Data Mining
Lecture 3: Discovering Groups

Jo Houghton

ECS Southampton

January 11, 2019

Discovering Groups - Introduction

Understanding large datasets is hard, especially if it has high dimensional features
To help understand a dataset:

- Find similar data items
- Find similar features

Discovering Groups - Documents

For example: Documents
We use a 'Bag of Words'
e.g. "The quick brown fox jumped over the lazy dog."
becomes:

'The': 1, 'quick': 1, 'brown': 1, 'fox': 1, 'jumped': 1, 'over': 1, 'the': 1, 'lazy': 1, 'dog.': 1

Discovering Groups - Documents

For example: Documents
We use a 'Bag of Words'
e.g. "The quick brown fox jumped over the lazy dog."
becomes:

'The': 1, 'quick': 1, 'brown': 1, 'fox': 1, 'jumped': 1, 'over': 1, 'the': 1, 'lazy': 1, 'dog.': 1

After some tokenising and sorting, you could get:
'brown': 1, 'dog': 1, 'fox': 1, 'jumped': 1, 'lazy': 1,'over': 1, 'quick': 1, 'the': 2

Discovering Groups - Documents

For example: Documents
We use a 'Bag of Words'
e.g. "The quick brown fox jumped over the lazy dog."
becomes:

'The': 1, 'quick': 1, 'brown': 1, 'fox': 1, 'jumped': 1, 'over': 1, 'the': 1, 'lazy': 1, 'dog.': 1

After some tokenising and sorting, you could get:
'brown': 1, 'dog': 1, 'fox': 1, 'jumped': 1, 'lazy': 1,'over': 1, 'quick': 1, 'the': 2

Further stemming and removal of stop words could give:
'brown': 1, 'dog': 1, 'fox': 1, 'jump': 1, 'lazy': 1,'over': 1, 'quick': 1

Discovering Groups - Documents

The bag of words vector for a document will be very sparse

Each document will only have a small fraction of the total number of words in the English language there are over 250,000 words in the English language

In recommender systems, the sparsity was due to missing data. Here, those zero values are meaningful, i.e. that a word was not there.

"I love you", "I don't love you"
Have very different meanings

Discovering Groups - Documents

As an example dataset, we use the syllabus description pages for all ECS COMPXXXX modules.

- The text is tokenized by splitting on [^A-Z^a-z]+ (using regular expressions)
- Terms with document frequency $< 10\%$ and $> 70\%$ are removed
- Words of length 2 or less are removed

Discovering Groups - Documents

For example:

> Courses >

COMP6237 Data Mining

Module Overview

The challenge of data mining is to transform raw data into useful information and actionable knowledge. Data mining is the computational process of discovering patterns in data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and data management.

This course will introduce key concepts in data mining, information extraction and information indexing; including specific algorithms and techniques for feature extraction, clustering, outlier detection, topic modelling and prediction of complex unstructured data sets. By taking this course you will be given a broad view of the general issues surrounding unstructured and semi-structured data and the application of algorithms to such data. At a practical level you will have the chance to explore an assortment of data mining techniques which you will apply to problems involving real-world data.

Module Details

Semester 2
CATS points: **15**
ECTS points: **7.5**
Level 7
Module Lead: **Markus Brede**

Becomes:

COMP6237: Data Mining 1 0 10 3 1 0 0 12 3 0 0 1 1 0 0

2 0 0 0 0 0 5 0 0 0 0 0 0 0 21 0 2 0 1 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 1 0 0 0 1 2 0 0 0 0 0 0 2 0 2 0 0 0 0 0 1 0

1 1 3 0 0 0 0 1 0 0 0 0 0 0 1 2 2 0 1 0 0 0 0 0 0 0 0 1 0

0 0 0 2 0 0 0 0 0 0 2 0 3 0 1 7 0 0 0 0 0 3 1 0 0 0 0 0 0

0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 4 0 0 0 3 0 0 0 0 1 0 0

0 3 0 0 0 0 0 0 0 0 0 1 0 0 4 1 0 0 1 0 0 0 0 0 0 0 0 0 1

0 1 2 1 0 0 0 0 1 1 2 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

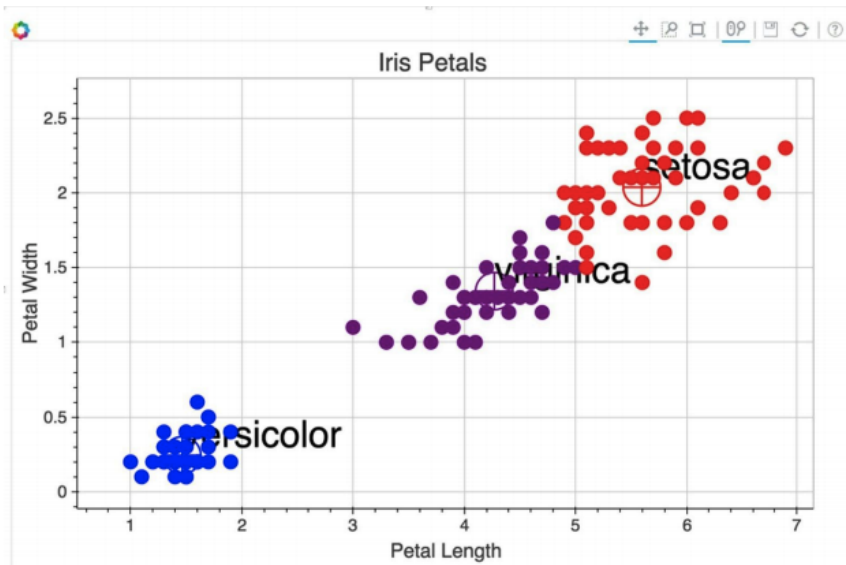
0 0 1 0 0 0 0 0 0 0 0 0 2 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0

0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 12

Discovering Groups - Clustering

Grouping data, just using the feature vectors

- Unsupervised
- Similar feature vectors grouped together
- Can be
  - Soft (allow overlapping groups)
  - Hard (each item assigned to one group)



Discovering Groups - Hierarchical Clustering

Hierarchical Clustering:

Creates a binary tree that recursively groups pairs of similar items or clusters

Can be:

- Agglomerative (bottom up)
- Divisive (top down)

Discovering Groups - Hierarchical Clustering

Algorithm 1: Hierarchical Agglomerative Clustering

Data: N data points with feature vectors X_i $i = 1 \ldots N$

$numClusters = N$;

while $numClusters > 1$ **do**

    cluster1, cluster2 = FindClosestClusters();

    merge(cluster1, cluster2);

end

The distance between the clusters is evaluated using a linkage criterion.

If each merge is recorded, a binary tree structure linking the clusters can be formed.

This gives a **dendrogram**

Discovering Groups - Hierarchical Clustering

Linkage criterion: A measure of dissimilarity between clusters
Centroid Based:

- Dissimilarity is equal to distance between centroids
- Needs numeric feature vectors

Distance-Based:

- Dissimilarity is a function of distance between items in clusters
- Only needs precomputed measure of similarity between items

Discovering Groups - Hierarchical Clustering

Centroid based linkage:

- WPGMC: Weighted Pair Group Method with Centroids
  When two clusters are combined into a new cluster, the average of the two centroids is the new centroid
- UPGMC: Unweighted Pair Group Method with Centroids
  When two clusters are combined into a new cluster, the new centroid is recalculated based on the positions of the items

Discovering Groups - Hierarchical Clustering

Distance based linkage:

- **Minimum**, or **single-linkage clustering** Distance between two closest members

$$\min d(a, b) : a \in A, b \in B$$

  Produces long, thin clusters
- **Maximum**, or **complete-linkage clustering** Distance between two most distant members

$$\max d(a, b) : a \in A, b \in B$$

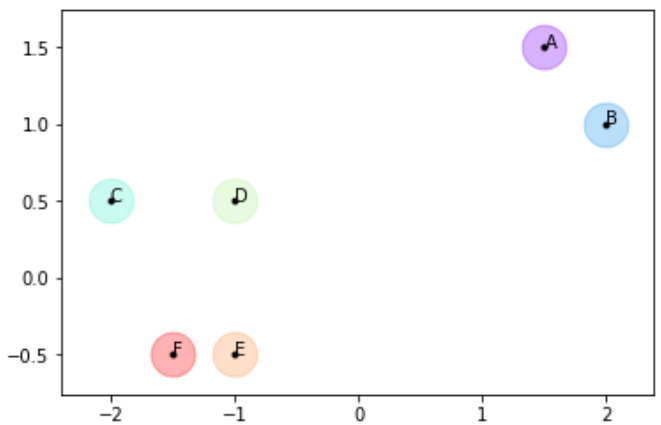
  Finds compact clusters, approximately equal diameter
- **Mean** or **Average Linkage Clustering** (**UPGMA**: Unweighted Pairwise Group Method with Arithmetic Mean):

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

Discovering Groups - Hierarchical Clustering

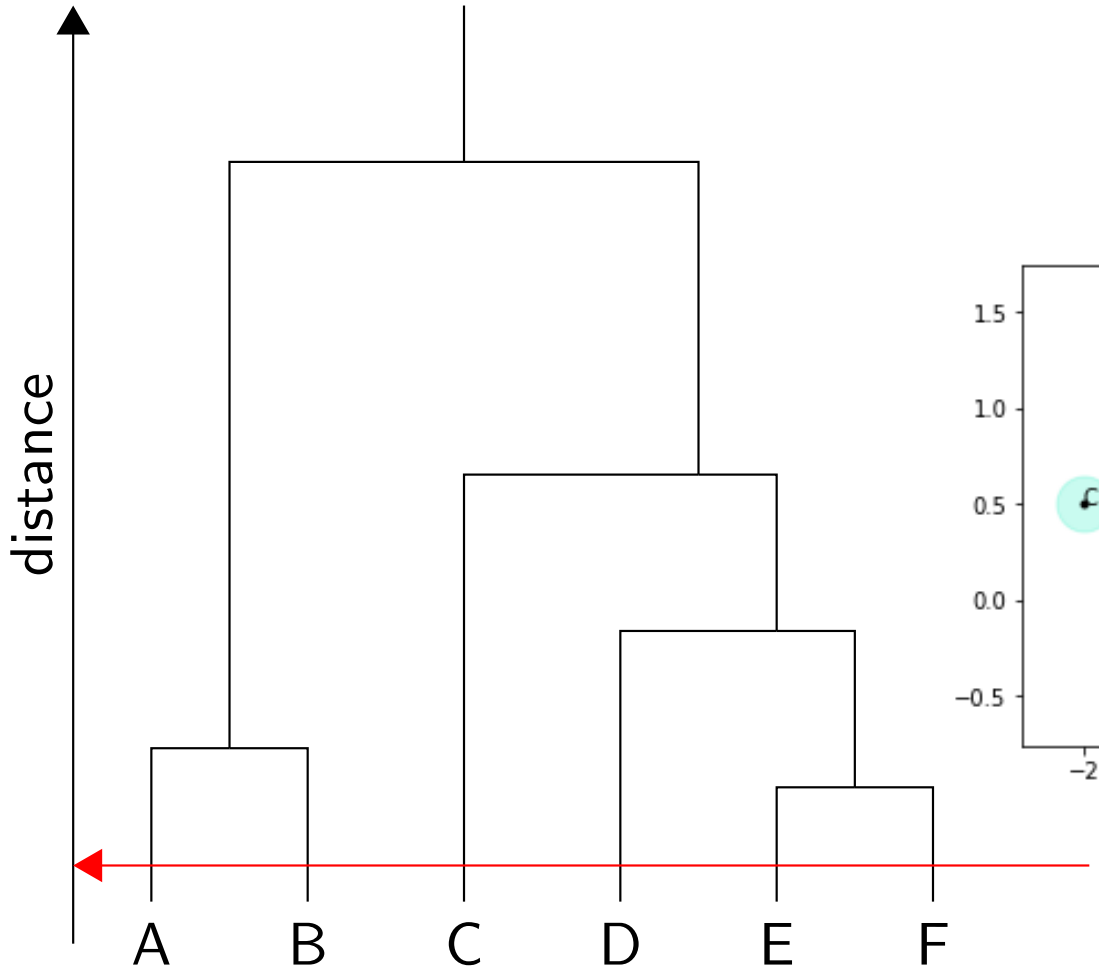
With sample data:

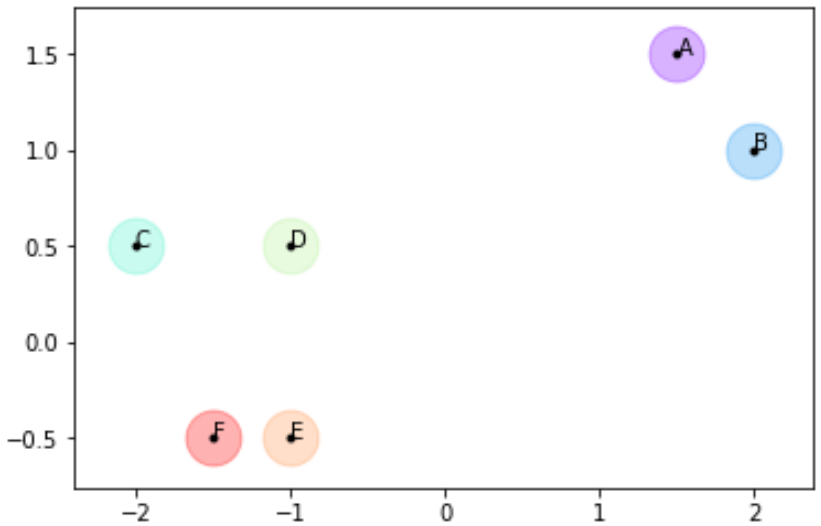
$$X = \begin{bmatrix} 1.5 & 1.5 \\ 2.0 & 1.0 \\ 2.0 & 0.5 \\ -1.0 & 0.5 \\ -1.5 & -0.5 \end{bmatrix}$$



Using minimum or single-linkage clustering

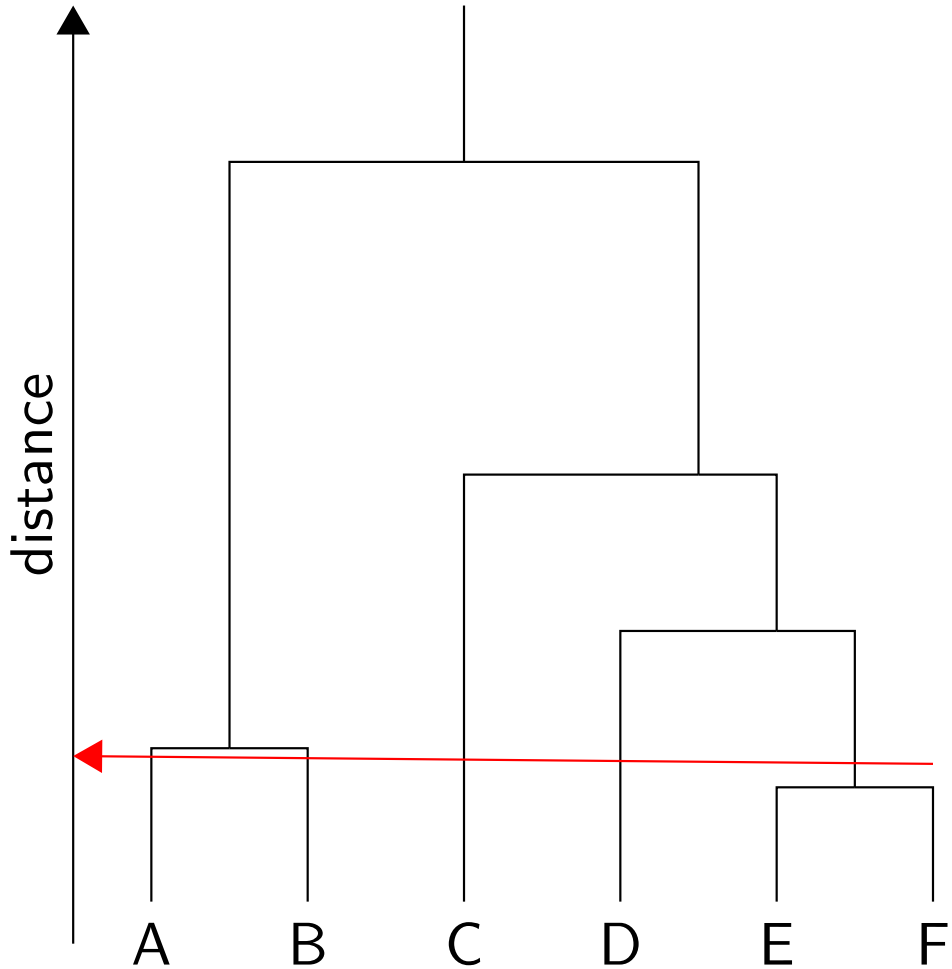
Discovering Groups - Centroid Clustering

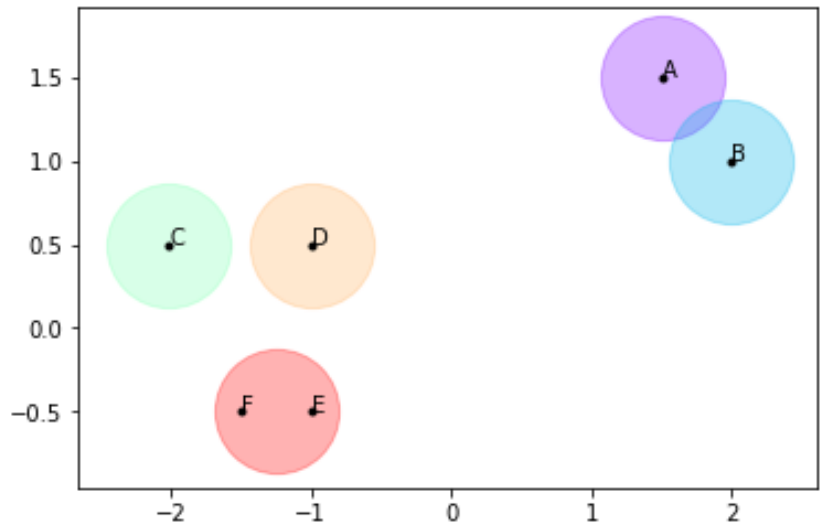




Using minimum or single-linkage clustering

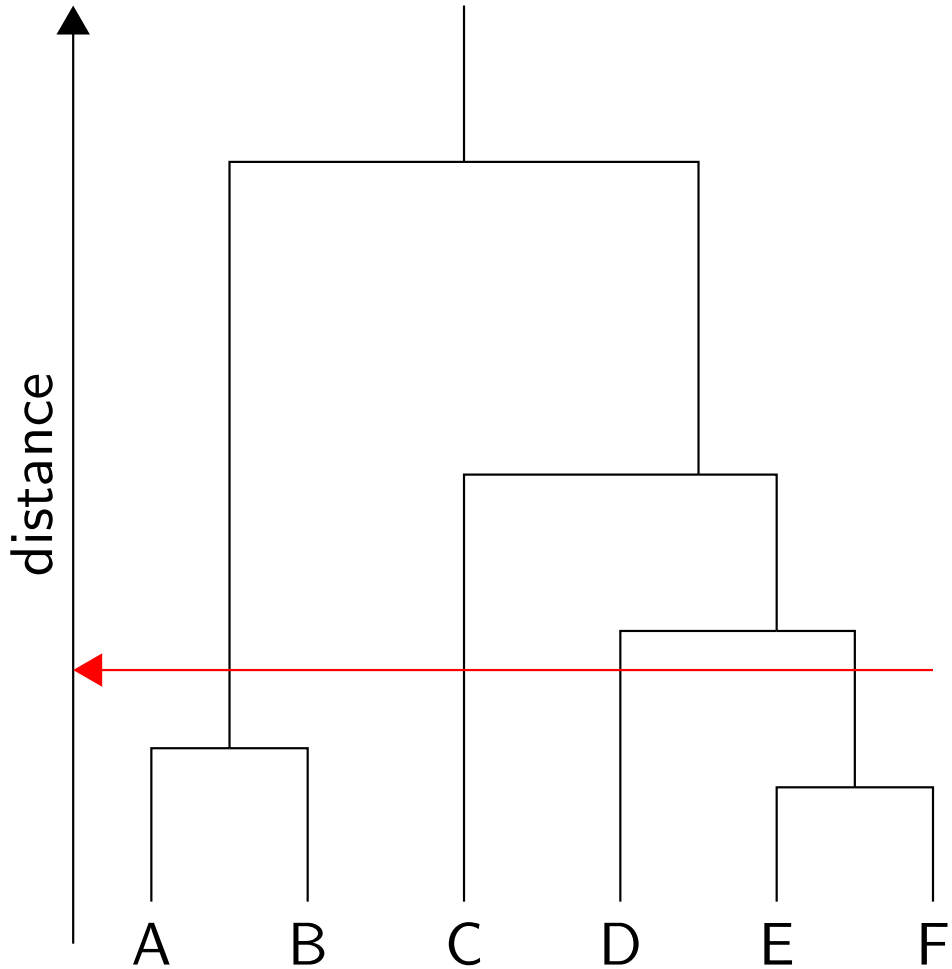
Discovering Groups - Centroid Clustering

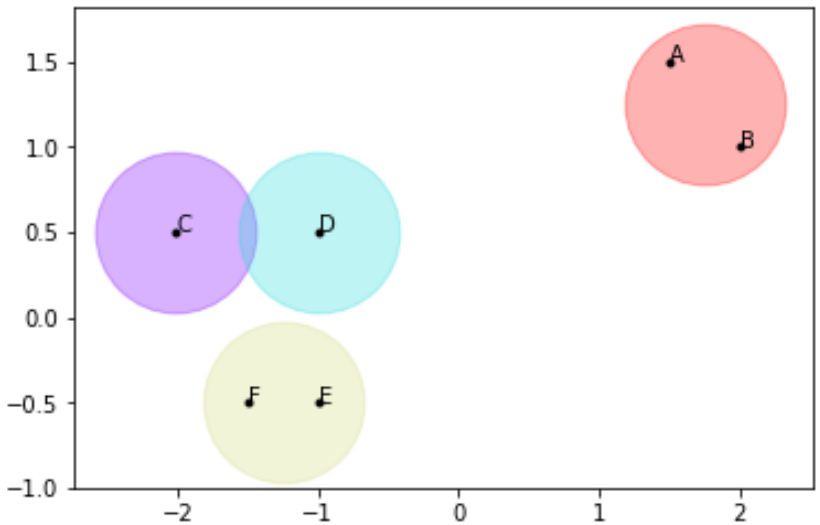




Using minimum or single-linkage clustering

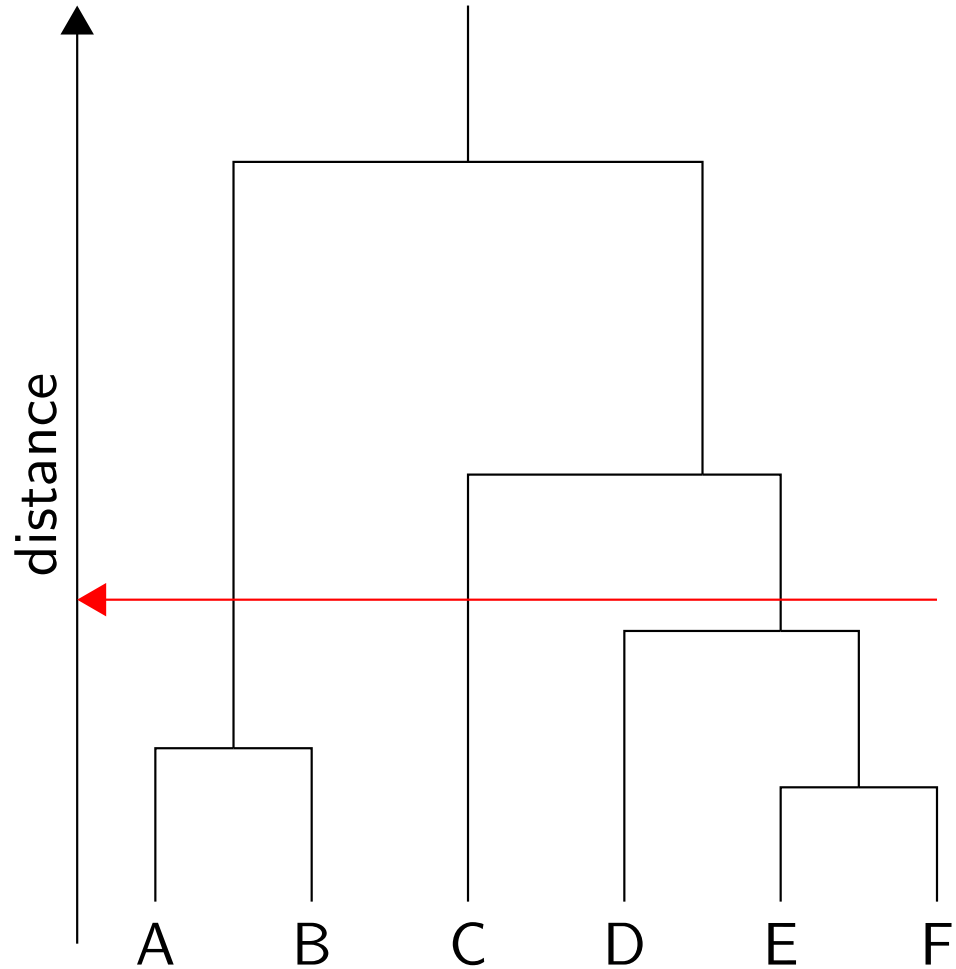
Discovering Groups - Centroid Clustering

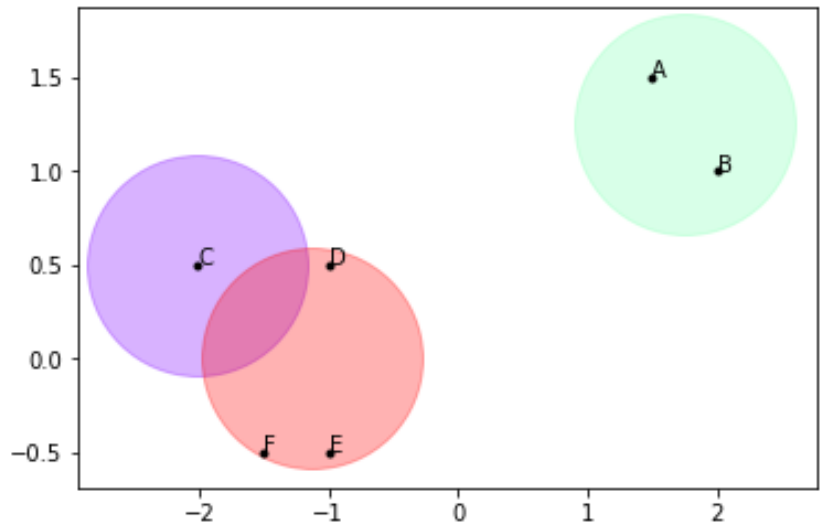




Using minimum or single-linkage clustering

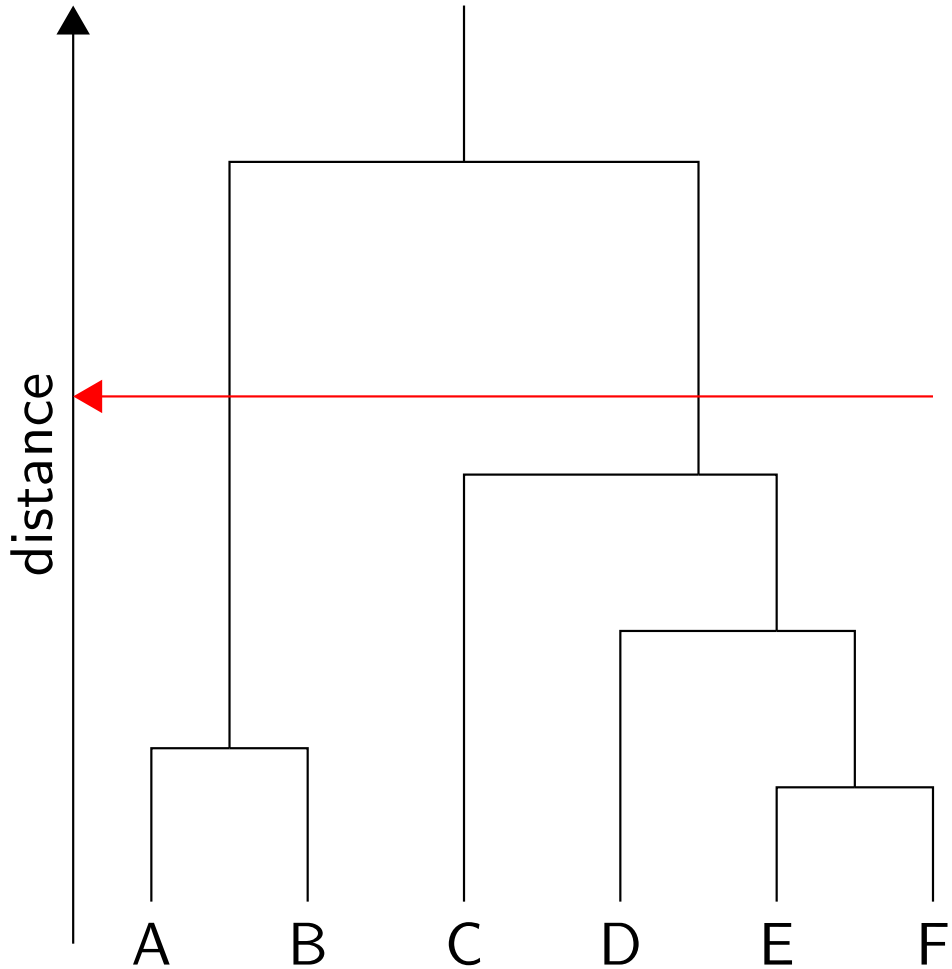
Discovering Groups - Centroid Clustering

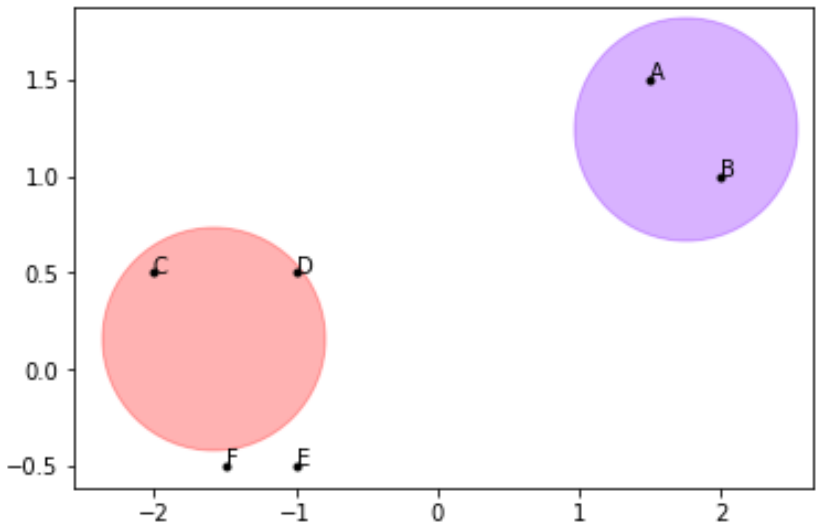




Using minimum or single-linkage clustering

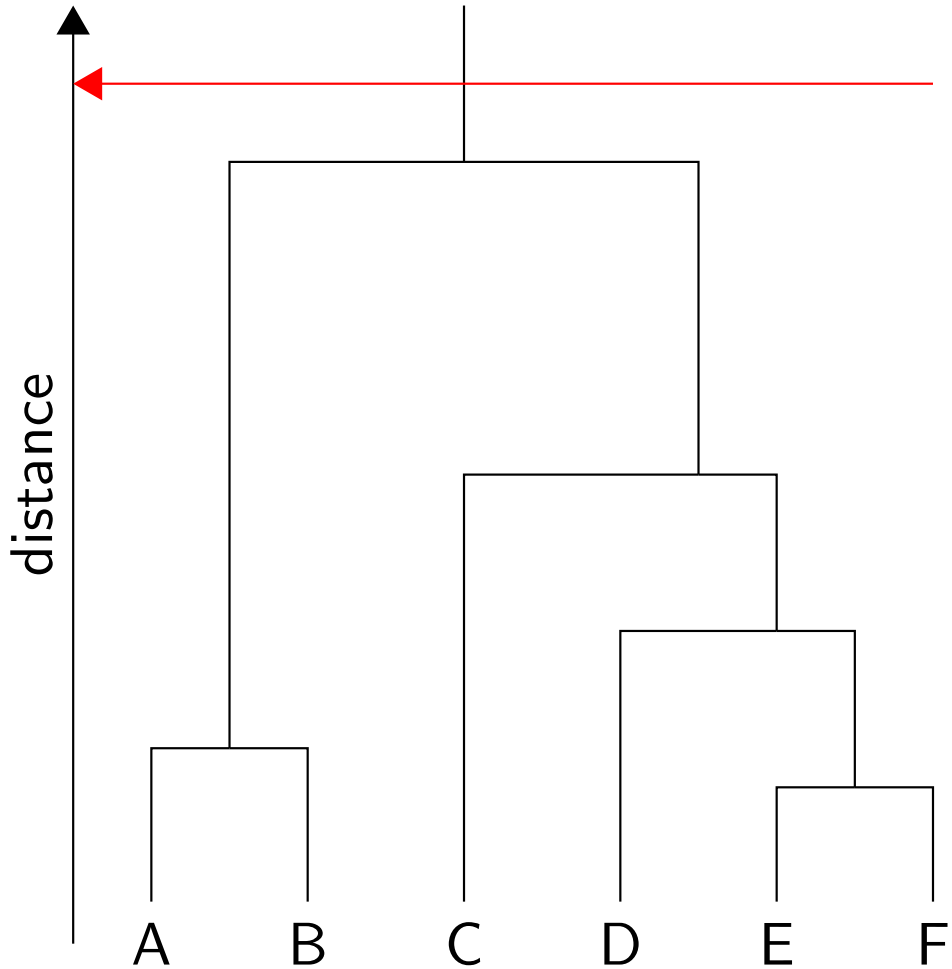
Discovering Groups - Centroid Clustering

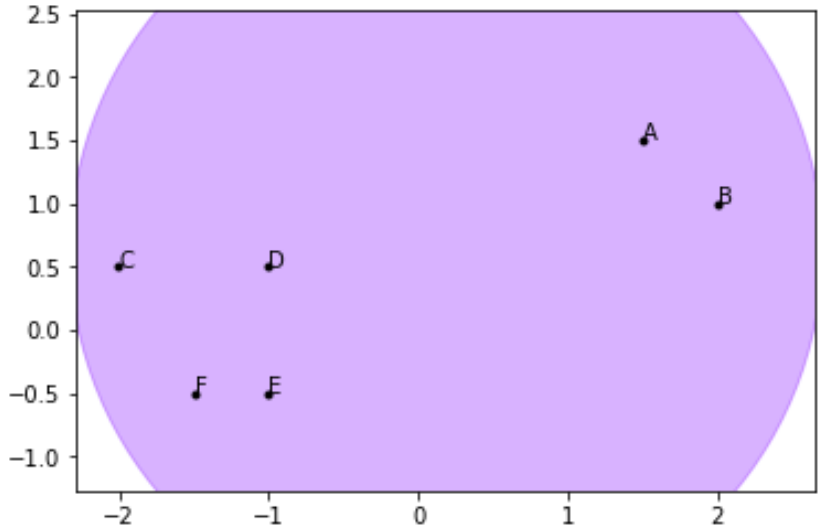




Using minimum or single-linkage clustering

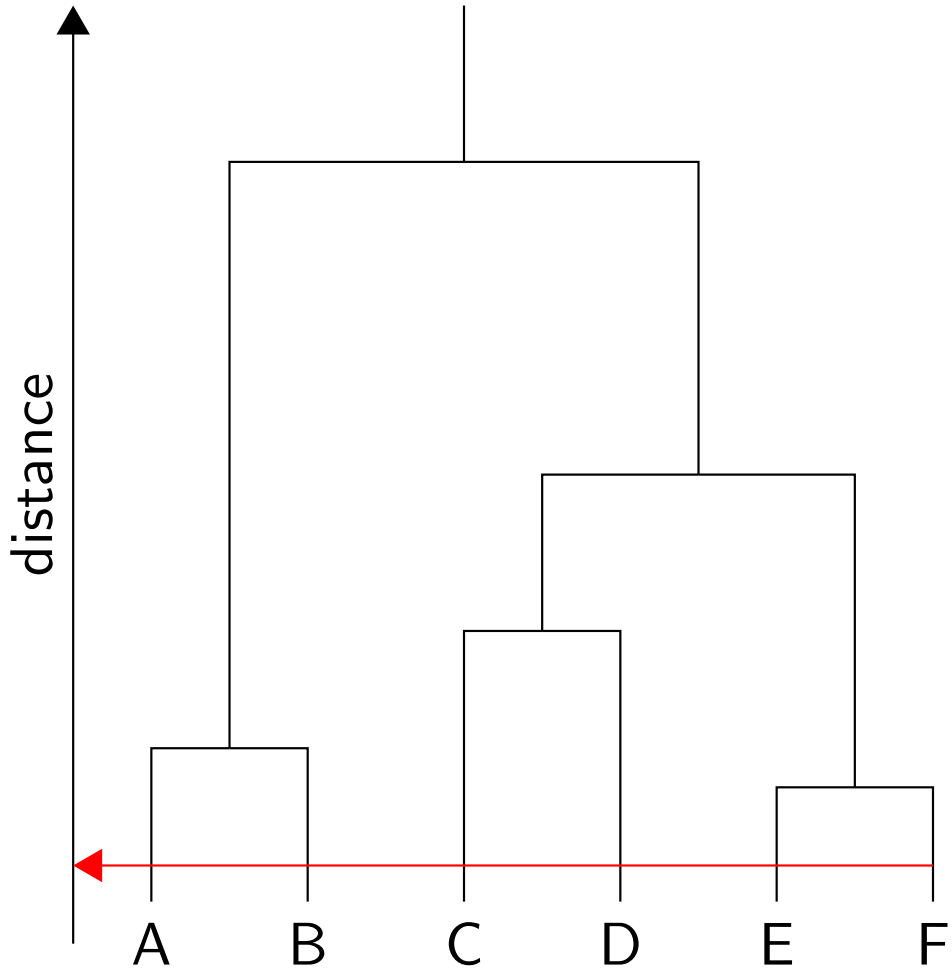
Discovering Groups - Centroid Clustering

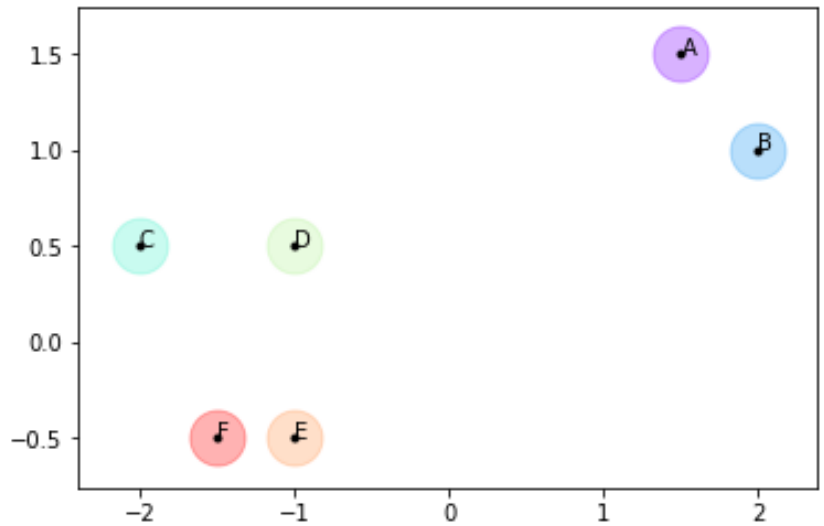




Using minimum or single-linkage clustering

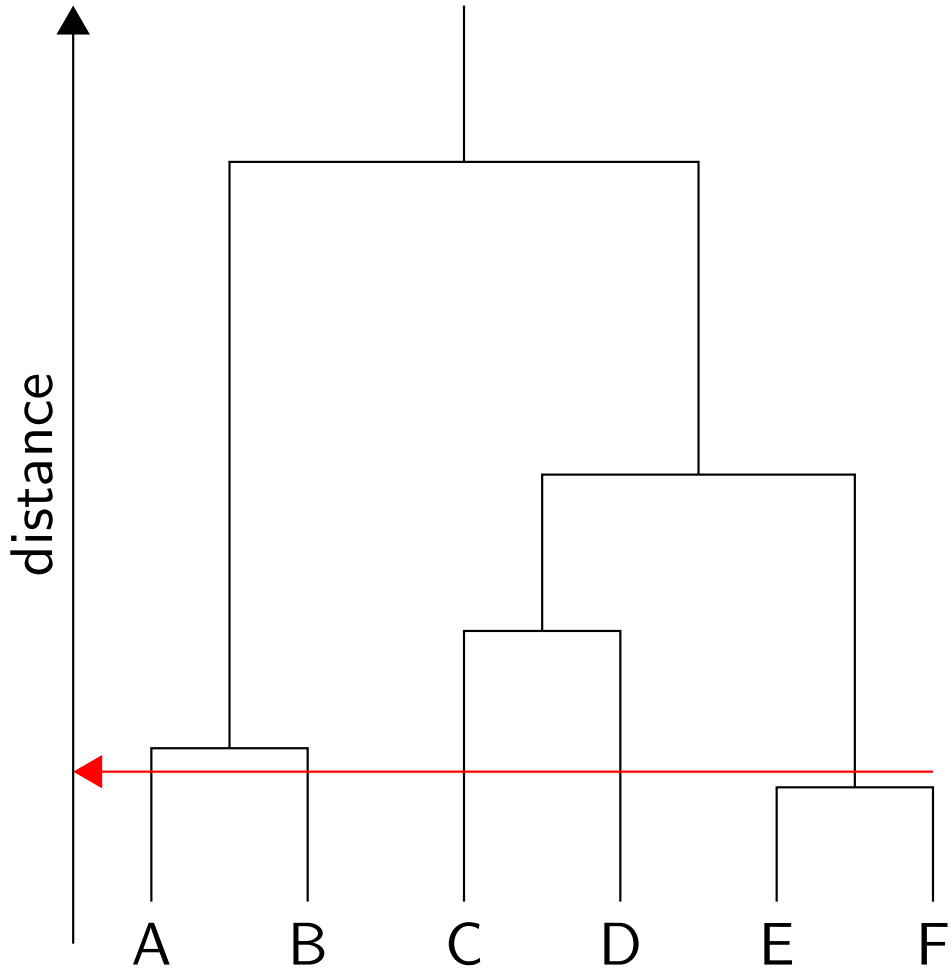
Discovering Groups - Centroid Clustering

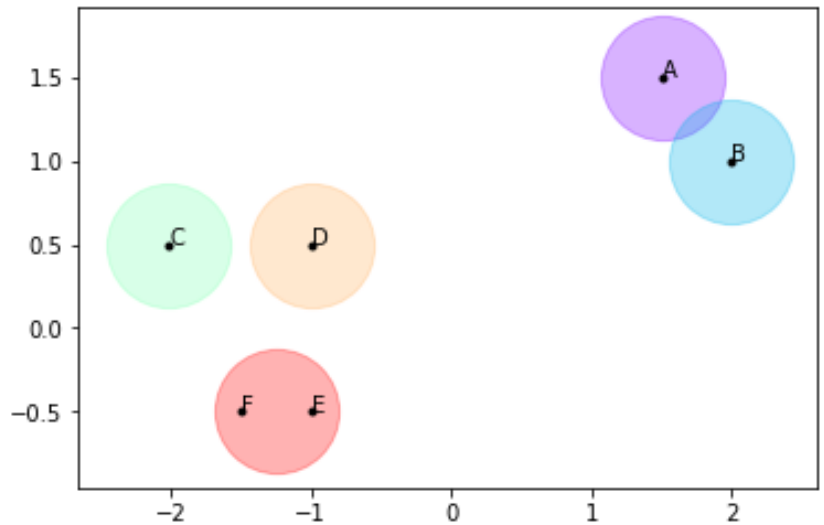




Using maximum or complete-linkage clustering

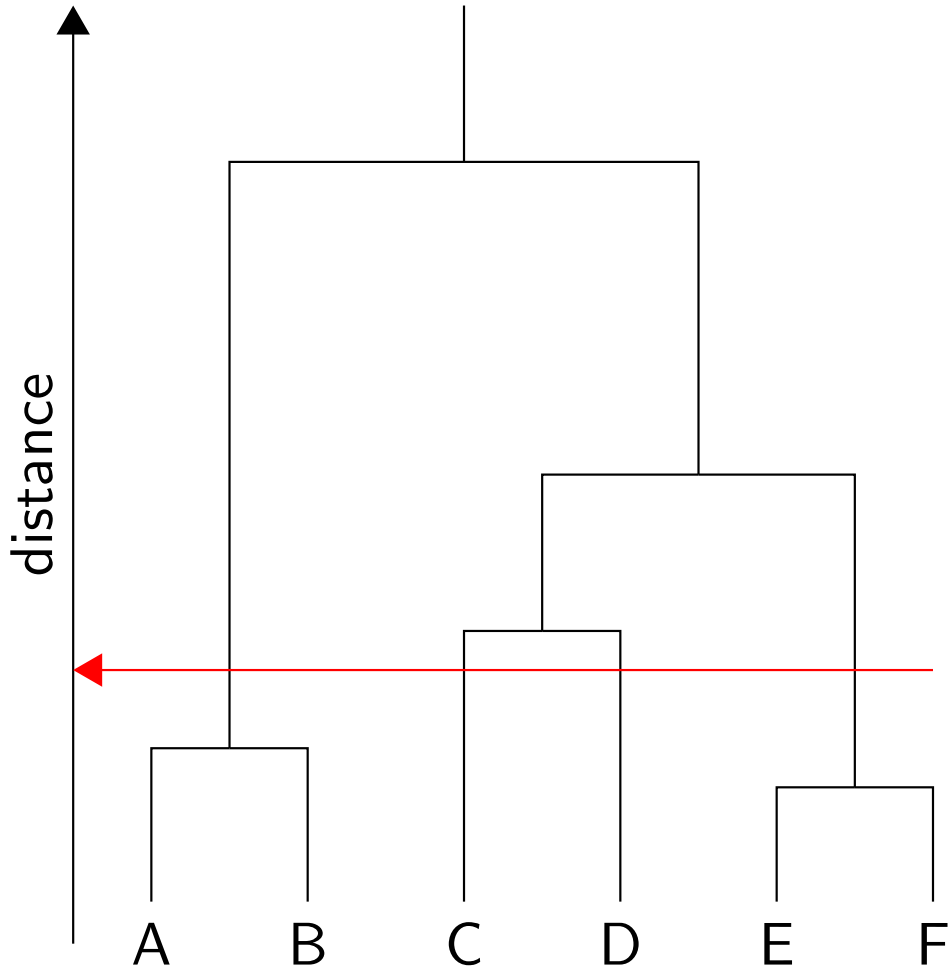
Discovering Groups - Centroid Clustering

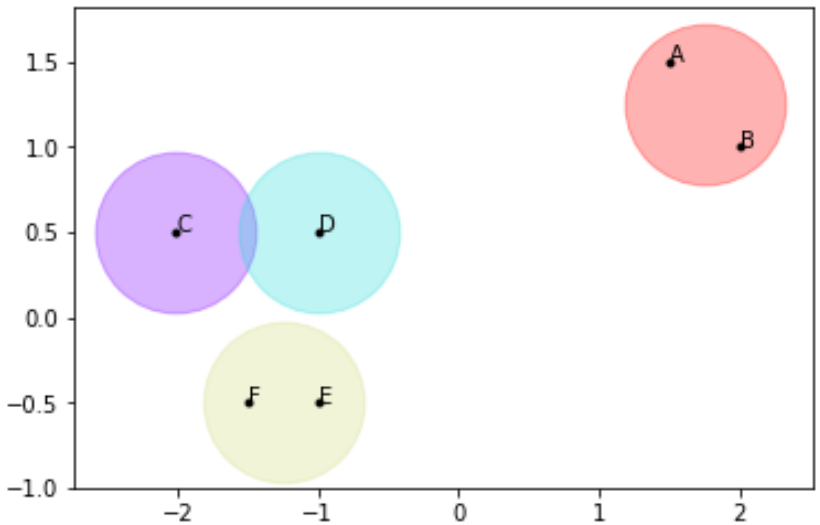




Using maximum or complete-linkage clustering

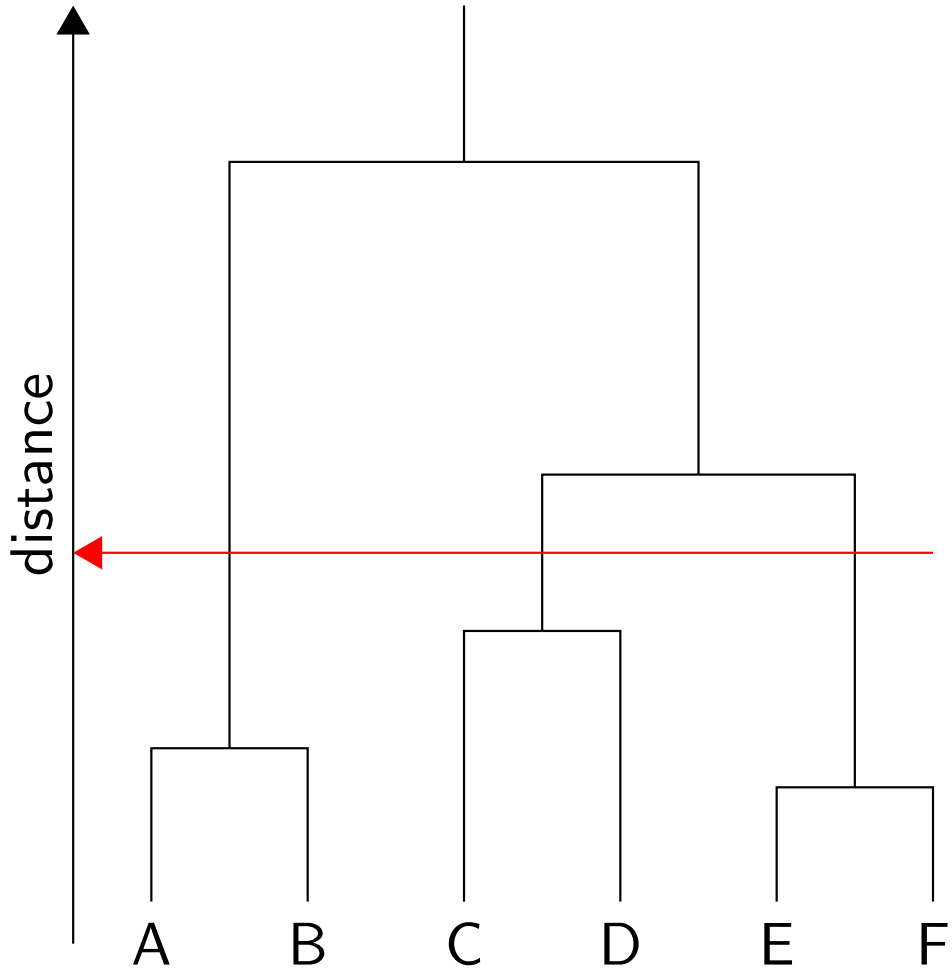
Discovering Groups - Centroid Clustering

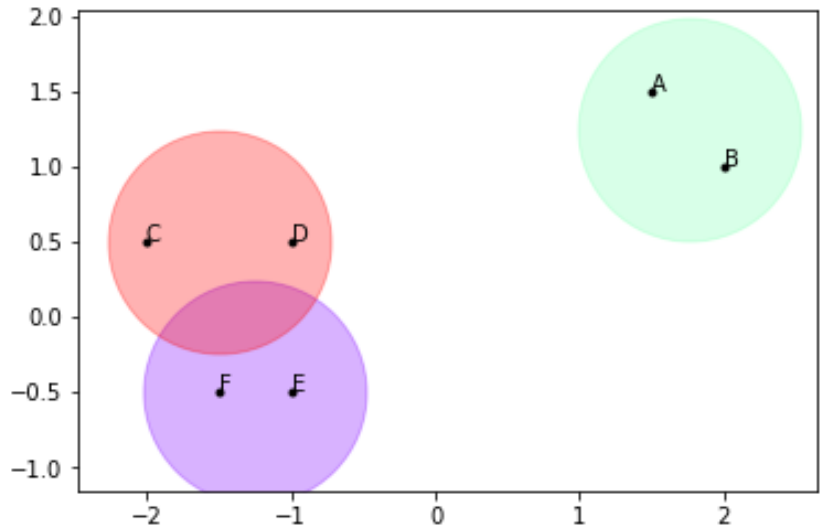




Using maximum or complete-linkage clustering

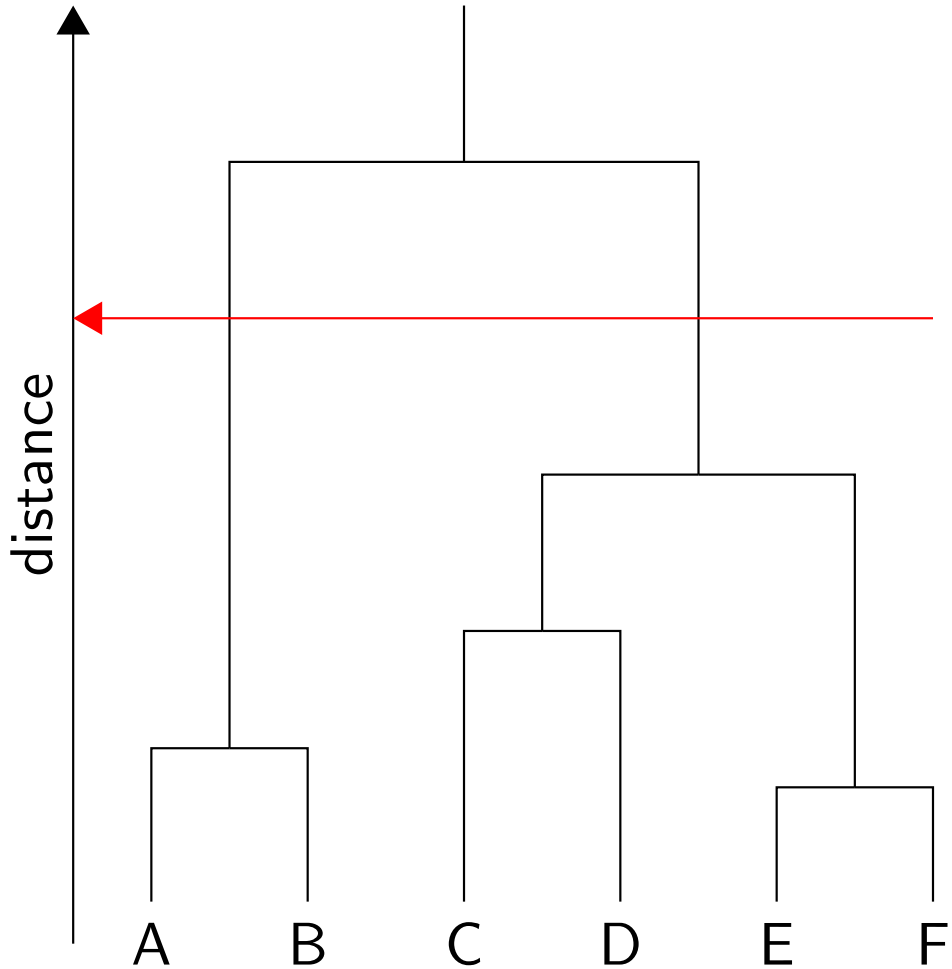
Discovering Groups - Centroid Clustering

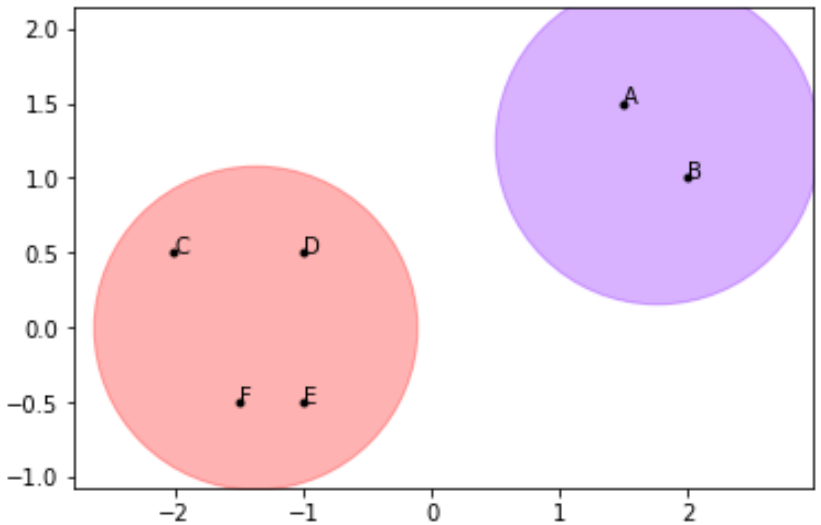




Using maximum or complete-linkage clustering

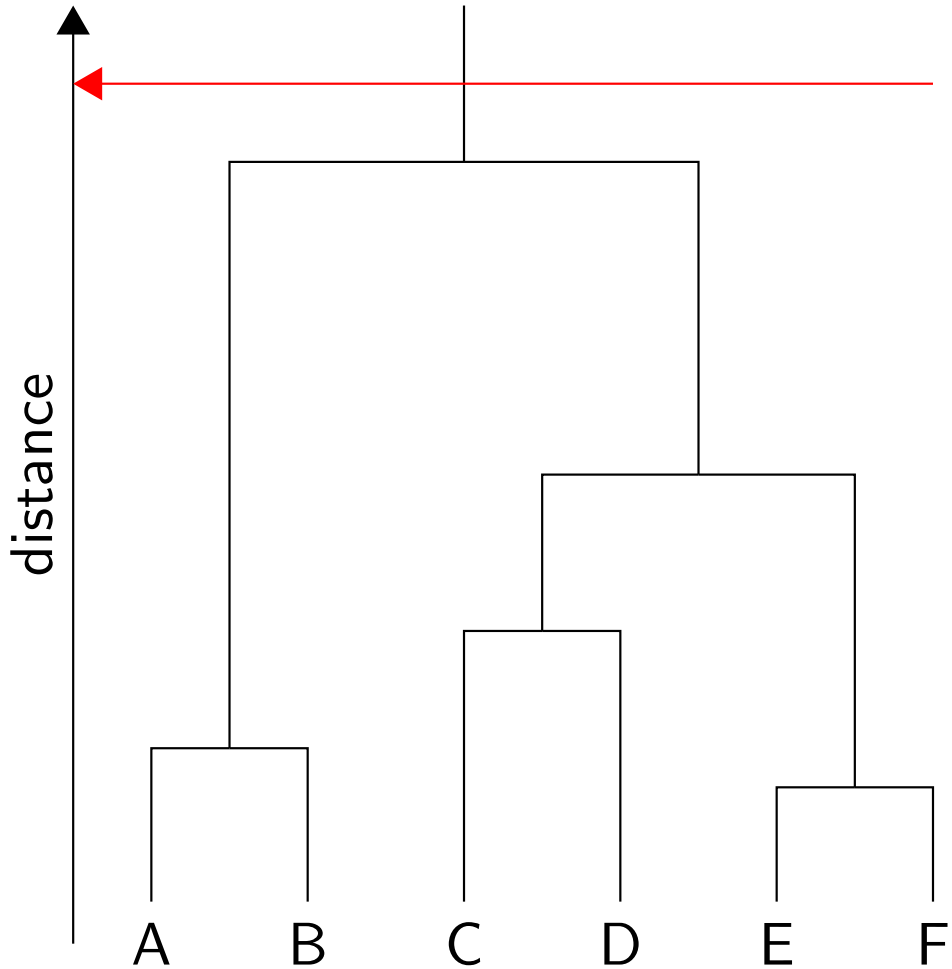
Discovering Groups - Centroid Clustering

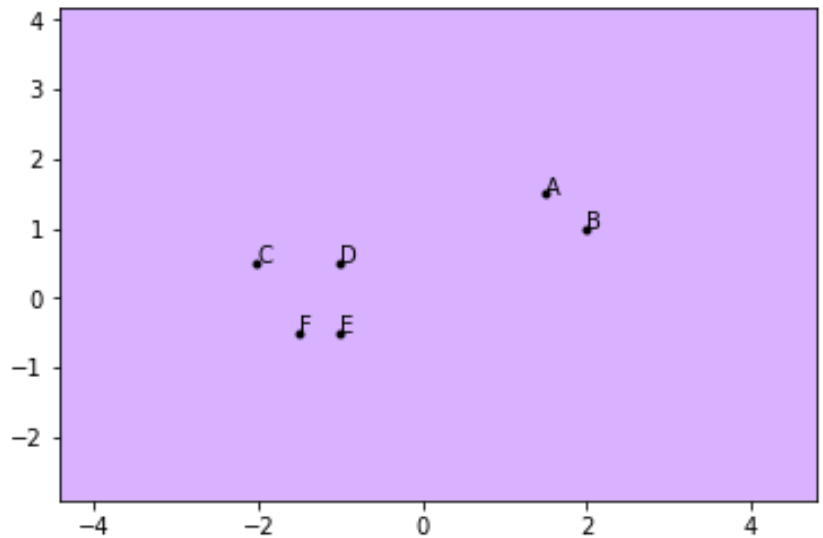




Using maximum or complete-linkage clustering

Discovering Groups - Centroid Clustering





Using maximum or complete-linkage clustering

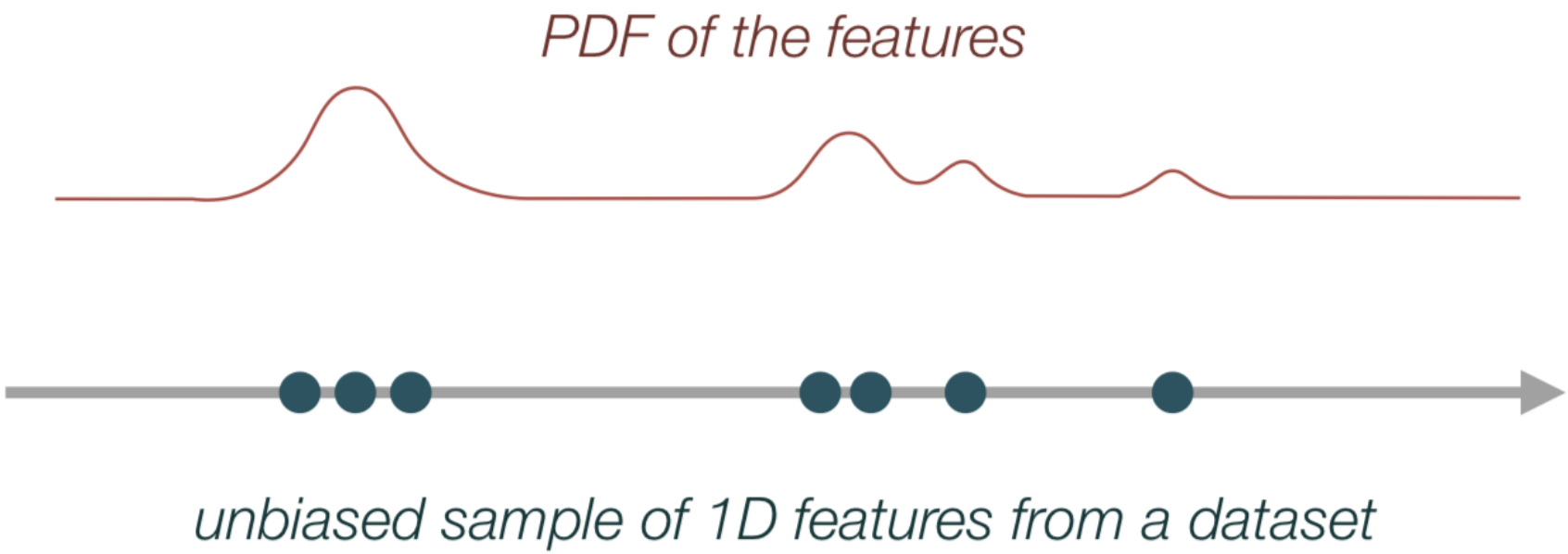
Discovering Groups - Mean Shift Clustering

Mean shift finds the *modes* of a probability density function.

This means if finds the points in feature space with the highest feature density, i.e. are the most likely given the dataset
Needs a kernel and a kernel bandwidth.

It is a hill climbing algorithm that

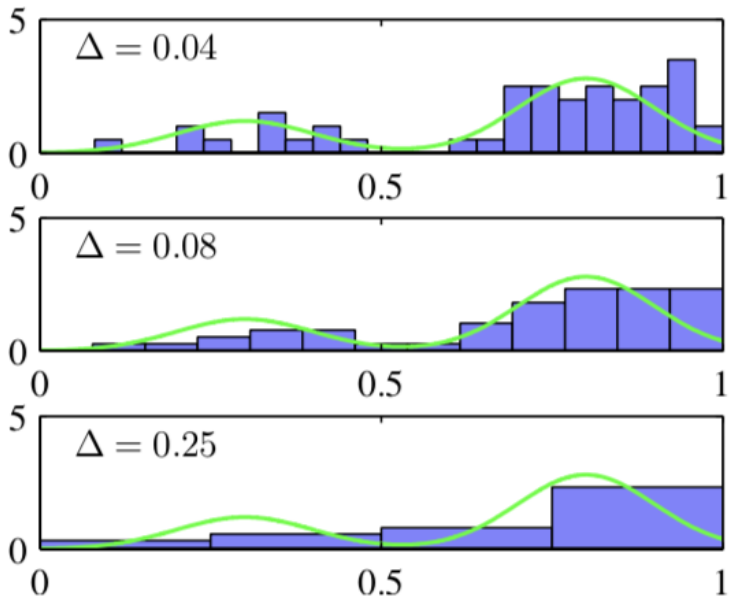
Discovering Groups - Mean Shift Clustering



Discovering Groups - Mean Shift Clustering

How can we estimate the PDF?

Could use a histogram, need to guess number of bins



Changing bin size affecting accuracy of probability density estimation[1]

Can be too crude

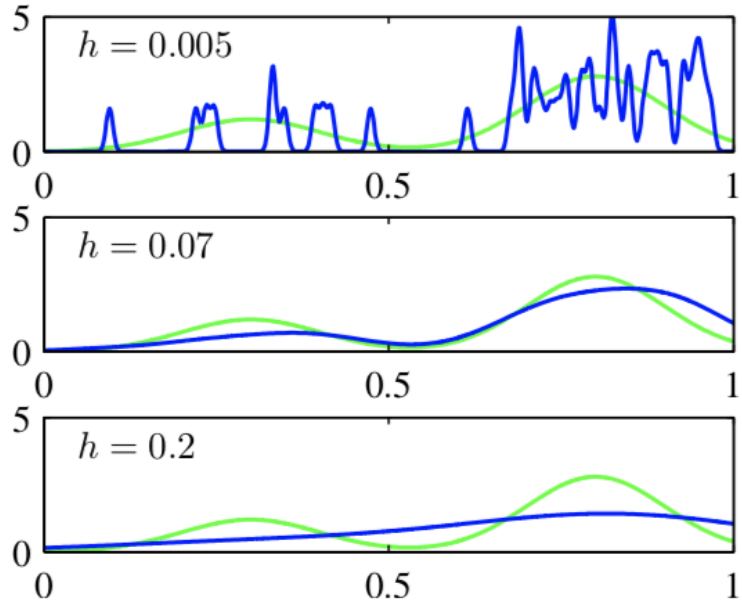
[1] C. Bishop, Pattern Recognition and Machine Learning

Discovering Groups - Mean Shift Clustering

Kernel Density Estimation (aka Parzen Window)
Gives a smooth continuous estimate

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

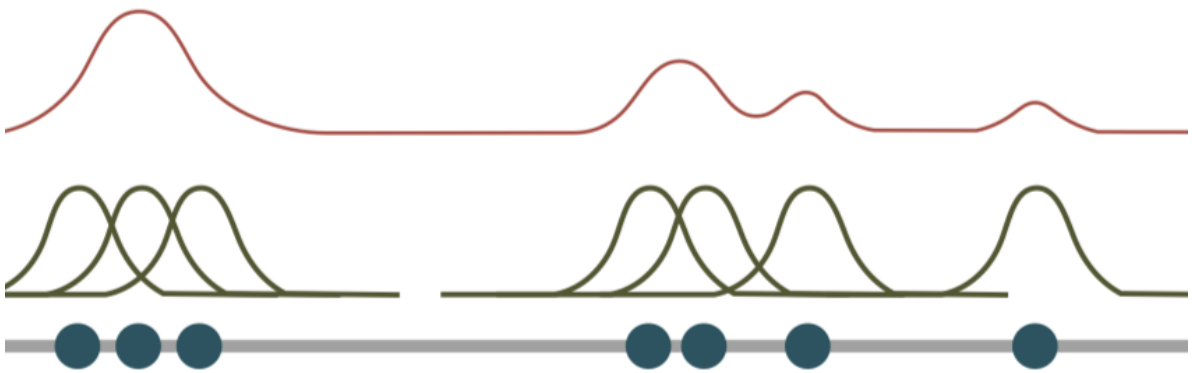
Where nh is the number of items, d is the dimensionality of the feature space, K is the kernel function, x is an arbitrary position in feature space, h is the kernel bandwidth



Changing bandwidth affecting accuracy of probability density estimation

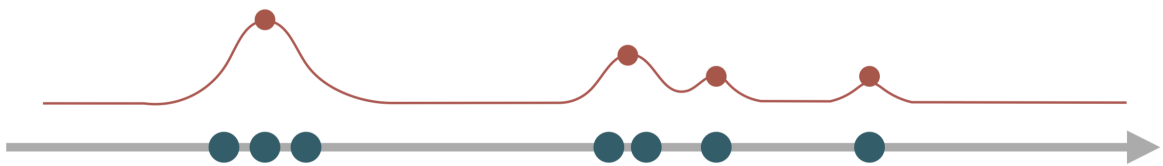
Discovering Groups - Mean Shift Clustering

Usually use a Gaussian kernel with $\sigma = 1$
If kernel is radially symmetric, then only need profile of kernel, $k(x)$ that satisfies $K(x) = C_{k,d}k(||x||^2)$



Discovering Groups - Mean Shift Clustering

FInd the modes of the probability density function (PDF), i.e. where the gradient is zero. $\Delta f(x) = 0$



Discovering Groups - Mean Shift Clustering

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

$$K(x) = c_{k,d} k(||x||^2)$$

Where $c_{k,d}$ is a normalisation constant

$$f(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k(||\frac{x - x_i}{h}||^2)$$

Assuming a radially symmetric kernel:

$$\Delta f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x - x_i) g(||\frac{x - x_i}{h}||^2) \quad g(x) = -k'(x)$$

$$\Delta f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} g(||\frac{x - x_i}{h}||^2) \frac{\sum_{i=1}^{n} x_i g(||\frac{x - x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x - x_i}{h}||^2)} - x$$

Discovering Groups - Mean Shift Clustering

$$\Delta f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} g\left(||\frac{x-x_i}{h}||^2\right) \frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x$$

The first part is a probability density estimate with kernel $G(x) = x_{g,d} g(||x||^2)$

Discovering Groups - Mean Shift Clustering

$$\Delta f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} g\left(||\frac{x - x_i}{h}||^2\right) \frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x$$

The first part is a probability density estimate with kernel $G(x) = x_{g,d} g(||x||^2)$

The second part is the mean shift, the vector that always points in the direction of maximum density

Discovering Groups - Mean Shift Clustering

Mean shift algorithm:

Algorithm 2: Mean Shift Procedure

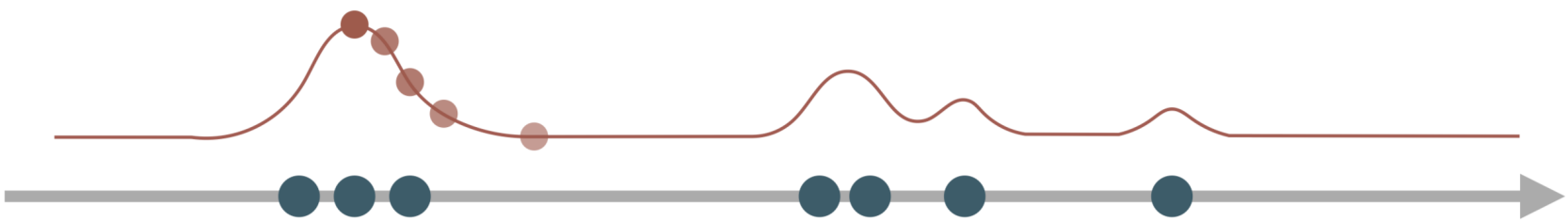
Data: N data points with feature vectors X_i $i = 1 \ldots N$

while $x_t not = x_{t+1}$ **do**

$m_h(x_t) =$ computeMeanShiftVect();

$x_{t+1} = x_t + m_h(x_t)$;

end

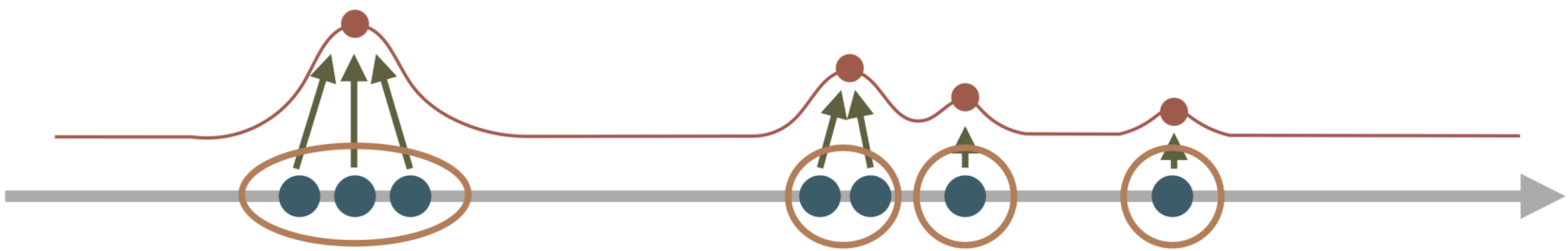


Discovering Groups - Mean Shift Clustering

For each feature vector:

- apply mean shift procedure until convergence
- store resultant mode

Set of feature vectors that converge to the same mode define the basin of attraction of that mode



Discovering Groups - Summary

Clustering is a key way to understand your data.

There are many different approaches

They are a very good way to start exploring a dataset