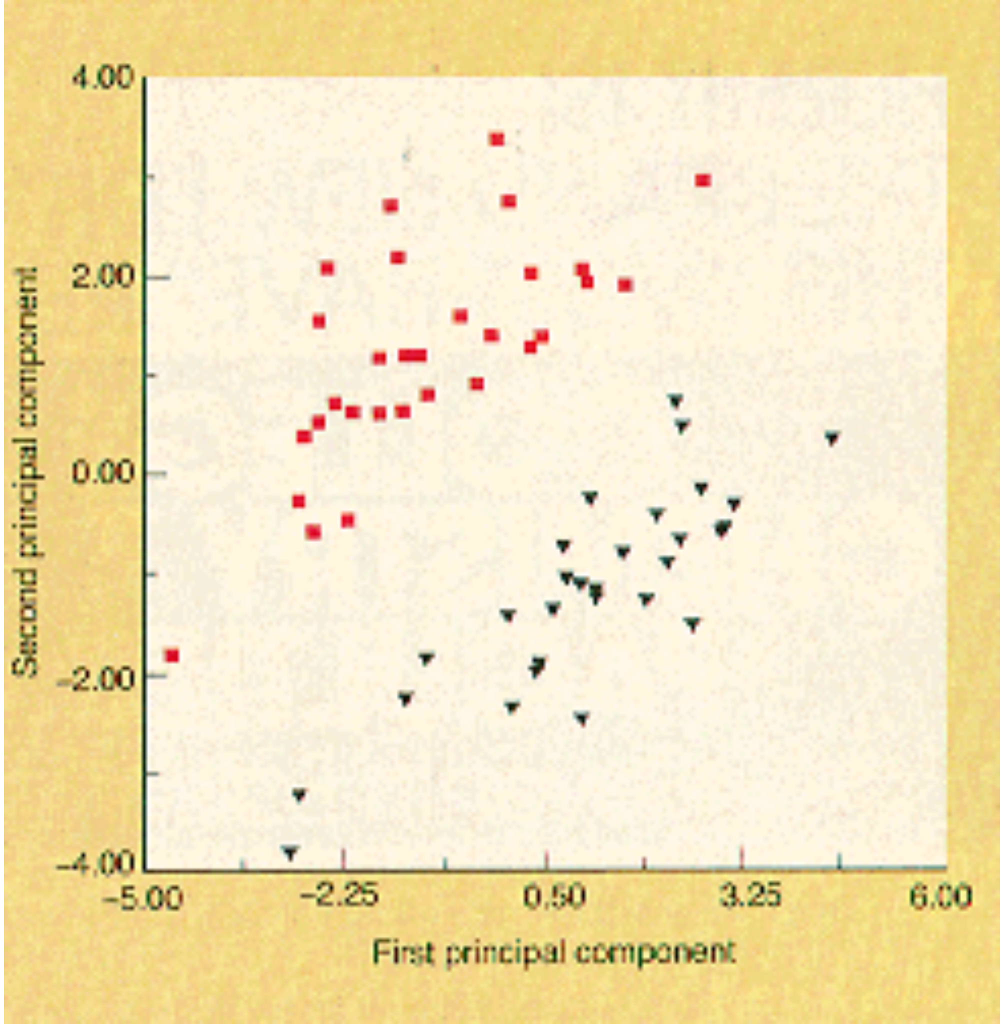
u = V.x

Matrix **V** is called coefficient matrix or loading

European Bee or Africanized Bee?

- A study to differentiate two species of bees, European or Africanized
- The Africanized bees are more aggressive. It is difficult to differentiate the two species just by looking at physical features.
- Hydrocarbon in the bees cuticle were extracted and their concentration were determined by GC-MS.
- From the chromatogram, 10 peaks were selected to represent the features of the bees samples.
- B.K. Lavine, *Anal. Chem*. **59**, 1987, 468A.

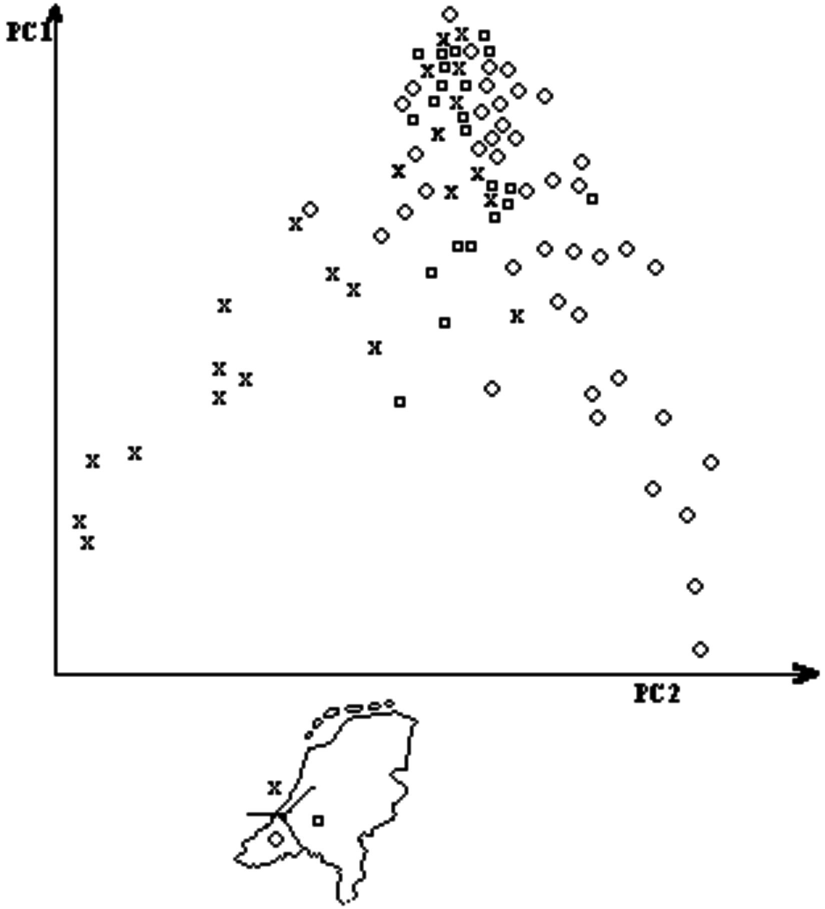
Principal Components Plot



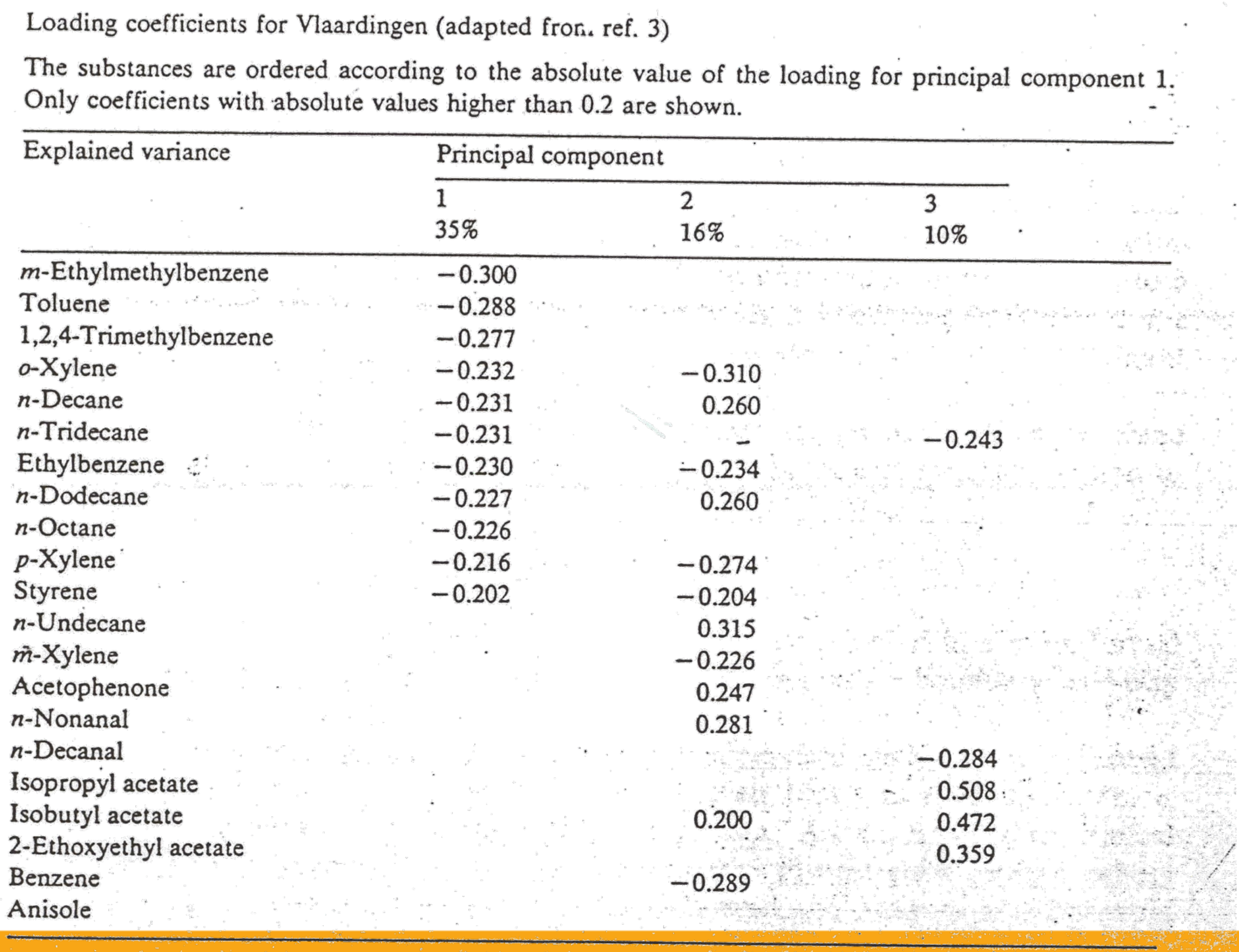
Air Pollution Study

- Samples were taken every week for 3 years
- Concentration of 300 compounds were determined. 35 compounds were found in all samples
- Total number of samples 150.
- Weather data were also recorded
- Massart *et al. Atmos. Environ.* **18** (1984) 2741.

Principal component plot



Loadings for PC1, PC2, PC3



Loading coefficients for Vlaardingen (adapted fron. ref. 3)

The substances are ordered according to the absolute value of the loading for principal component 1. Only coefficients with absolute values higher than 0.2 are shown.

Explained variance	Principal component		
	1	2	3
	35%	16%	10%
m-Ethylmethylbenzene	−0.300		
Toluene	−0.288		
1,2,4-Trimethylbenzene	−0.277		
o-Xylene	−0.232	−0.310	
n-Decane	−0.231	0.260	
n-Tridecane	−0.231		−0.243
Ethylbenzene	−0.230	−0.234	
n-Dodecane	−0.227	0.260	
n-Octane	−0.226		
p-Xylene	−0.216	−0.274	
Styrene	−0.202	−0.204	
n-Undecane		0.315	
m-Xylene		−0.226	
Acetophenone		0.247	
n-Nonanal		0.281	
n-Decanal			−0.284
Isopropyl acetate			0.508
Isobutyl acetate		0.200	0.472
2-Ethoxyethyl acetate			0.359
Benzene		−0.289	
Anisole			

Conclusions

- PC1 – a measure of general pollution
  PC2 - aromatic – negative; aliphatic - positive
- 3 conditions:
  - Normal condition
  - Pollution when wind comes from industrial area - aromatic
  - Pollution when wind comes from highways and industrial areas - aliphatic

Other Examples

- **Determination of Chemical Structures**
- Chemical structures can be determined based on MS, IR, NMR spectral information. Can determine functional groups and substructures present in the structures

Classification of Complex Mixtures

- Mixtures can be classified into different categories using pattern recognition techniques.
- Olive oils – determine origin from percentage of fatty acids – will determine price.

More examples

- **Oil Pollution Study**
- From the types of hydrocarbons in the oil spills, determine the source.

Predicting biological properties from molecular structures

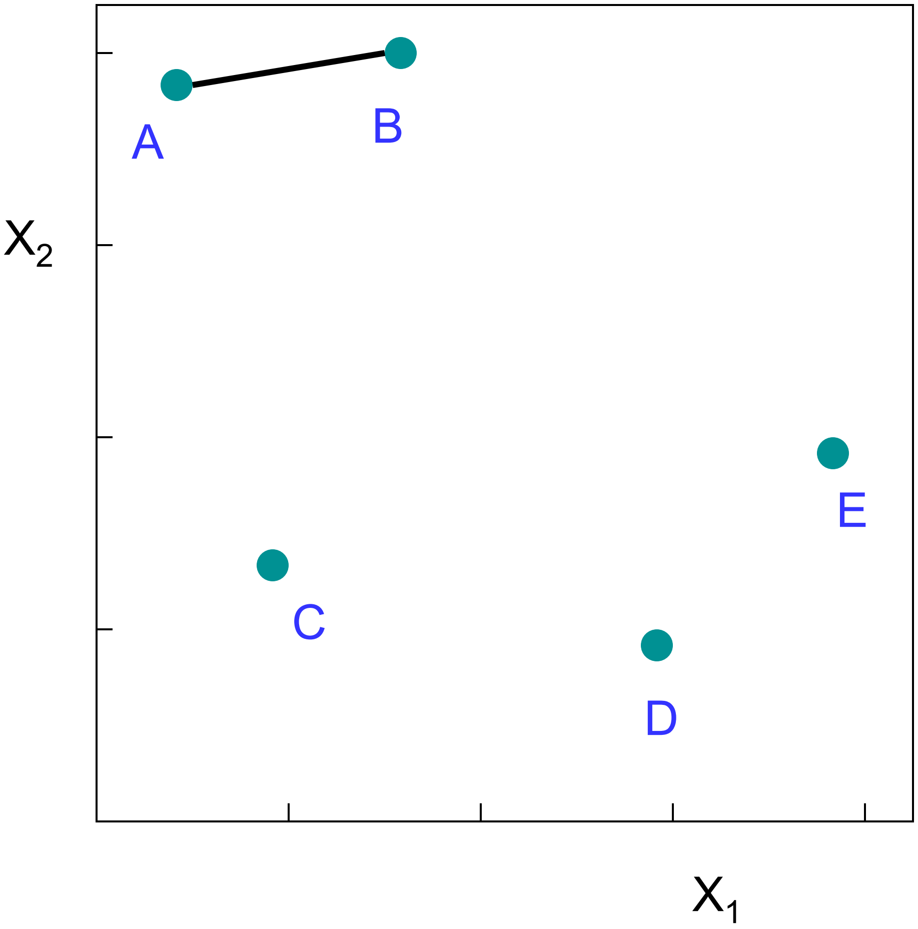
- Structure-activity relationships study (SAR).
- Biological activity: toxicity, carcinogen, medicinal property, aroma
- Generate descriptors from molecular structures
- Based on training set, develop classifier. Use classifier to predict activity of unknown.

Cluster Analysis

- Unsupervised technique
- Use distance between two samples (example Euclidean distance) to determine degree of similarity
- Shorter distance means higher similarity
- The cluster produced represented in a two dimensional plot - dendogram

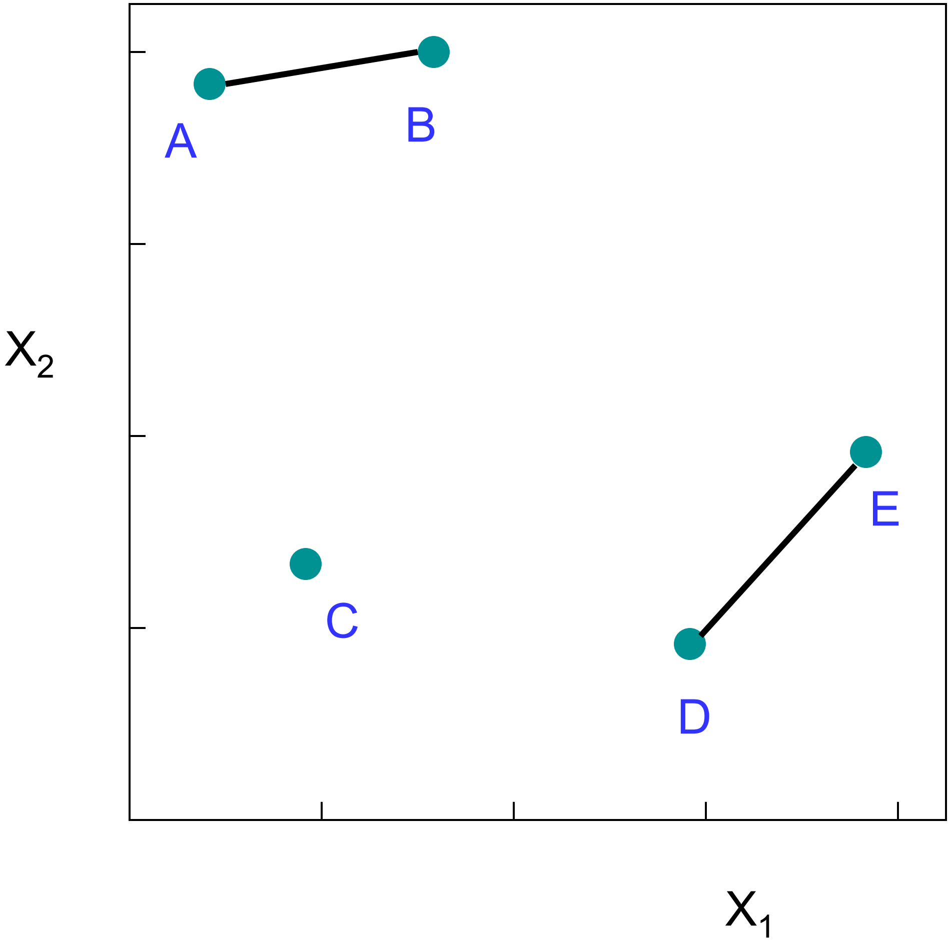
Step 1

Sample A and B are the closest, form the first cluster



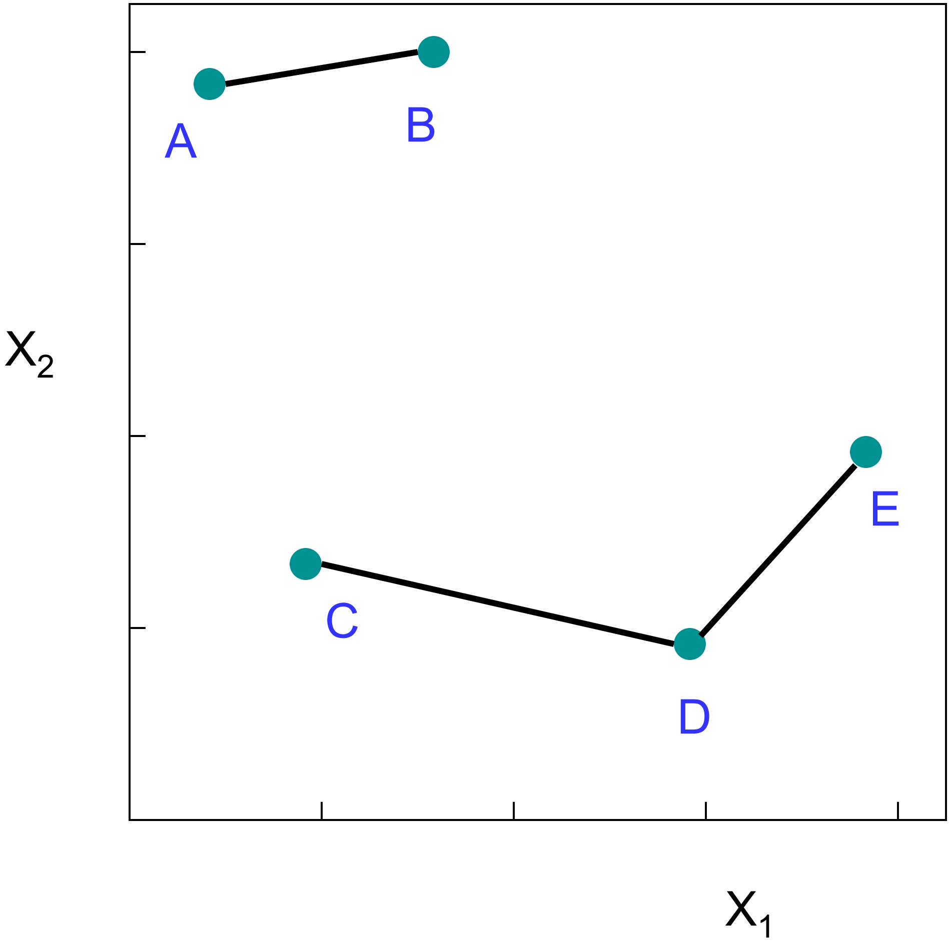
Step 2

Sample D and E are the closest. Form the second cluster.



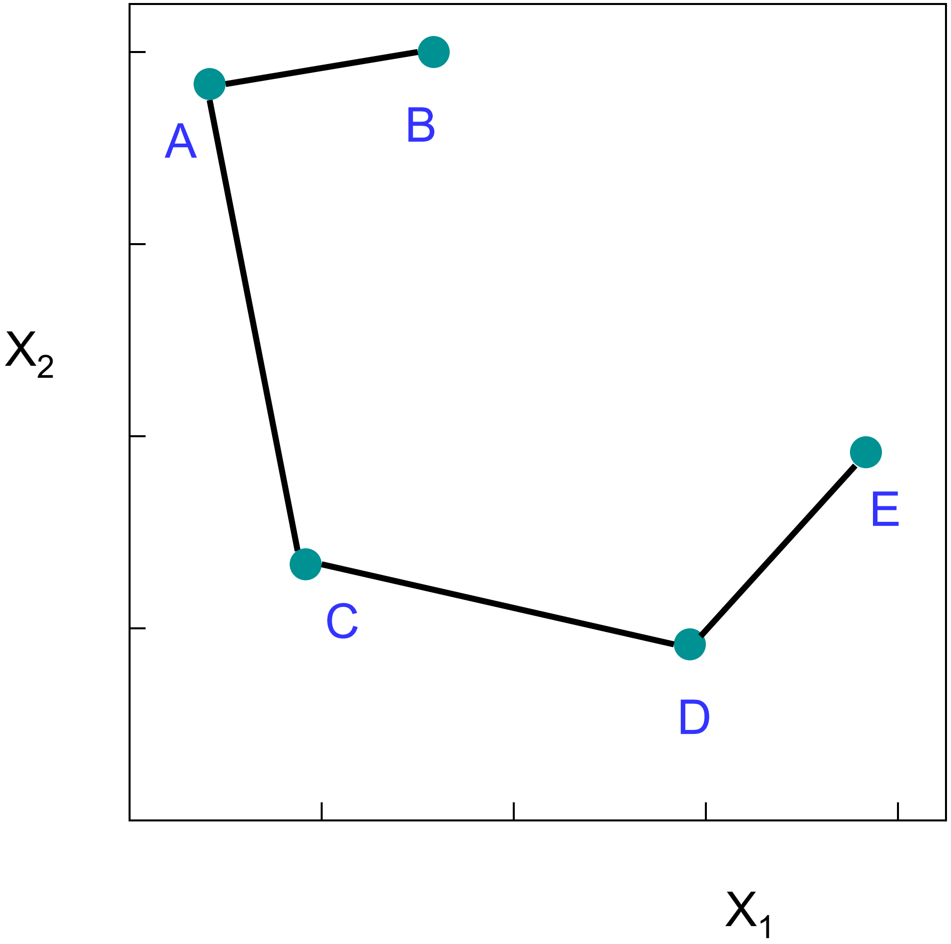
Step 3

Sample C is linked to cluster DE because D is closest to C

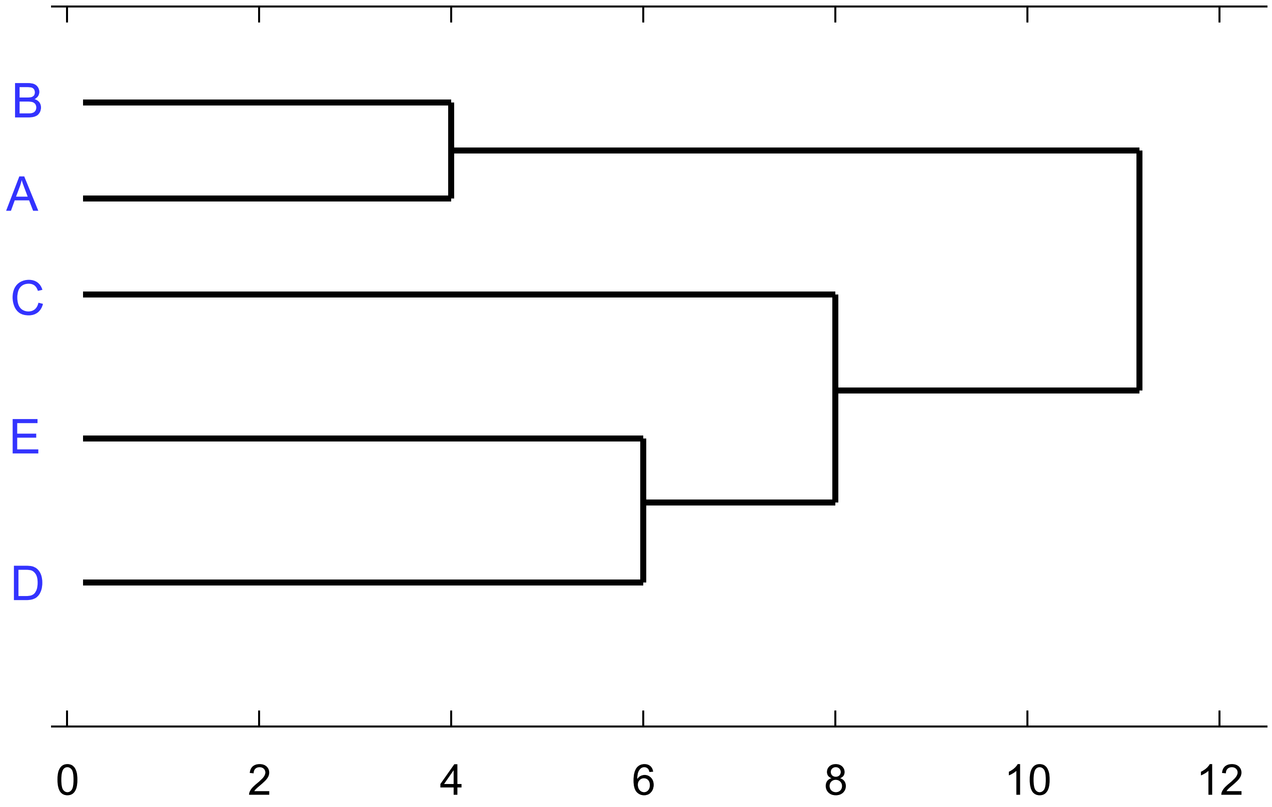


Step 4

The two clusters are joined together to form a big cluster.



Dendogram – Single link



Centroid link method

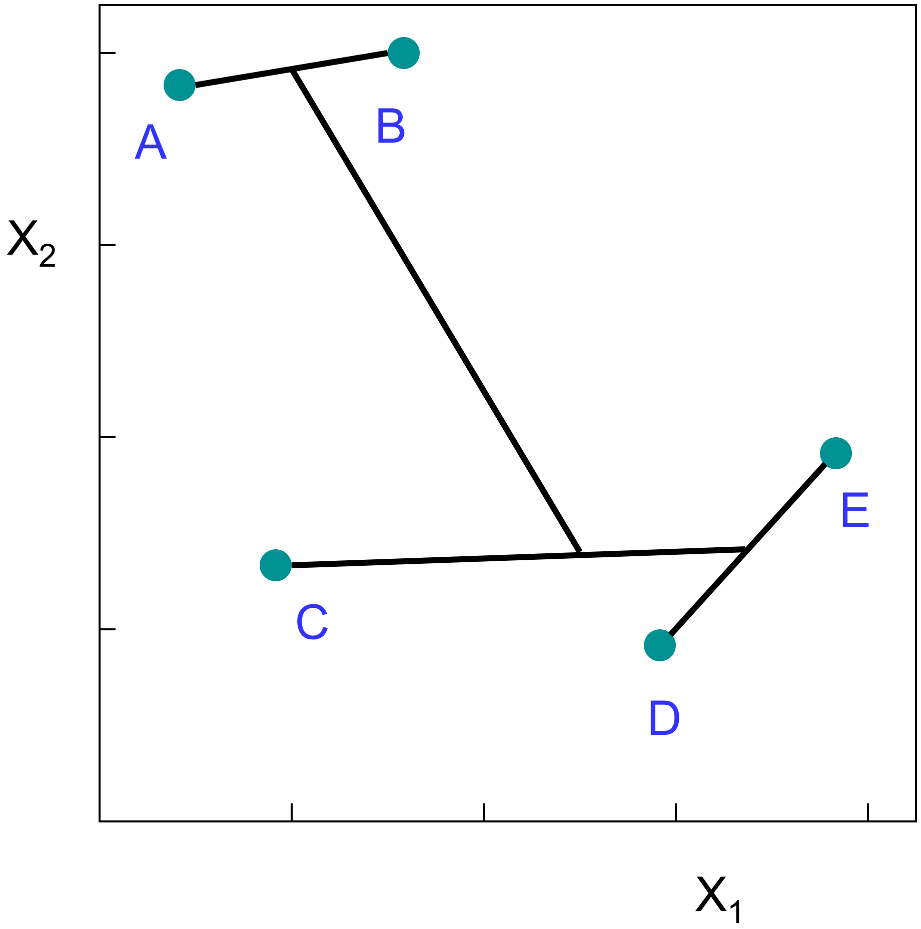
- Connecting the center of a cluster and not nearest neighbours.
- The centroid of a cluster is the average point in the cluster.

Centroid link method

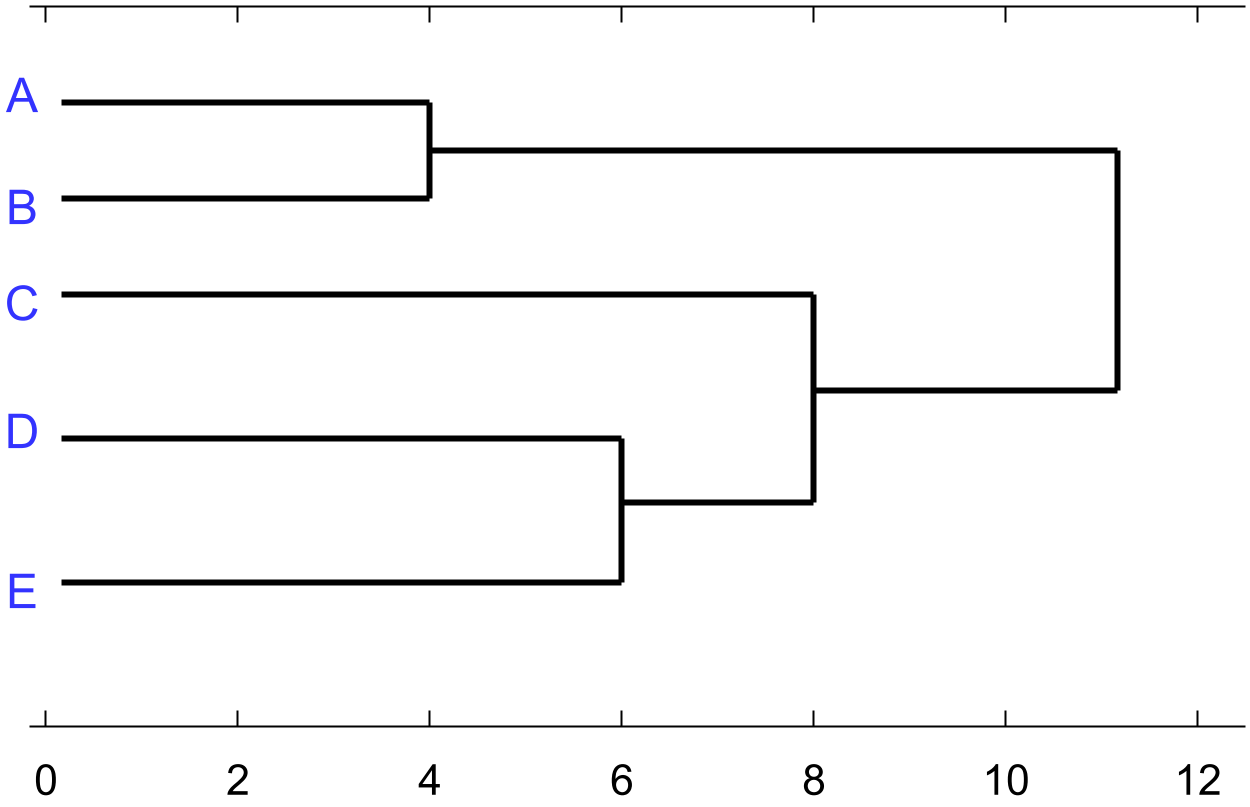
The first and second steps are similar to the single link method

In the third step, C is connected to the centroid of cluster DE

In the final step, centroid of AB is connected to the centroid of CDE



Dendogram – centroid link


A
B
C
D
E
0
2
4
6
8
10
12