Cluster analysis

Classification and prediction

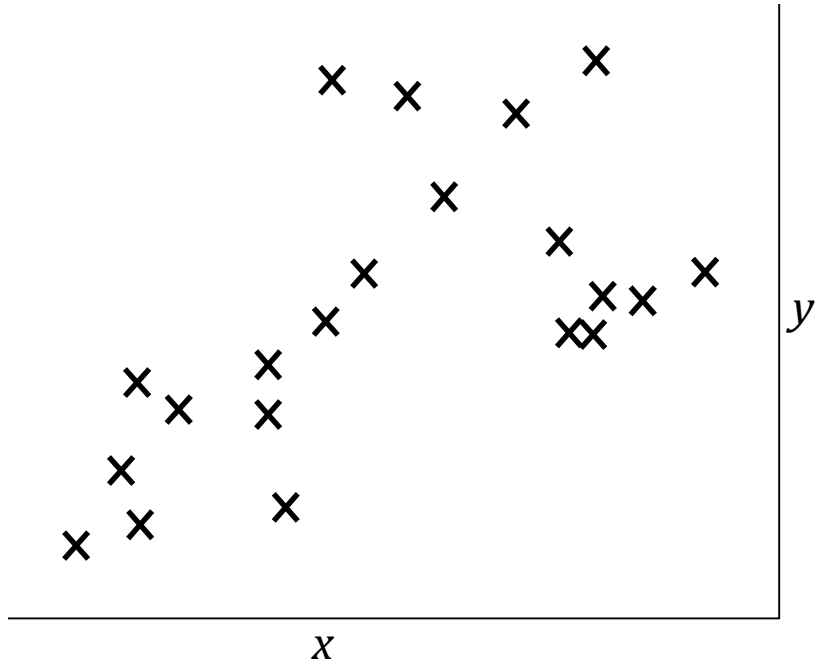
Methods

- *k*-means
- hierarchical
  - nearest neighbour
  - divisive
- Übung

- Many methods – no perfect answers

Classification versus prediction ?

Easy data two clusters

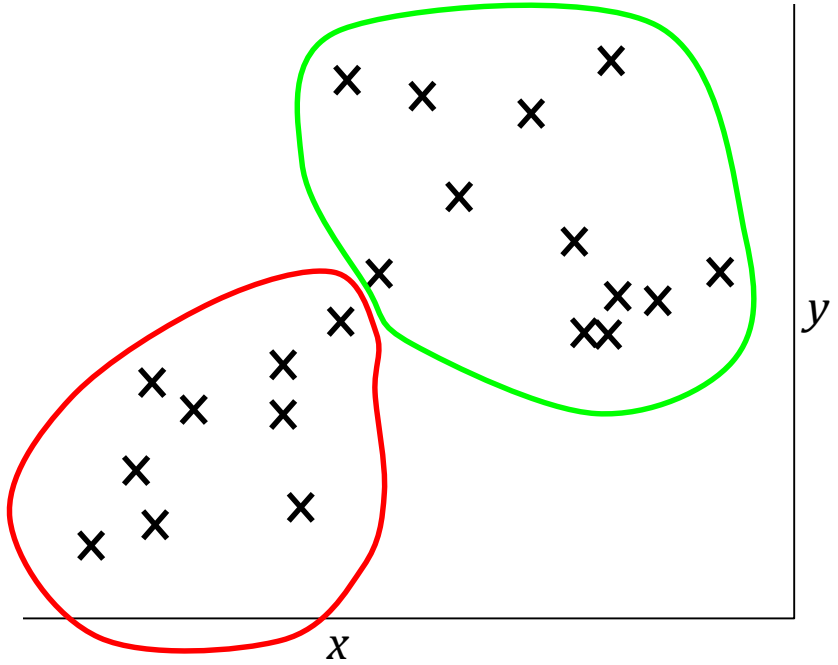


Classification versus prediction ?

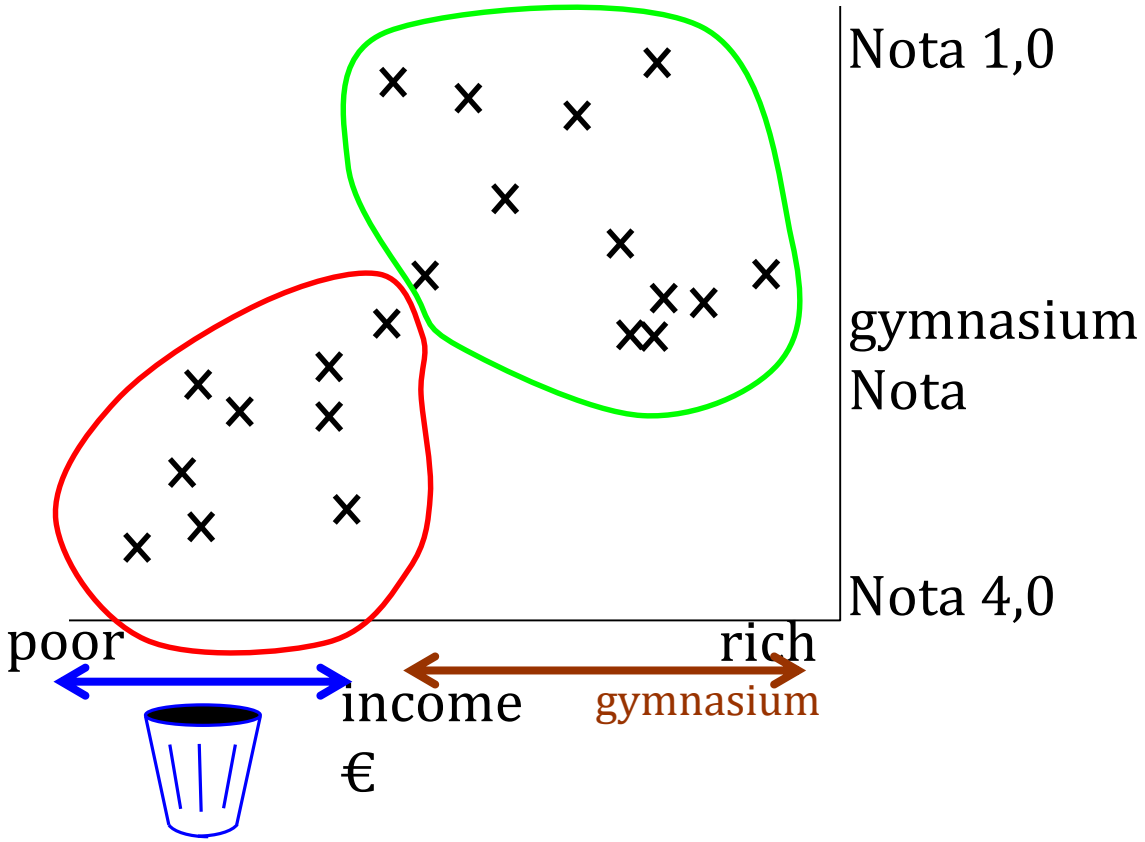
Easy data two clusters

Can this be predictive ?

- put labels on



Classification versus prediction ?



Try a prediction based on income

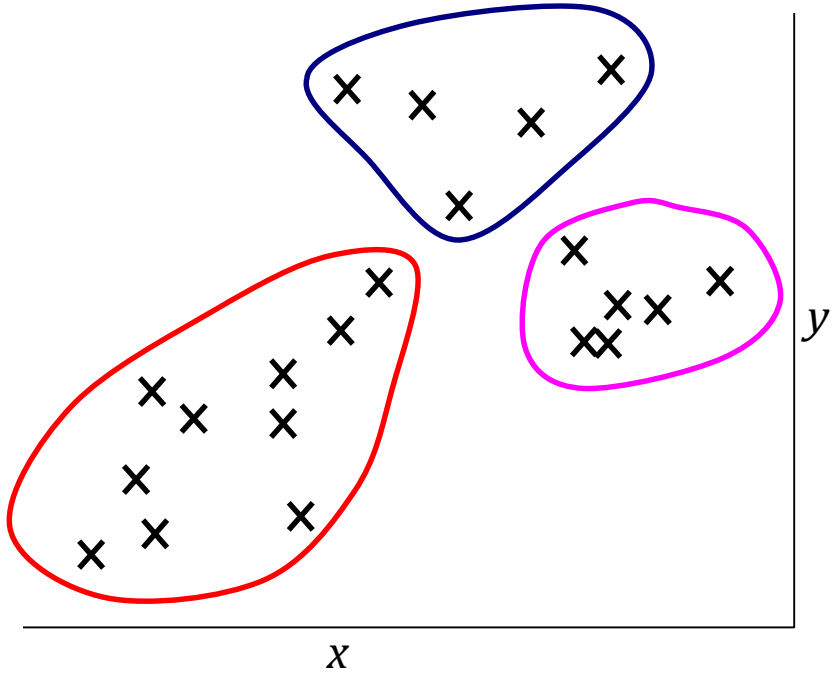
- the main goal
  - if we know of some properties, we can guess others

Problems

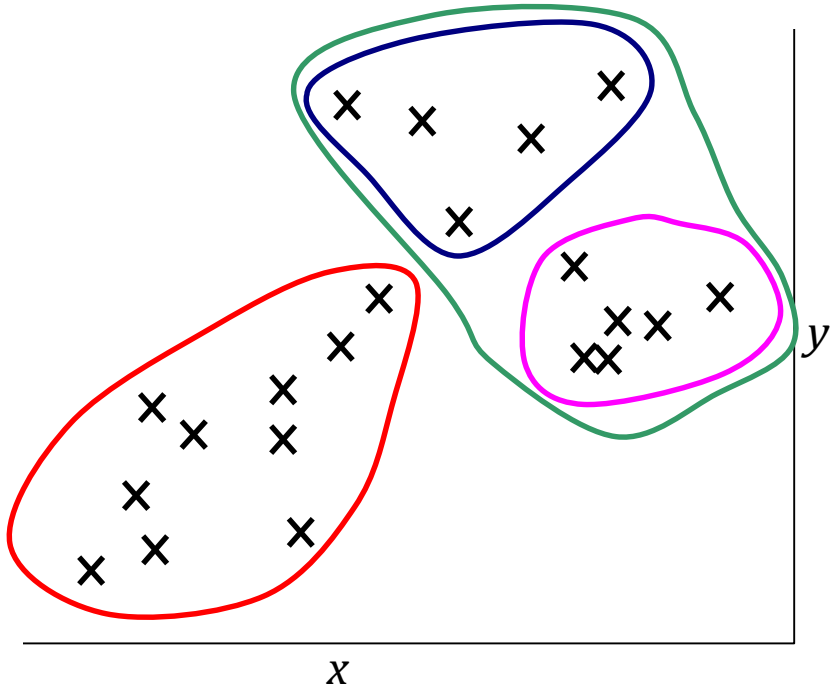
Easy data two clusters

- is it really ?

Alternative ?

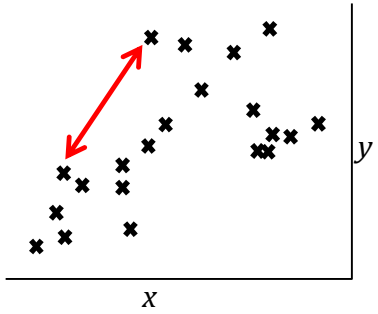


Problems



Two clusters with sub-clusters ?

Distance Measures (Euclidean)



For any two points

- want a distance /dissimilarity

Euclidean distance (easy in two dimensions)

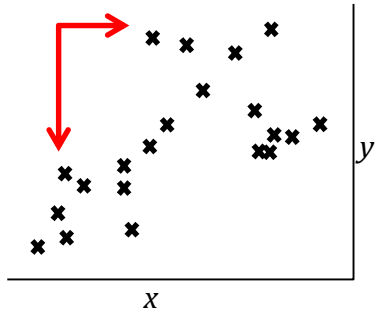
$$d_{ij} = \left(\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 \right)^{1/2}$$

in 3D $$d_{ij} = \left(\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2 \right)^{1/2}$$

in nD $$d_{ij} = \left(\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2 + .. \right)^{1/2}$$

Distance Measures (Manhattan)

- 2D $d_{ij} = |x_i - x_j| + |y_i - y_j|$
- nD $d_{ij} = |x_i - x_j| + |y_i - y_j| + |z_i - z_j| + \cdots$



Euclidean versus Manhattan versus …

- depends on belief
- if one is lucky, results will not be too different

Worse cases

- category data
  - cars have
    - speeds, size, **colour, 2 door/4door**

not x, y continuous descriptor

- a possible Manhattan measure

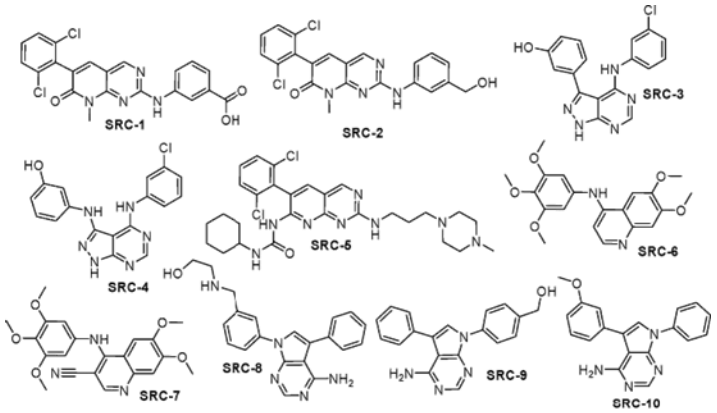
A set of discrete descriptors

Identify properties

- make long bit-vector
- dissimilarity ?
  - count matching bits
- typically 10^2 - 10^3 properties
- crude ?
  - enough properties that mistakes do not matter

Is this a Manhattan distance ?

- probably



$n_{polar} < 5$	0
$5 < n_{polar} < 10$	0
$n_{polar} \geq 10$	1
contains sulfur	0
acidic (can lose H^+)	0
has ether O	1
… lots more ...	
mol. wt. < 300	0
300 < mol. wt. < 500	1
mol. wt. ≥ 500	0

ligands from, Wang, N., DeLisle, R. K. and Diller, D.J. (2005), J. Med. Chem., 48, 6980-6990

General versus Specific

When I know nothing

- invent distance / dissimilarity based on descriptors x, y, ..

If I know more, use an appropriate distance

- sequence example
  - Jukes-Cantor distance, p-value measure
- protein structures, metabolic pathways, small molecules
  - (geometric differences, similar reactions, bit strings)

Given some distances what are the methods ?

Clustering Methods

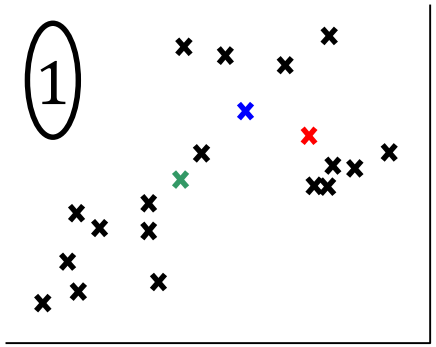
- *k*-means
- hierarchical
- fuzzy (not here)
- large data sets (not here)

k-means

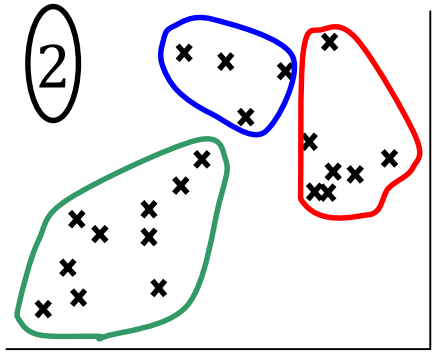
Pick *k* points - call them cluster centres

```
while (there is substantial change)
   assign each data-point to nearest centre
   re-calculate centres
```

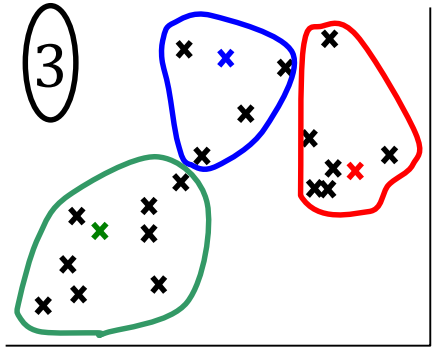
k-means steps



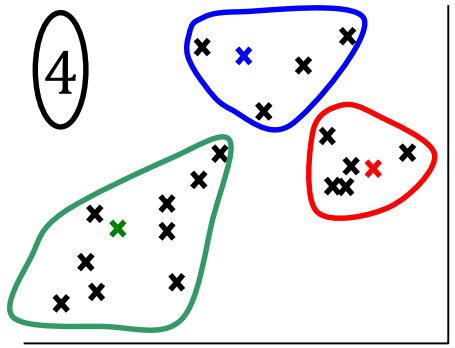
pick 3 points



allocate all other points



pick new centres



allocate all other points

k-means problems

What is *k* ?

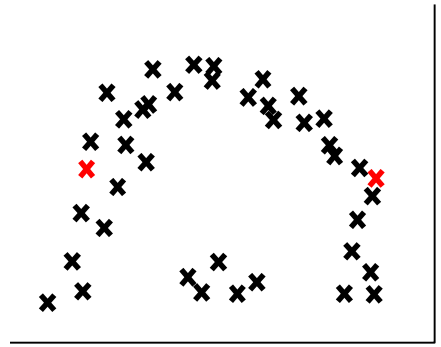
- guess, experiment, preconception

Initial choice of cluster centres

- requires concept of cluster centre (mean)

- non deterministic

- convergence



Cluster shape

- what if red points become centres ?

Hierarchical

Two flavours

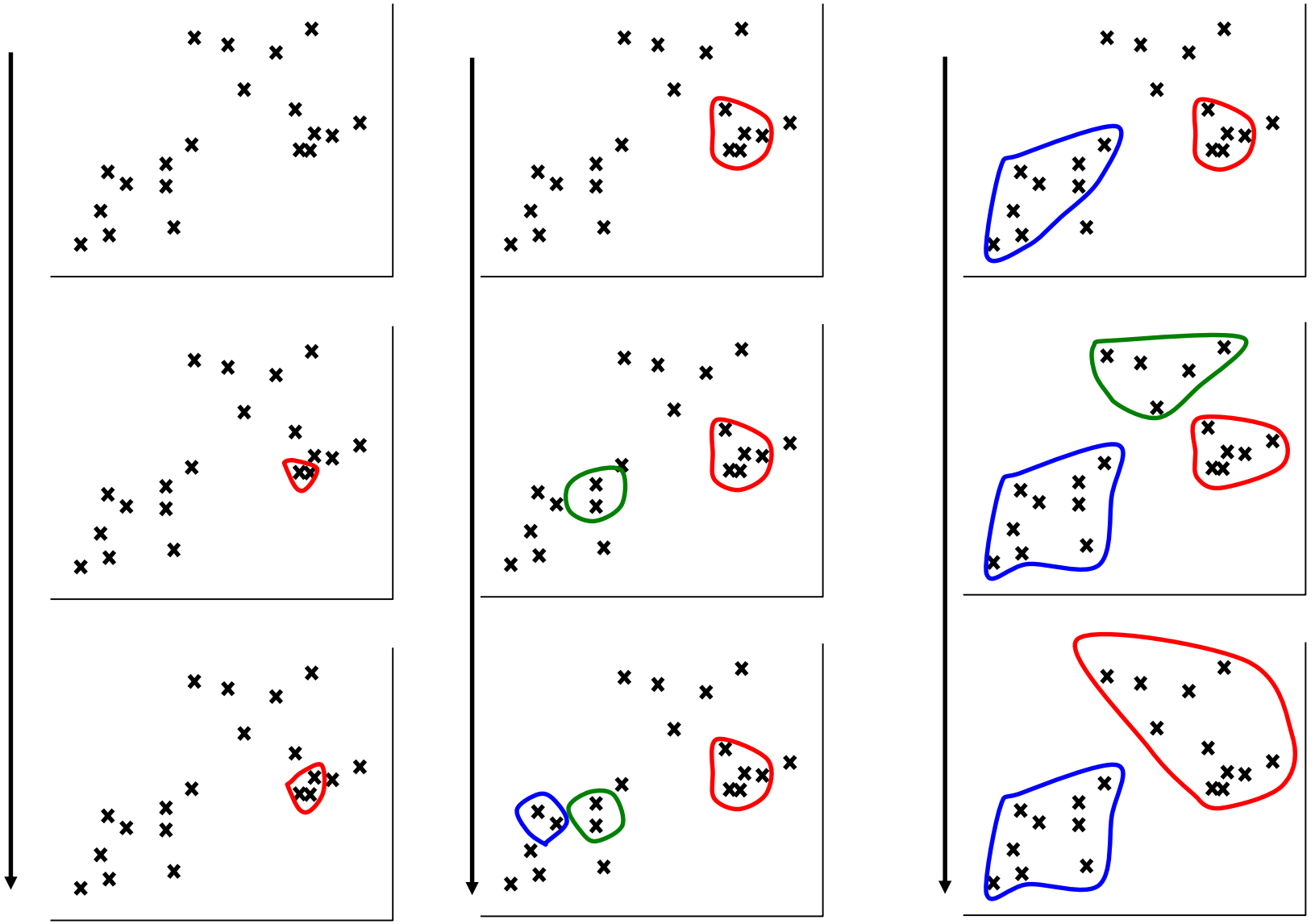
- divisive
- agglomerative / joining / nearest neighbour

agglomerative / joining / nearest neighbour

For *n* observations form *n* clusters (each point is separate)

```
while (not finished)
      find two nearest clusters (details later)
      join
```

agglomerative / joining example



Divisive

split_into_two (cluster)

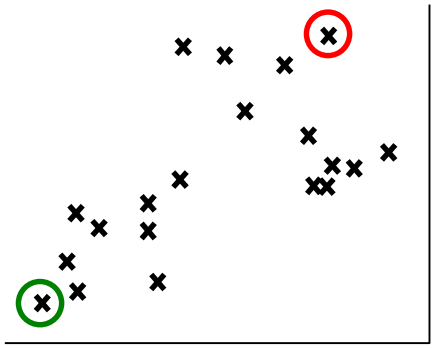
```
split_into_two (cluster)
   select two most separated points as centres of new clusters
   for each point in cluster
      allocate to nearest cluster centre
```

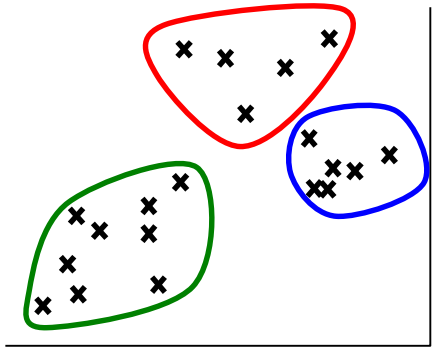
main procedure

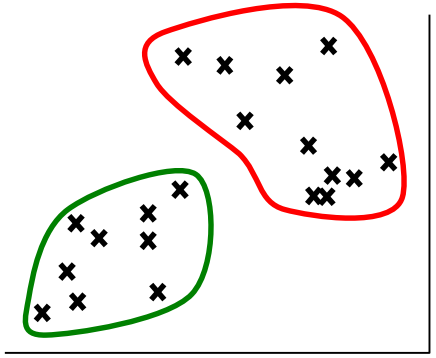
```
all points in one cluster
while (not finished)
    find largest cluster
    split_into_two (cluster)
```

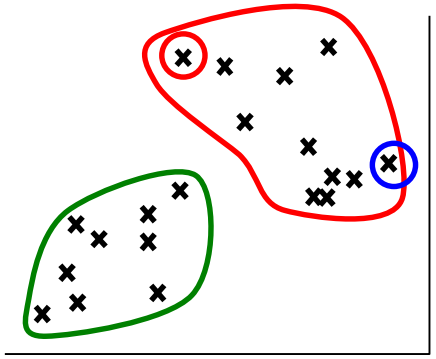
example

Divisive example

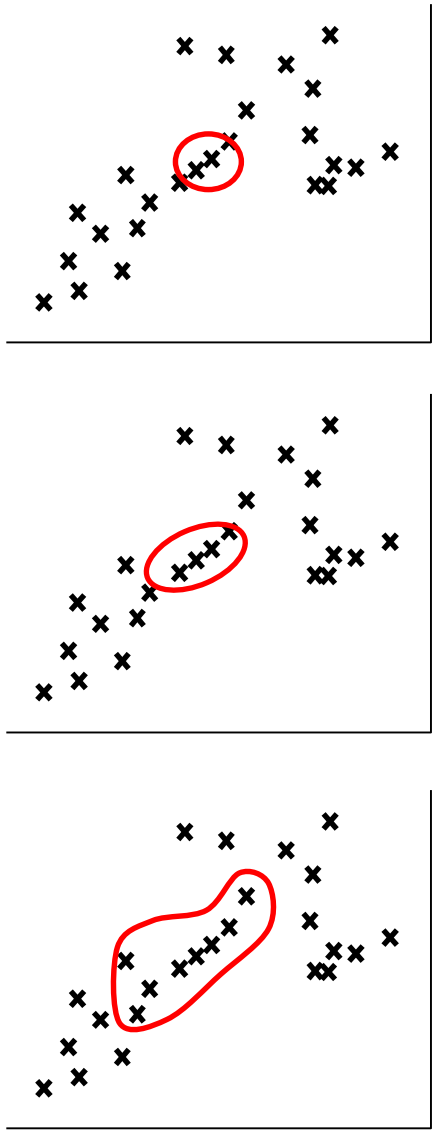








Breaking a joining method

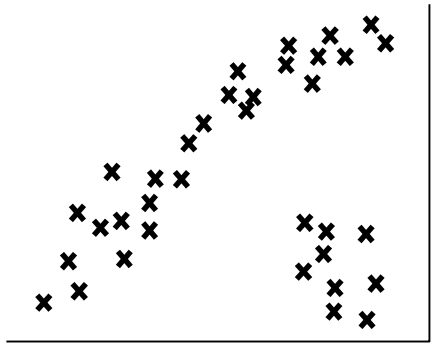


Consider this data with an agglomerative method

- distances are important, not compactness

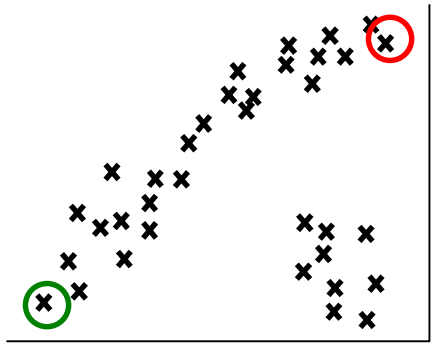
- is this always true ?

breaking a divisive method



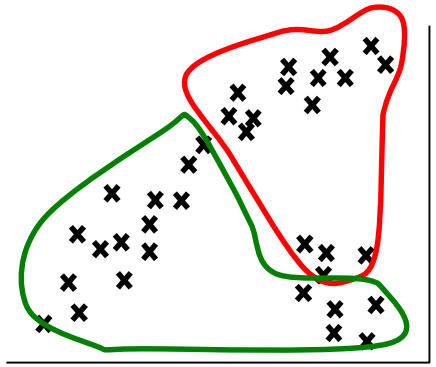
Method considers distances

- with this data
  - compactness of points is more important

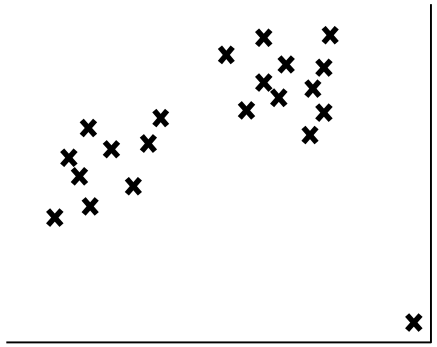


In many problems

- we only trust measures of high similarity
- example
  - molecular similarity
    - very different versus very very different

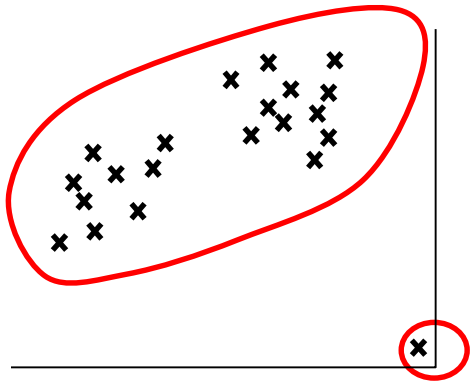


More ways to break joining methods



Different forms of neighbour joining

- earlier – "single linkage"
- sometimes "complete linkage" – use biggest distance between clusters

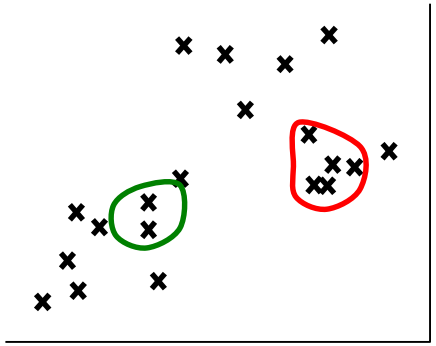


- susceptible to outliers
- relevant to Übung

cluster distances

Many details not touched

- what is cluster distance ? cluster centre ?
- distance between clusters ?



Distance between points is clear, but

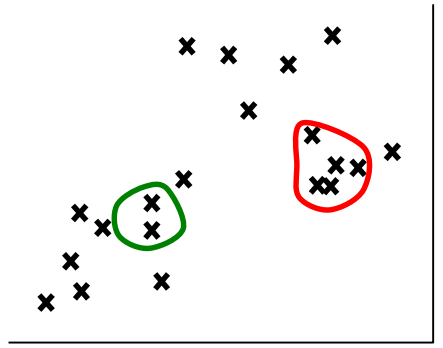
- between point and cluster
- between clusters ?

Sensible choices

- from cluster to nearest point
- from cluster to most typical point in other cluster

UPGMA

- in many bioinformatics texts
- unweighted pair group method using arithmetic averages
- take red points (5)
- take green points (2)
  - take average of all 2×5 distances
- debate over distance measures
  - similar to agglomerative versus divisive discussion
  - depends on structure of data



How complicated is clustering ?

In practice

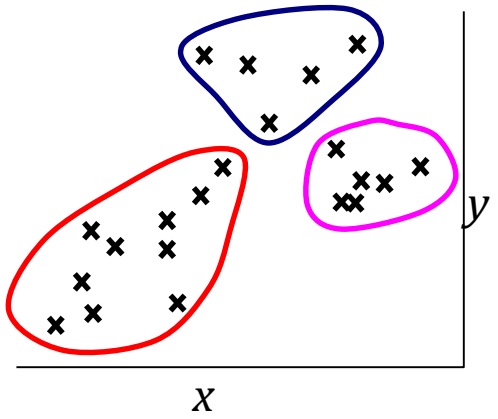
- distance based methods are best when a table of distances is available $O(n^2)$

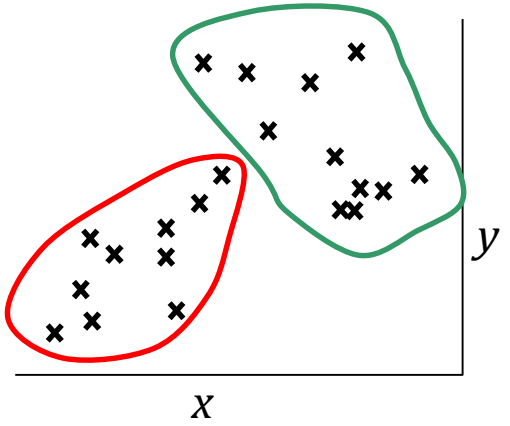
Problem in most fundamental form

- unknown k-clusters
- combinatorial possibilities huge

Formalise our goal

- maximise density within clusters
- maximise distance between clusters
- should be able to distinguish
  - 2 from 3 cluster answers





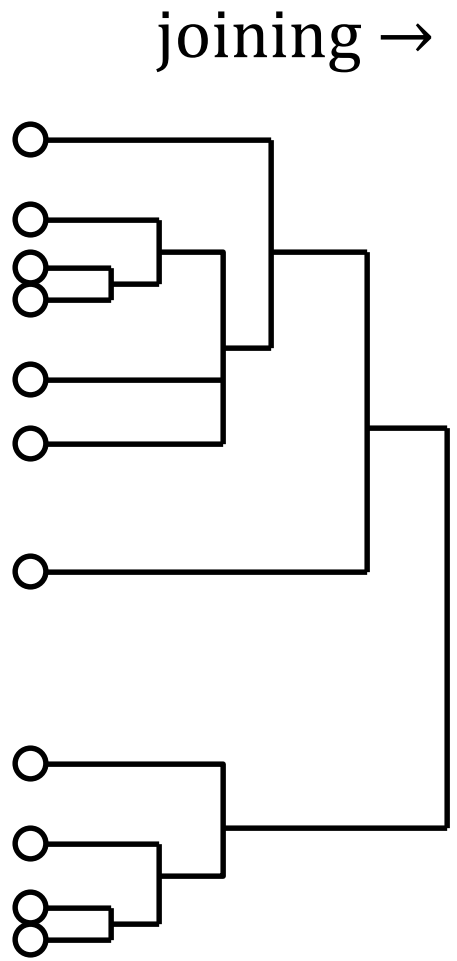
Are we finished ?

Lots of decorations

- iterations over cluster memberships
- different definitions of distances, centres

Mixing x, y, z continuous descriptors and categories (red/blue/..)

Dendrograms



- assumption of hierarchy
- what you call the "classification" depends on where you want to cut tree
- protein shape example
  - most detailed level
    - very similar protein sequences

Applications - sequences

Sequence comparison

- distances ?
  - evolutionary estimates or
  - similarity based on statistics (*p*-values)
  - clear model (evolution) - suits hierarchy

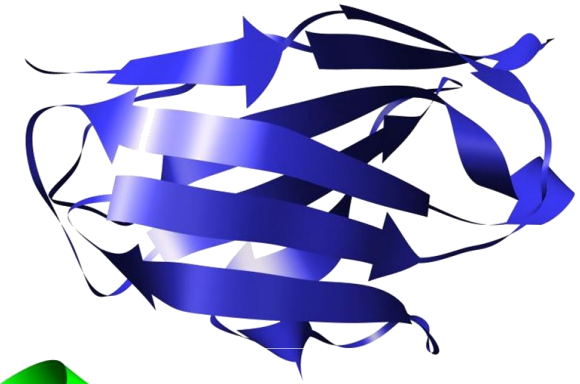
  - related sequences
    - distances OK
  - less related sequences
    - time estimates unreliable (J-C model)
    - alignments unreliable

Applications - protein structure

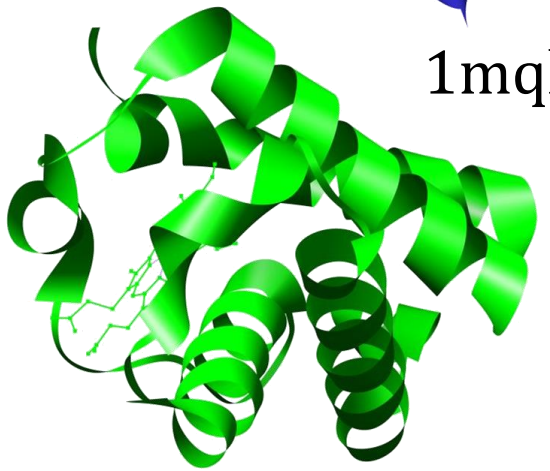
3 proteins of similar size

- 1bww and 1mqk easy (immunoglobulins human/mouse)
  - not easy to compare against 1dlw (globin shape)

1bww



1mqk



1dlw

	1bww	1mqk	1dlw
1bww	0	easy	?
1mqk		0	?
1dlw			0

Applications - protein structure

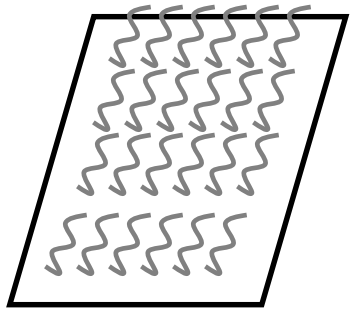
Are we lost ?

- easiest to tackle problems with joining methods

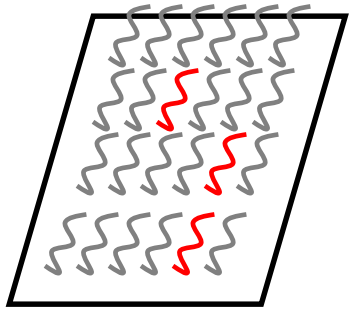
Applications - microarray data

What are microarrays ?

- little slabs with pieces of DNA bound



microarrays



- lots of bits of DNA from known genes (complementary)
- pour on a sample from cells with mRNA
  - some binds
  - detect by fluorescence
  - have a look which bits of DNA on chip were affected - tells us which genes were involved
  - we know which genes were activated in the original soup

microarrays

- feed sugar to cells
  - pour on to microarray - who lights up ?
  - boring
- feed lipids to cells
  - who lights up
- feed … to cells
- starve cells, heat cells, find cells with disease

Are there groups of genes whose regulation is similar ?

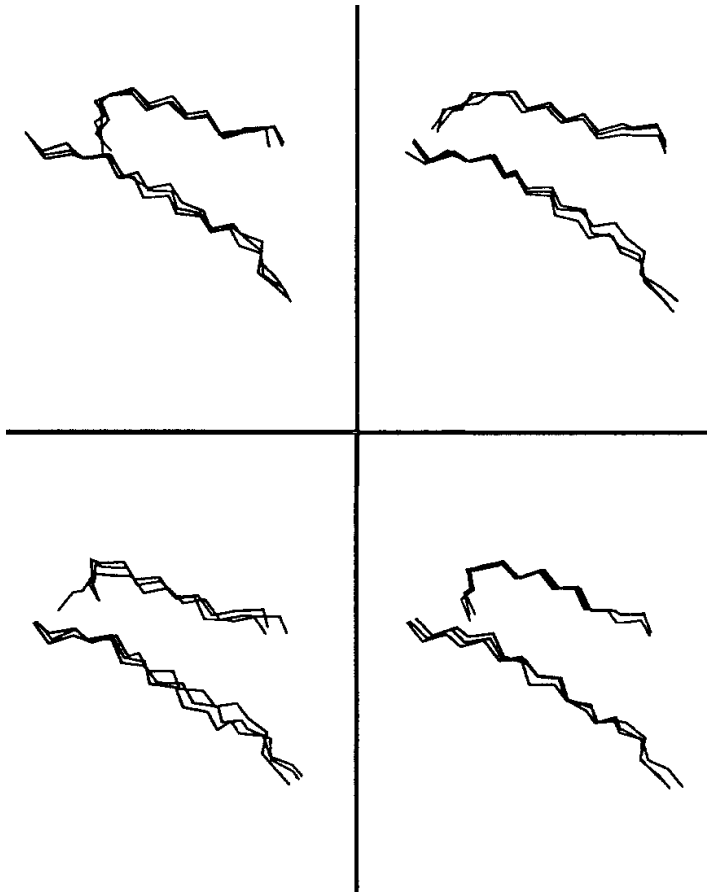
- should let you find genes in pathways / regulation mechanisms

protein structure

Simulate a protein molecule and see 10^8 configurations

- is the molecule constantly changing or sometimes leaving and returning to conformations ?

- does not look like much..
  backbone atoms only
- long molecular dynamics simulation
  - 4 major clusters selected
  - each represented by centre
    + two outliers



Distance measures – common theme

- similar protein structures
- similar sequences
- cells with similar behaviour

- closer relationships are more reliable

Summary

- Rarely is there a correct answer
- Method of choice may depend on data

Best case

- reliable distances known between all points

Real problems

- noise / outliers

Distance measures

- close relationships are usually more reliable

Running time ?

- O(n^2) for dissimilarity matrix
- method dependent - usually less than O(n^2)