Unsupervised Learning

Martha Lewis
(slides created by Jonathan Lawry)

Department of Engineering Mathematics
University of Bristol

Introduction

- Aim: Given unlabelled data the aim is to partition the data set into distinct clusters of similar elements.
- Similarity: Often based on Euclidean distance or set distance.
- Challenges:
  - Knowing the optimal number of clusters.
  - Stability and convergence of algorithms.

Similarity and Distance

- The notions of similarity and difference are dual.
- A suitable measure of difference can easily be transformed into a corresponding measure of similarity.
- A distance metric on a set Ω is a function $d : \Omega \to \mathbb{R}^+$ such that:
  - Reflexive: $\forall x \in \Omega,\ d(x,x) = 0$.
  - Symmetric: $\forall x, y \in \Omega,\ d(x,y) = d(y,x)$.
  - Triangular Inequality: $\forall x, y, z \in \Omega,\ d(x,y) \leq d(x,z) + d(z,y)$.

Example Distance Metrics

- Euclidean distance: For $\Omega = \mathbb{R}^n$, $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$;

$$d(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

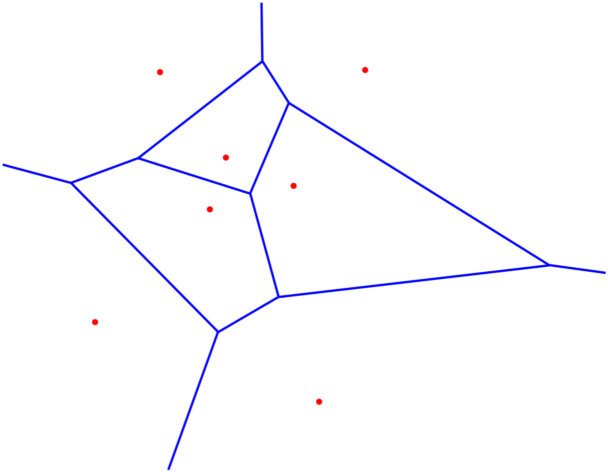
- Manhattan distance: For $\Omega = \mathbb{R}^n$, $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$;

$$d(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}|| = \sum_{i=1}^{n} |x_i - y_i|$$

- Hamming Distance: the distance between two strings of the same length is the number of positions in which they differ.

Voronoi Partitions

- A set of points naturally generates a partition on $\mathbb{R}^n$ called a Voronoi partition or tessellation.



k-means Clustering

- Perhaps the simplest clustering algorithm is k-means where k refers to the number of prototypes.
- Given a set of vectors drawn from $\Omega = \mathbb{R}^n$:
  1. Randomly partition the set of vectors into k sets.
  2. For each set P calculate its mean vector:

$$\hat{x}_P = \left(\frac{\sum_{\vec{x} \in P} x_1}{|P|}, \ldots, \frac{\sum_{\vec{x} \in P} x_i}{|P|} \ldots, \frac{\sum_{\vec{x} \in P} x_n}{|P|} \right)$$

  3. For each vector evaluate its Euclidean distance from each of the mean vectors e.g. $||\vec{x} - \hat{x}_P||$. Reallocate the vector to the partition set the mean of which it is closest to.
  4. If the partition sets remain unchanged then stop. Else go to 2.

Degree of Dissimilarity

- We can measure the degree of dissimilarity across the partition $\mathbf{P} = \{P_1, \ldots, P_k\}$ by:

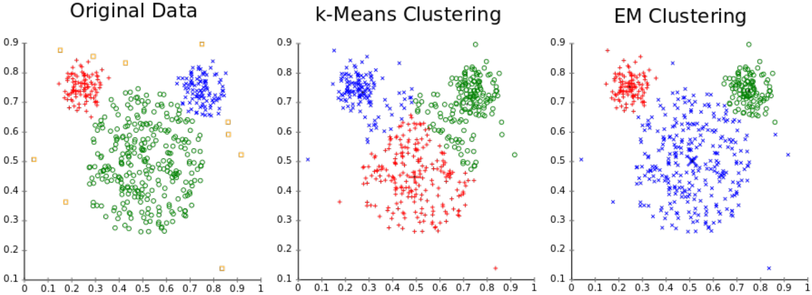
  $$J(\mathbf{P}) = \sum_{i=1}^{k} \sum_{\vec{x} \in P_i} ||\vec{x} - \hat{x}_{P_i}||^2 = \sum_{\vec{x} \in P_i} |P_i| Var(P_i)$$

  We call this the within-cluster sum of squares. Each of the inner terms is equal to the number of points in the cluster multiplied by the variance of the cluster.
- Since we intend that the mean vectors should serve as prototypes representing their corresponding partition set it is desirable that this value should be minimal.
- The k-means algorithm minimizes $J(\mathbf{P})$ terminating when its value stops decreasing.

Drawbacks of k-means

- Assumption that clusters are spherical and equally sized
- Number of clusters is an input parameter
- Can converge to local minima

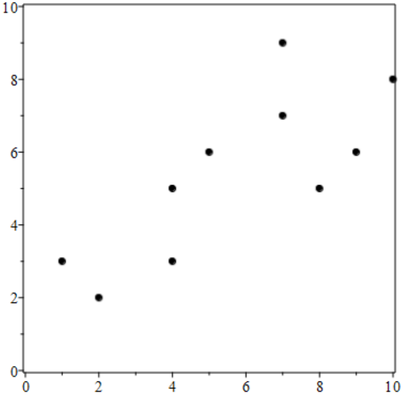
Different cluster analysis results on "mouse" data set:



Example: k-means

- Consider the following set of elements from $\mathbb{R}^2$;

$$\{(10,8),(7,9),(1,3),(2,2),(4,3),(8,5),(7,7),(5,6),(4,5),(9,6)\}$$

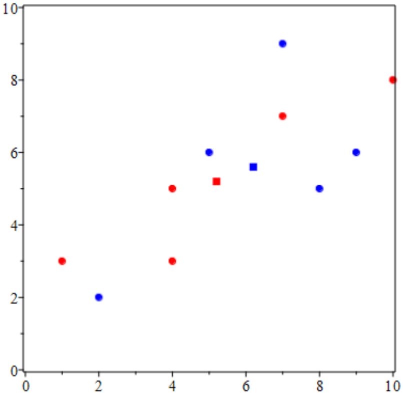


Example: k-means

- Let $k = 2$:
- Let the initial partition be:

$$P_1 = \{(10,8),(1,3),(4,3),(7,7),(4,5)\}$$
$$P_2 = \{(7,9),(2,2),(8,5),(5,6),(9,6)\}$$

- With means $\hat{x}_{P_1} = (5.2, 5.2)$ and $\hat{x}_{P_2} = (6.2, 5.6)$

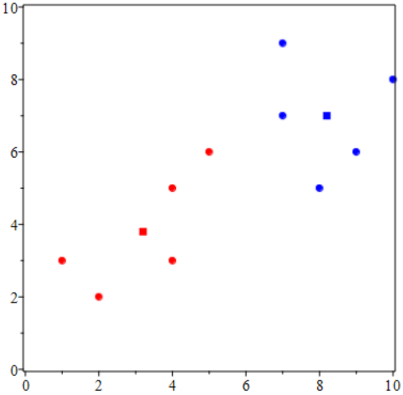


Example: k-means

- Reallocating points according to distance from means gives;

$$P_1 = \{(1,3),(2,2),(4,3),(5,6),(4,5)\}$$
$$P_2 = \{(10,8),(7,9),(8,5),(7,7),(9,6)\}$$

- Calculating new means and updating gives no change so terminate.

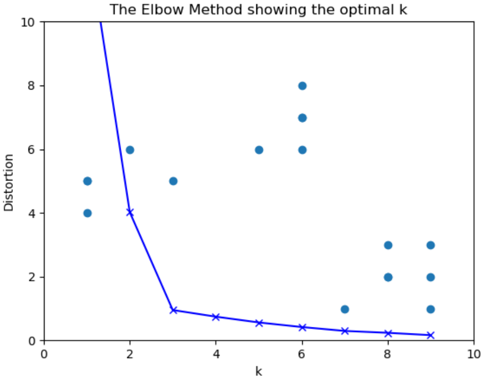


An alternative algorithm

- There are a number of different algorithms for k-means. Many work to assign the initial points so the the algorithm has less chance of getting stuck in a local optimum.
- A simple change is to allocate the first centroids by randomly choosing points in the dataset, and proceed from there.

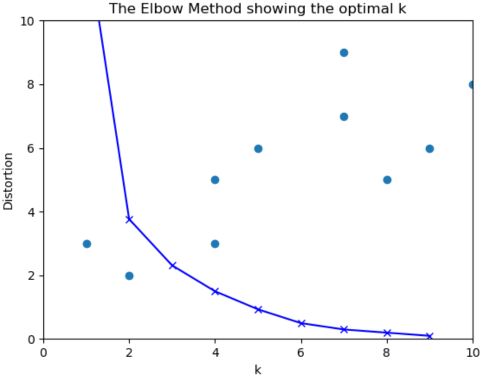
Elbow Plots

- Elbow plots help to identify the optimal value of k.
- Plot $J(\mathbf{P})$ against k
- If the plot looks like an arm, then the elbow on the arm is optimal k.



Elbow Plots Problems

- Sometimes it can be difficult to identify a clear elbow.
- In this case it is hard to identify the optimal k.

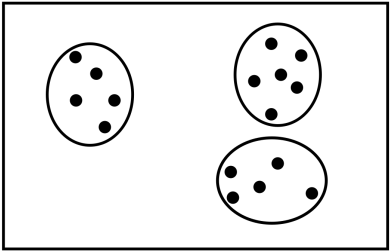


Summary: k-means

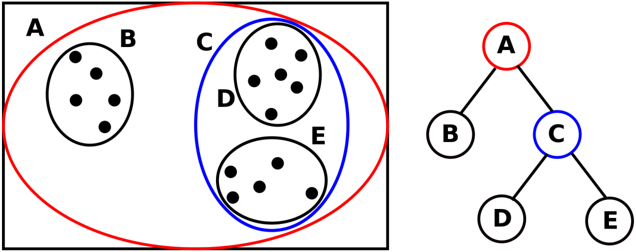
- k-means is a simple and intuitive algorithm to cluster data.
- k-means works by minimizing the variance of clusters of data that are determined by the cluster centroids.
- At each point in the algorithm, the cluster centroids are updated, and the datapoints reallocated to the clusters
- k-means suffers from a number of drawbacks: it assumes that the clusters are spherical and equally sized, it requires that we choose the number of clusters up front, and it can converge to local minima

Hierarchical Clustering

Flat Clustering



Hierarchical Clustering

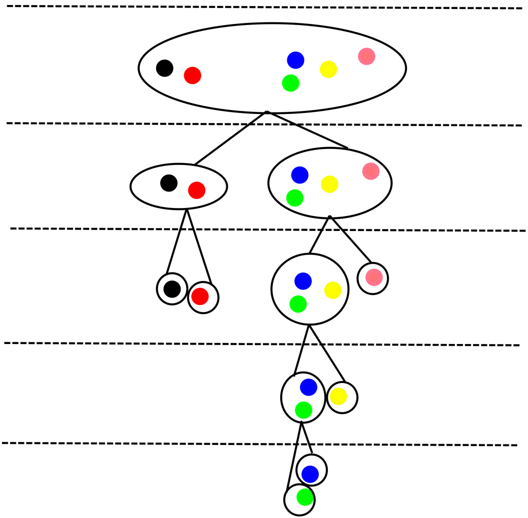


Types of Hierarchical Clustering

- *Top Down (Divisive) Clustering*: Begin with all data points in one cluster and then divide in child clusters.
- Recursively divide each child cluster and stop only when clusters contain a single point.
- *Bottom Up (Agglomerative) Clustering*: Clusters are built up from individual data points by merging.
- The most similar clusters are merged and the algorithm stops when all data points are merged into a single cluster.

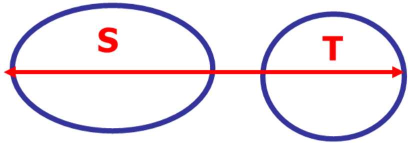
Hierarchical k-means

- For fixed k recursively run k-means on each child cluster until only single element clusters remain.



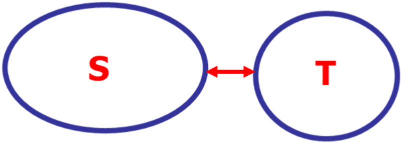
Distance Between Sets

max difference



$$d(S, T) = \max\{d(x, y) : x \in S, y \in T\}$$

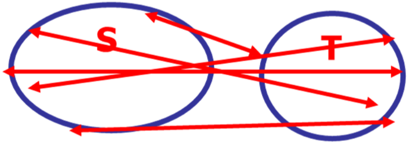
min difference



$$d(S, T) = \min\{d(x, y) : x \in S, y \in T\}$$

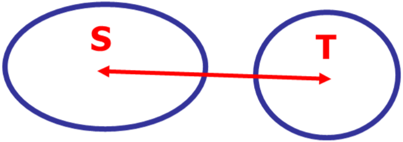
Distance Between Sets

Average difference



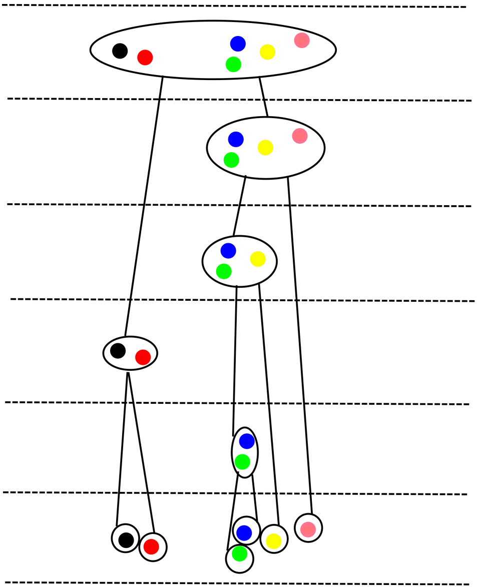
$$d(S,T) = \frac{1}{|S||T|} \sum_{x \in S} \sum_{y \in T} d(x,y)$$

Centroid difference

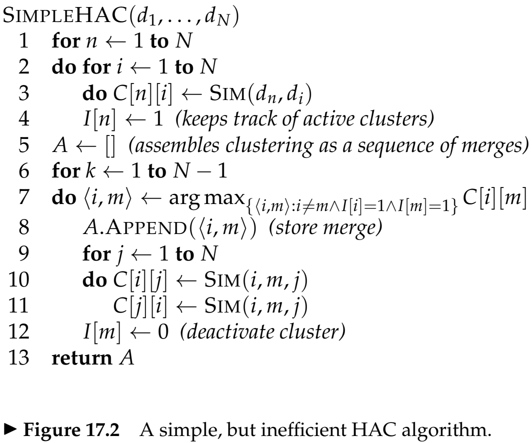


$$d(S,T) = d\left(\frac{\sum_{x \in S} x}{|S|}, \frac{\sum_{x \in T} x}{|T|}\right)$$

Agglomerative Clustering



Agglomerative Clustering



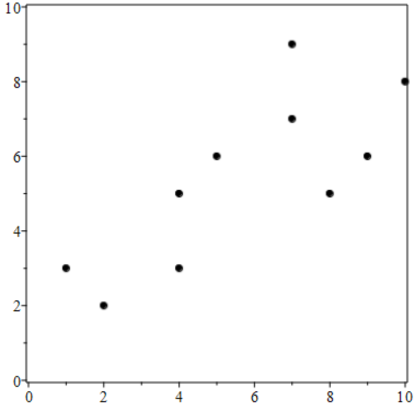
```
SIMPLEHAC($d_1, \ldots, d_N$)
for $n \leftarrow 1$ to $N$
do for $i \leftarrow 1$ to $N$
    do $C[n][i] \leftarrow$ SIM($d_n, d_i$)
    $I[n] \leftarrow 1$  (keeps track of active clusters)
$A \leftarrow []$  (assembles clustering as a sequence of merges)
for $k \leftarrow 1$ to $N - 1$
do $\langle i, m \rangle \leftarrow \arg\max_{\{\langle i,m \rangle : i \neq m \wedge I[i]=1 \wedge I[m]=1\}} C[i][m]$
    $A$.APPEND($\langle i, m \rangle$)  (store merge)
    for $j \leftarrow 1$ to $N$
    do $C[i][j] \leftarrow$ SIM($i, m, j$)
        $C[j][i] \leftarrow$ SIM($i, m, j$)
    $I[m] \leftarrow 0$  (deactivate cluster)
return $A$
```

► **Figure 17.2** A simple, but inefficient HAC algorithm.

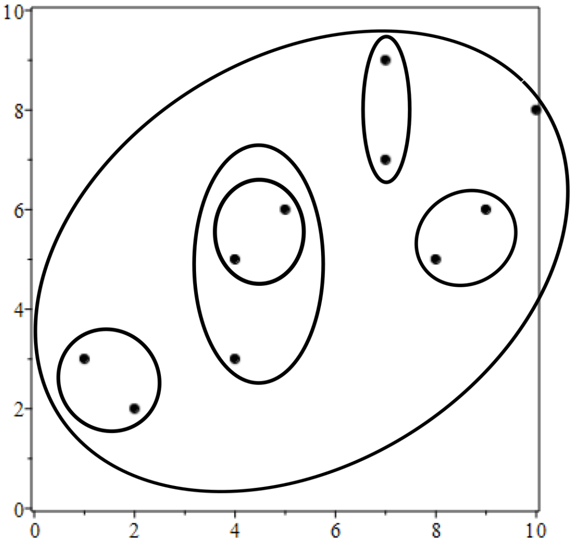
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. https://nlp.stanford.edu/IR-book/

Agglomerative Clustering Example

- Consider the set of data points $\{(10,8),(7,9),(1,3),(2,2),(4,3),(8,5),(7,7),(5,6),(4,5),(9,6)\}$

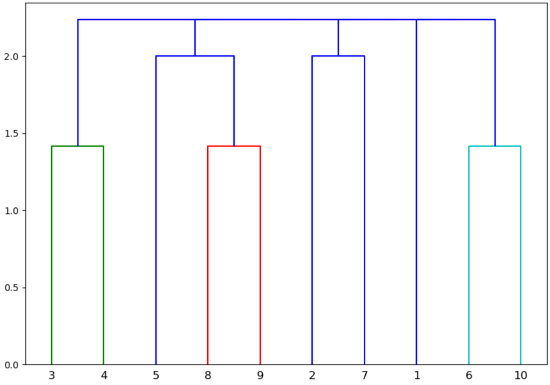


Agglomerative Clustering Example



Agglomerative Clustering Example

- The outcome of the clustering algorithm is represented in a dendrogram.
- The y-axis of the dendrogram indicates cluster similarity
- ‘Natural’ clusters of the data can be formed by cutting the dendrogram where the distance between clusters changes most.



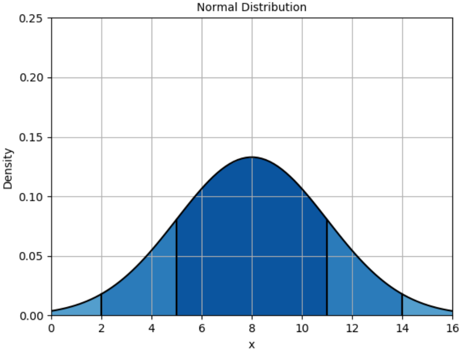
Summary

- Hierarchical clustering gives you a set of clusters that can be applied at different levels of hierarchy.
- Hierarchical clustering can be done top-down, or divisively, or bottom-up, using agglomeration.
- The results of the clustering can be visualized in a dendrogram.
- The dendrogram shows the order of clustering and distance between clusters. A ‘natural’ division of the data into clusters can be inferred by cutting the dendrogram where the distance between clusters is greatest.

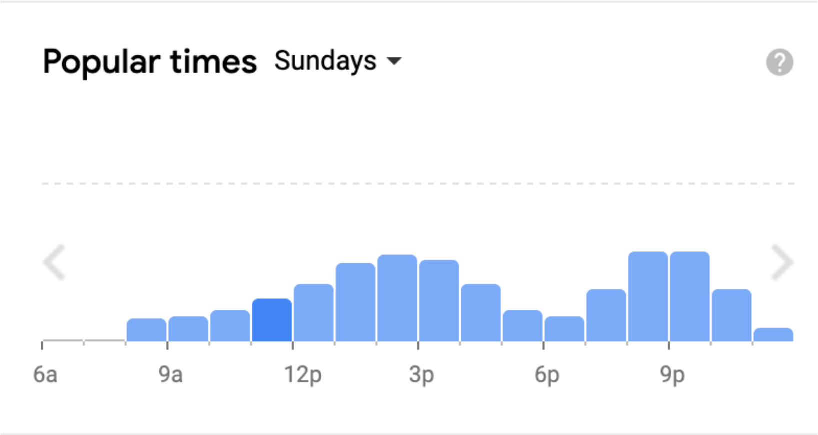
Gaussian (Normal) Distribution

- If $E(x) = \mu$ and $Var(x) = \sigma^2$ then;

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



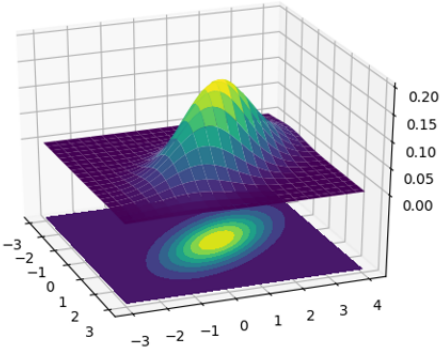
Multimodal distributions



Multivariate Gaussian Distribution

- Let $x_i : i = 1, \ldots, x_n$ be random variables where $\vec{x} = (x_1, \ldots, x_n)$, and let $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ where $E(x_i) = \mu_i$.
- Let Σ be the $n \times n$ covariance matrix such that $\Sigma_{i,j} = Cov(x_i, x_j) = E((x_i - \mu_i)(x_j - \mu_j))$.

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

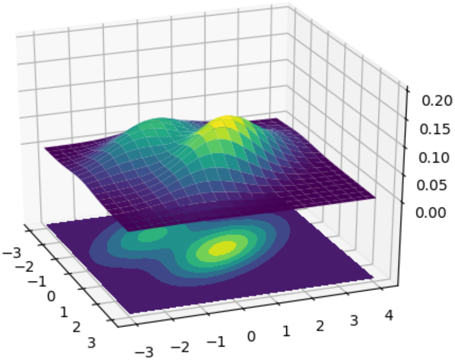


$$\vec{\mu} = (0, 1)$$

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

Gaussian Mixture Distributions

- A Gaussian mixture distribution has the form $f(\vec{x}) = \sum_{i=1}^{k} w_i f_i(\vec{x})$
- Each $f_i : i = 1, \ldots, k$ is a Gaussian distribution, $\sum_{i=1}^{k} w_i = 1$ and $w_i > 0$ for $i = 1, \ldots, k$.



Gaussian Mixture Clustering

- 1) Initialise algorithm by guessing $\vec{\mu}_i$ and Σ_i for $i = 1, \ldots, k$
- 2) Cluster membership: The membership of cluster j is $m_j(\vec{x}) = \frac{f_j(\vec{x}) w_j}{\sum_{i=1}^{k} f_i(\vec{x}) w_i}$
- 3) Compute new mean vectors, covariance matrices and weights according to;

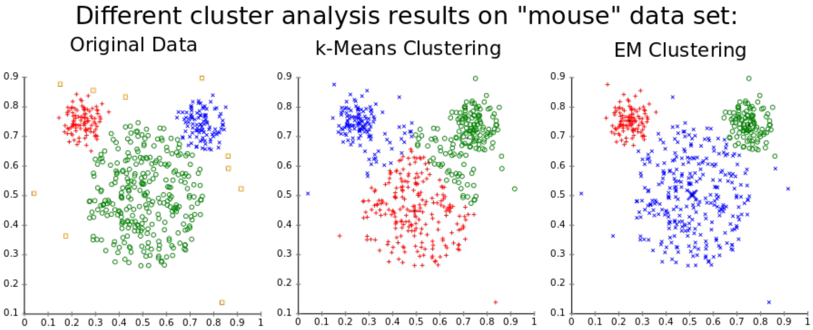
$$\vec{\mu}'_j = \frac{\sum_{\vec{x}} m_j(\vec{x}) \vec{x}}{\sum_{\vec{x}} m_j(\vec{x})}$$

$$\Sigma'_j = \frac{\sum_{\vec{x}} m_j(\vec{x}) (\vec{x} - \vec{\mu}'_j)(\vec{x} - \vec{\mu}'_j)^T}{\sum_{\vec{x}} m_j(\vec{x})}$$

$$w'_j = \frac{\sum_{\vec{x}} m_j(\vec{x})}{N}$$

- Repeat steps 2) and 3) until convergence.

Gaussian Mixture Clustering



Summary

- Gaussian mixture models can be used to cluster data in a probabilistic way
- We represent the data as a weighted sum of multivariate Gaussians
- The model parameters are learnt using an iterative algorithm, which is a kind of expectation maximisation algorithm.

Worksheet

- Covering k-means, hierarchical clustering, and Gaussian mixture models
- Interpretation of results and the effect of different distance metrics