TEXAS A&M UNIVERSITY

CSCE 633: Machine Learning

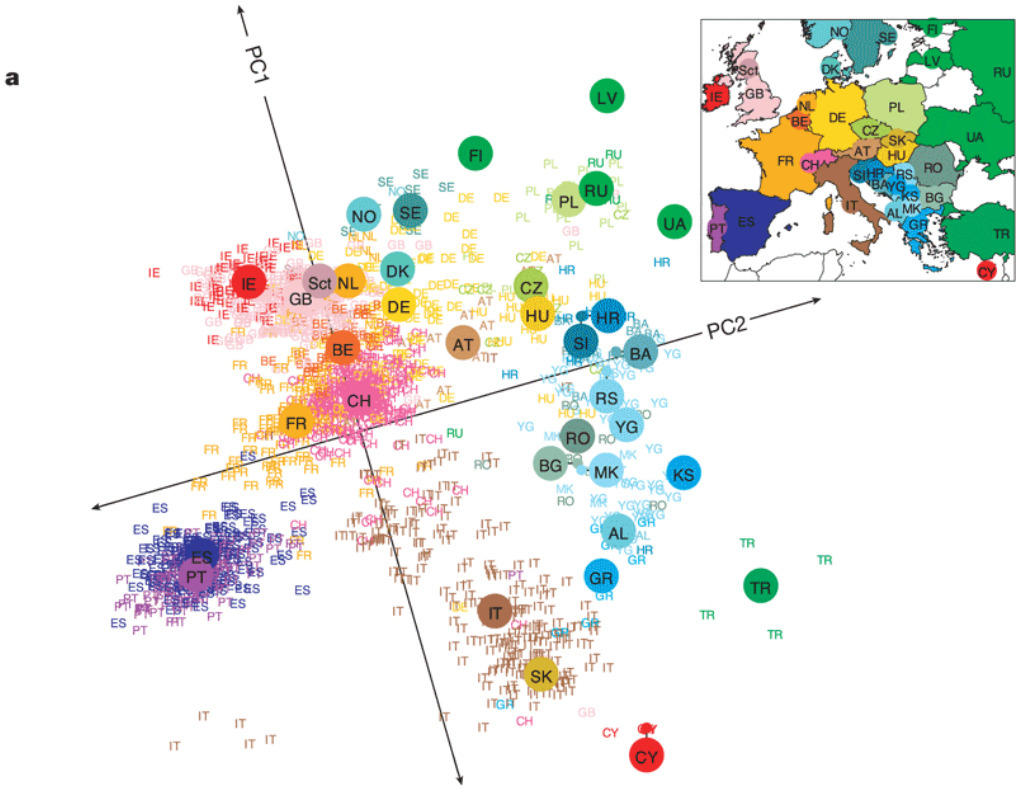
Lecture 11

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)
- Hierarchical clustering

Clustering

(1) Understanding: Finding patterns/structure/sub-populations in data



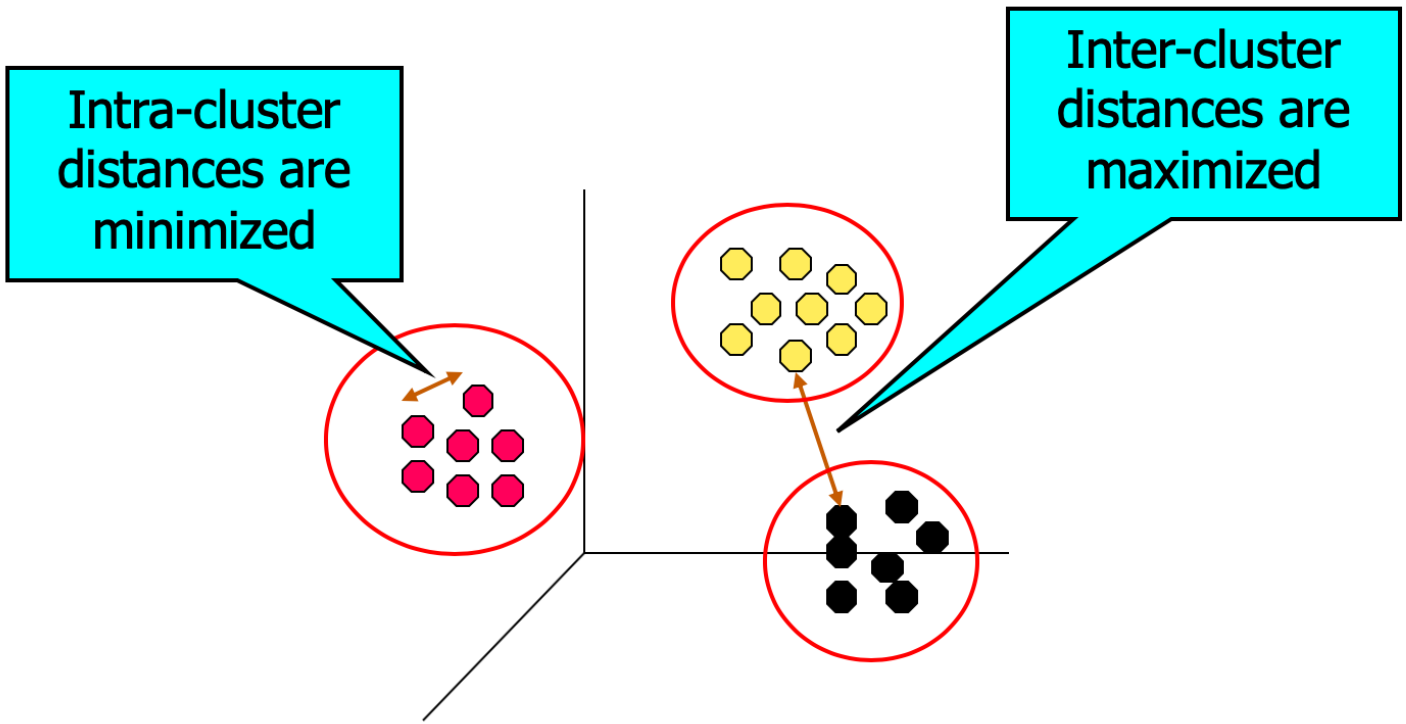
(2) Summarization: Reducing the size of large datasets

Clustering

- find patterns/structure/sub-populations in data ("knowledge discovery")
- training data does not include desired outputs
- less well-defined problem with no obvious error metrics
- topic modeling, market segmentation, clustering of hand-written digits, news clustering (e.g. Google news)

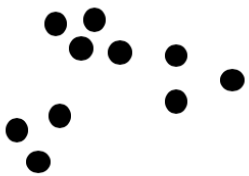
Clustering

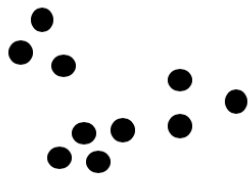
Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



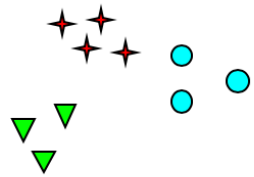
Clustering

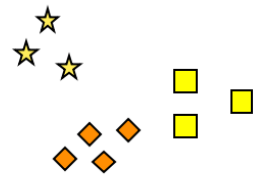
Notion of clustering can be ambiguous



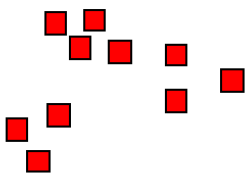


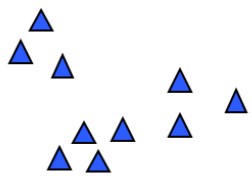
How many clusters?



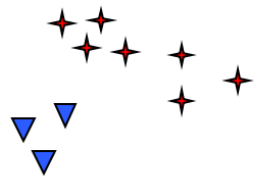


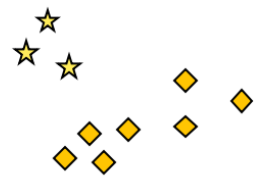
Six Clusters





Two Clusters





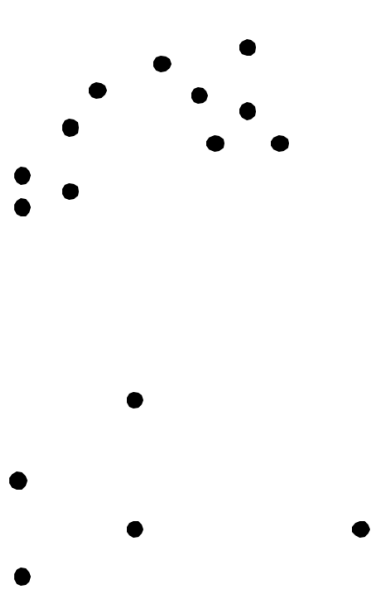
Four Clusters

Types of clustering

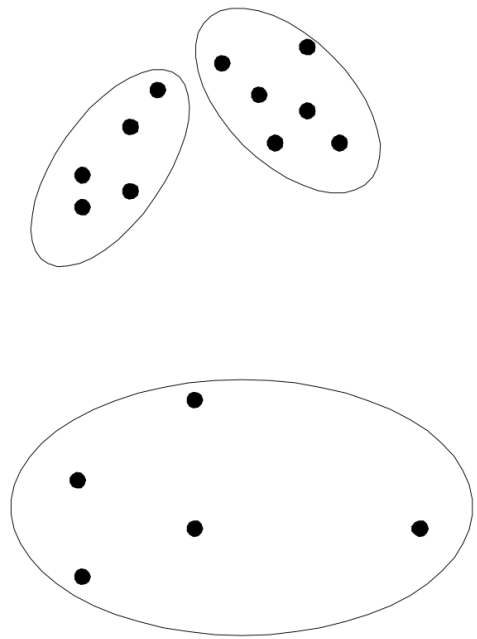
- Partitional clustering
  - non-hierarchical clusters
- Hierarchical clustering
  - a set of nested clusters organized as a hierarchical tree

Types of clustering

Partitional clustering



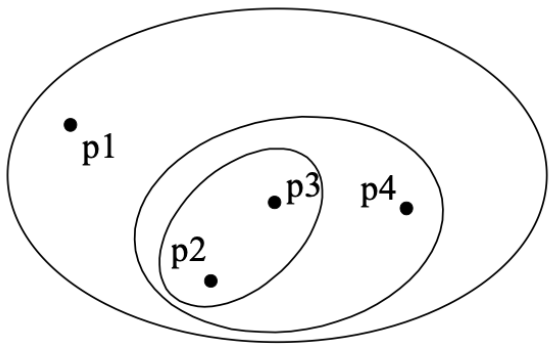
Original Points



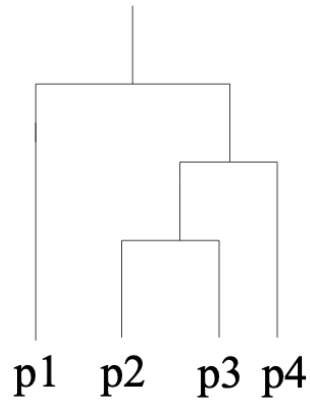
A Partitional Clustering

Types of clustering

Hierarchical clustering



Traditional Hierarchical Clustering



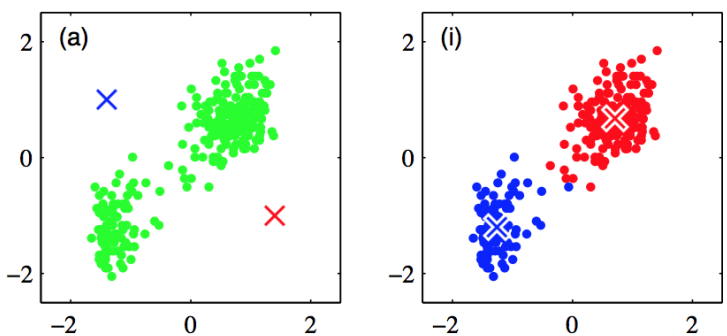
Traditional Dendrogram

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)
- Hierarchical clustering

K-means Clustering

- Input: Data $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$
- Output: Clusters $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}$
- Decision: Cluster membership, the cluster id assigned to sample $\mathbf{x_n}$, i.e. $A(\mathbf{x_n}) \in \{1, \ldots, K\}$
- Evaluation metric: Distortion measure
  $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$, where $r_{nk} = 1$ if $A(\mathbf{x_n}) = k$, 0 otherwise
- Intuition: Data points assigned to cluster k should be close to centroid $\boldsymbol{\mu_k}$



K-means Clustering

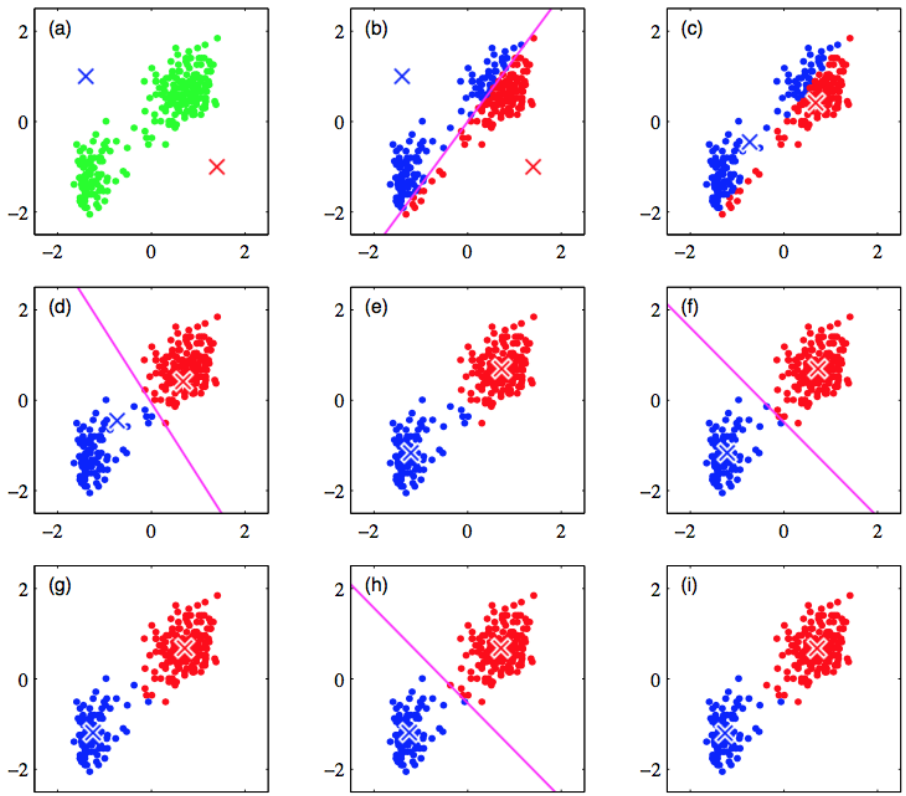
Evaluation metric: $\min_{r_{nk}} J = \min_{r_{nk}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$

Optimization:

- Step 0: Initialize $\boldsymbol{\mu_k}$ to some values
- Step 1: Assume the current value of $\boldsymbol{\mu_k}$ fixed, minimize J over r_{nk}, which leads to the following cluster assignment rule
  $r_{nk} = \begin{cases} 1, & \text{if } k = arg\min_j \|\mathbf{x_n} - \boldsymbol{\mu_j}\|_2^2 \\ 0, & \text{otherwise} \end{cases}$
- Step 2: Assume the current value of r_{nk} fixed, minimize J over $\boldsymbol{\mu_k}$, which leads to the following rule to update the prototypes of the clusters $\boldsymbol{\mu_k} = \frac{\sum_n r_{nk} \mathbf{x_n}}{\sum_n r_{nk}}$
- Step 3: Determine whether to stop or return to Step 1

K-means Clustering

Example



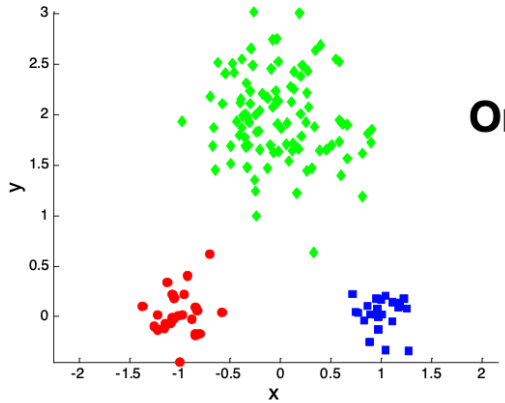
K-means Clustering

Remarks

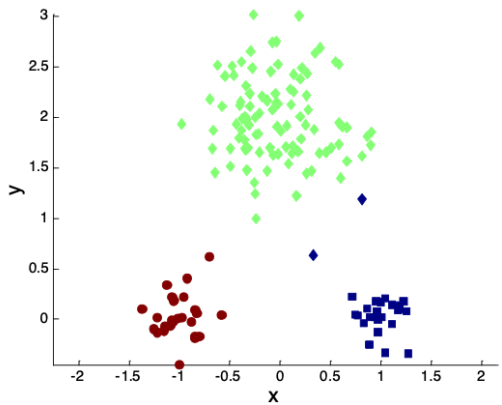
- The centroid $\boldsymbol{\mu_k}$ is the means of data points assigned to the cluster k, hence the name K-means clustering.
- The procedure terminates after a finite number of steps, as the procedure reduces J in both Step 1 and Step 2
- There is no guarantee the procedure terminates at the global optimum of J. In most cases, the algorithm stops at a local optimum, which depends on the initial values in Step 0 $\rightarrow$ random restarts to improve chances of getting closer to global optima

K-means Clustering

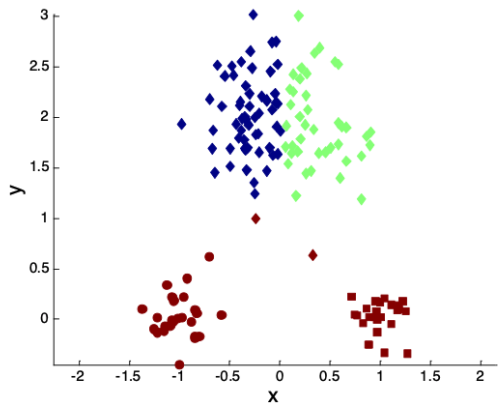
Initialization of K-Means is important



Original Points



Optimal Clustering



Sub-optimal Clustering

K-means Clustering

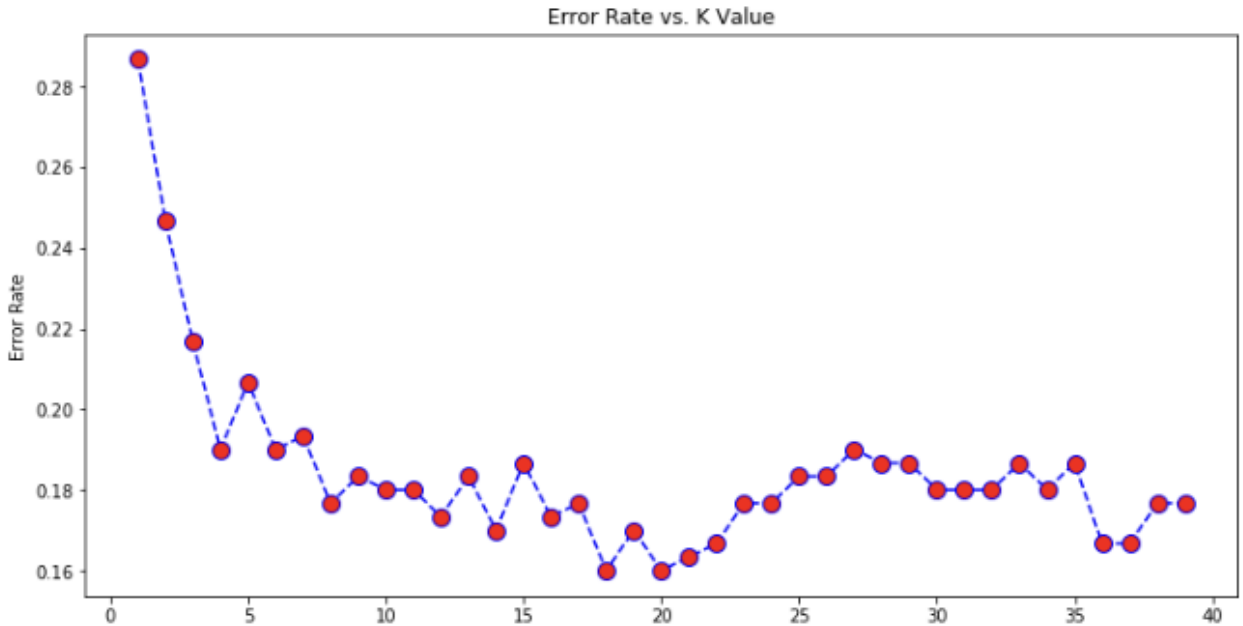
Solutions to Initial Centroids Problem

- Multiple random initializations
- Start with hierarchical clustering to determine initial centroids
- Select more than K initial centroids and then select among these initial centroids

K-means Clustering

How to know when to stop - Elbow Method

- Plot the error (i.e., distance of each sample to the corresponding centroid) against the number of clusters
- Stop when the decrease in error becomes almost flat



K-means Clustering

Application: vector quantization

- We can replace our data points with the centroids $\boldsymbol{\mu_k}$ from the clusters they are assigned to $\rightarrow$ vector quantization
- We have compressed the data points into
  - a codebook of all the centroids $\{\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}\}$
  - a list of indices to the codebook for the data points (created based on r_{nk})
- This compression is obviously lossy as certain information will be lost if we use a very small K

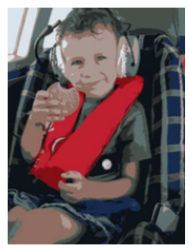
K-means Clustering

Question: vector quantization with K-means

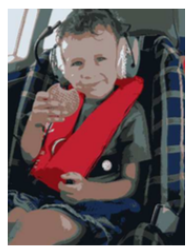
Assume that the images bellow are created by vectoring the original image with K-means using different values of K. What is the correct combination?

Original Image

A) $K = 25$ $K = 10$ $K = 3$



B) $K = 3$ $K = 10$ $K = 25$



Correct answer is A of course :)

K-means Clustering

Limitations of K-Means

- Problems when clusters are of differing size, density, or non-spherical shapes (for Euclidean distances)
- Sensitive to outliers
- Number of clusters is difficult to determine

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)
- Hierarchical clustering

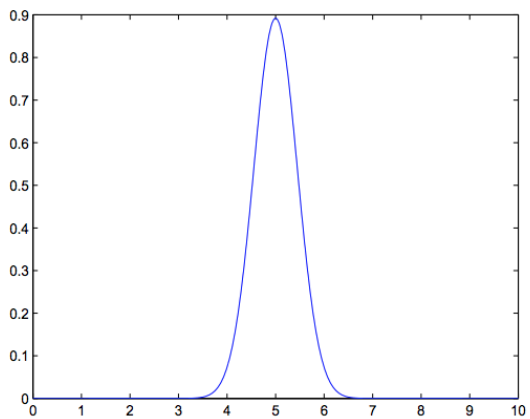
Multivariate Gaussian distribution

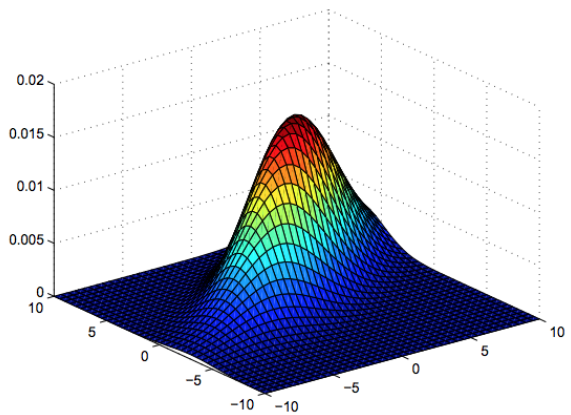
Univariate Gaussian distribution

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Multivariate Gaussian distribution

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$





Multivariate Gaussian distribution

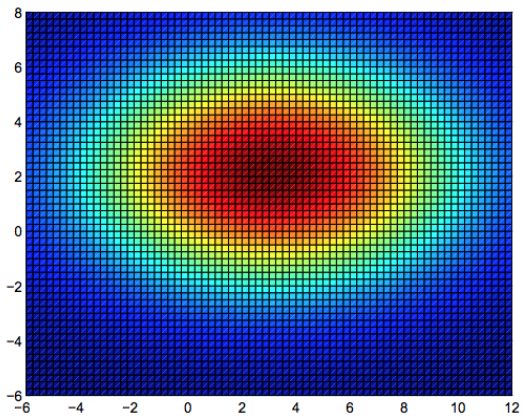
Covariance matrix

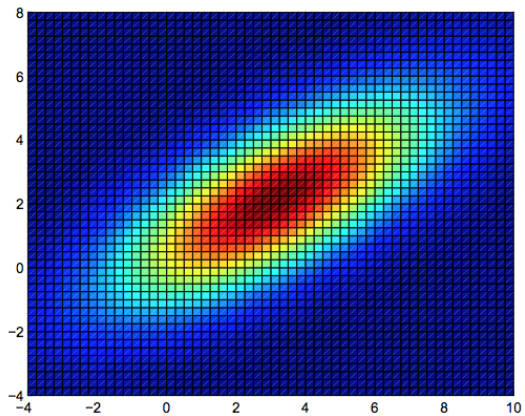
- Covariance between two random variables X and Y
  $Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- The covariance matrix provides a way to summarize the covariances of all pairs of variables $(\boldsymbol{\Sigma})_{ij} = Cov(X_i, X_j)$
- $\boldsymbol{\Sigma}$ is always positive definite

Multivariate Gaussian distribution

Isocontours

- For a function $f : \mathbb{R}^2 \to \mathbb{R}$ an isocontour is a set of the form $\{\mathbf{x} \in \mathbb{R}^2 : f(x) = c\}$

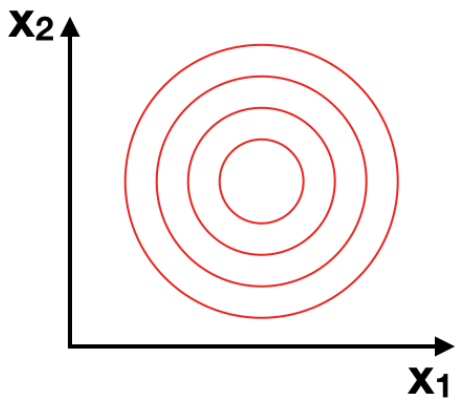




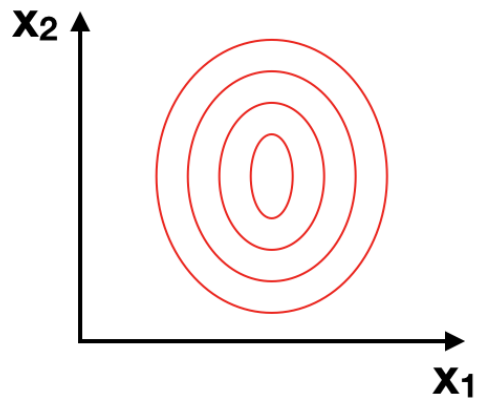
Multivariate Gaussian distribution

The diagonal covariance case

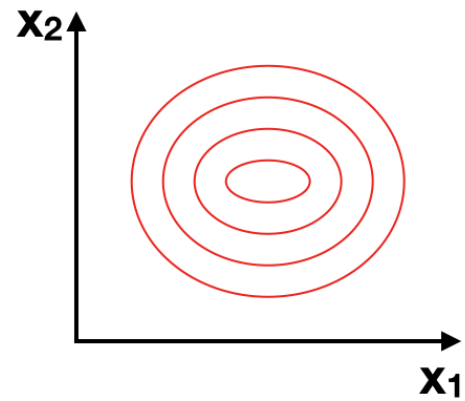
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



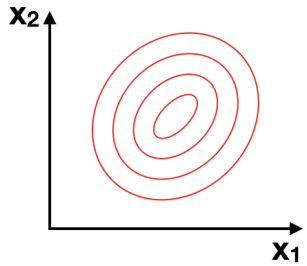
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

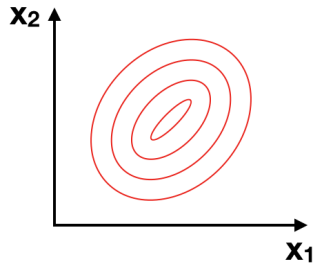


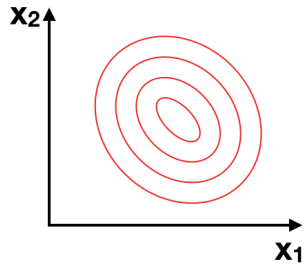
$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

Multivariate Gaussian distribution

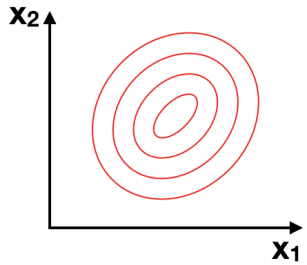
Question: Which is correct in this non-diagonal covariance case?

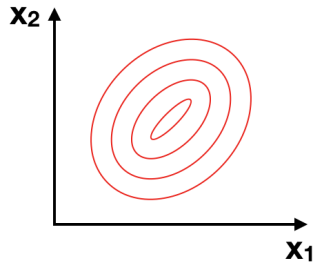


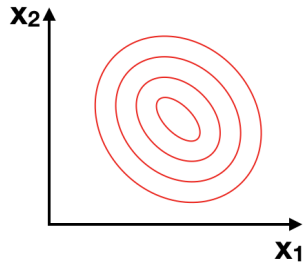




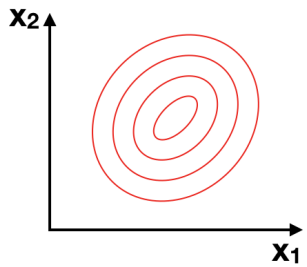
A) $\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

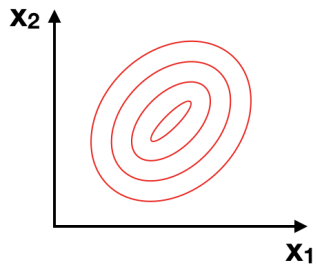


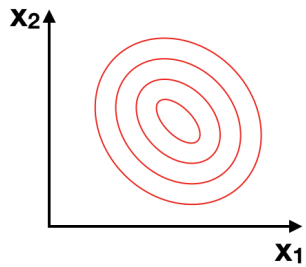




B) $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$







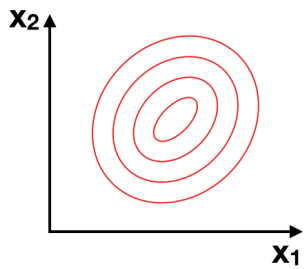
C) $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

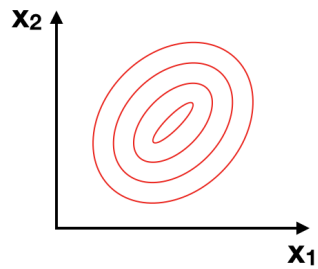
Multivariate Gaussian distribution

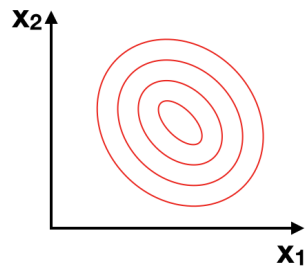
Question: Which is correct in this non-diagonal covariance case?

Correct answer is C

By increasing the off-diagonal elements from 0.5 to 0.8, the distribution is more thinly peaked along the line where x_1 is equal to x_2



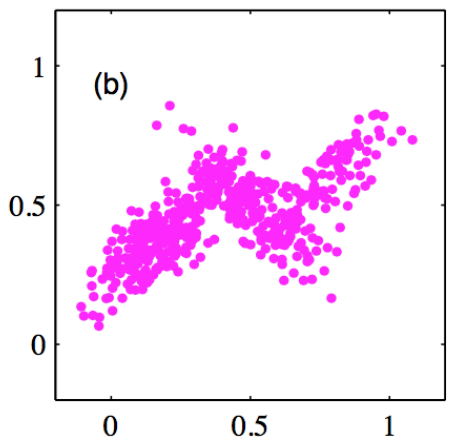




c) $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$ $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

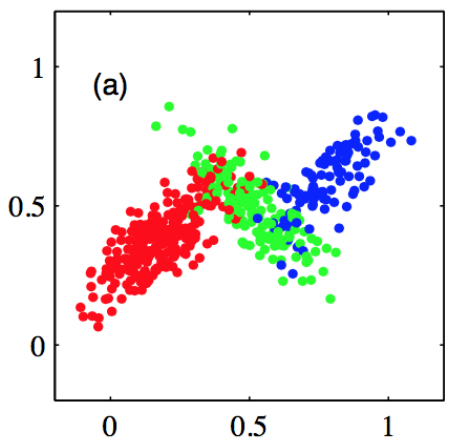
Probabilistic interpretation of clustering

- We want to find $p(\mathbf{x})$ that best describes our data
- The data points seem to form 3 clusters
- However, we cannot model $p(\mathbf{x})$ with simple and known distributions, e.g. one Gaussian



Probabilistic interpretation of clustering

- Instead, we will model each region with a Gaussian distribution $\rightarrow$ Gaussian mixture models (GMMs)
- Question 1: How do we know which (color) region a data point comes from?
- Question 2: What are the parameters of Gaussian distributions in each region?
- We will answer both in an unsupervised way from data $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$



GMM as the marginal distribution of a joint distribution

- The joint distribution between $\mathbf{x}$ and z (representing color) are

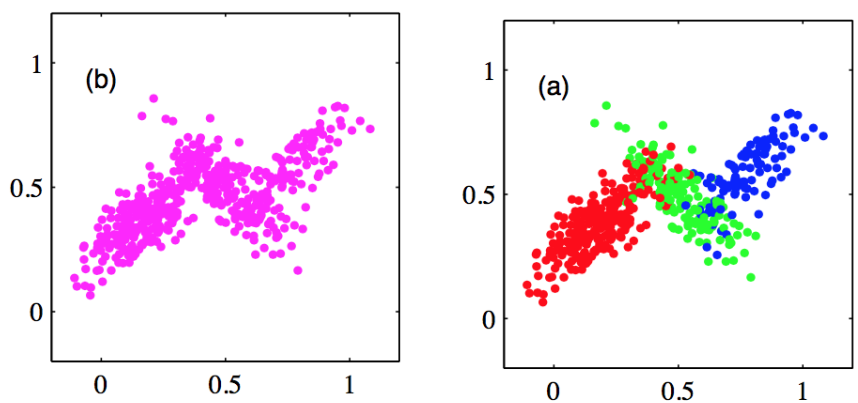
$$p(\mathbf{x}|z = red) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$$

$$p(\mathbf{x}|z = blue) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$$

$$p(\mathbf{x}|z = green) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_3}, \boldsymbol{\Sigma_3})$$

- The marginal distribution is thus

$$\begin{aligned} p(\mathbf{x}) =& p(red)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) + p(blue)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}) \\ &+ p(green)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu_3}, \boldsymbol{\Sigma_3}) \end{aligned}$$



Gaussian mixture models

A Gaussian mixture model has the following density function for $\mathbf{x}$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

- K: number of Gaussians
- $\boldsymbol{\mu_k}$,$\boldsymbol{\Sigma_k}$: mean & covariance of k^{th} component
- π_k: component weights

$$\pi_k > 0\ ,\quad \forall k \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1$$

- Estimate $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$, $\pi_k \Rightarrow$ Expectation Maximization

Parameter estimation for GMMs

If we know the probability of sample $\mathbf{x_n}$ belonging to Gaussian component k, i.e., responsibility $\gamma(z_{nk})$, we can estimate the parameters of each Gaussian distribution $\{\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}, \pi_k\}$ (Maximization Step)

$$\pi_k = \frac{\sum_n \gamma(z_{nk})}{\sum_k \sum_n \gamma(z_{nk})} \qquad \boldsymbol{\mu_k} = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk}) \mathbf{x_n}$$

$$\boldsymbol{\Sigma_k} = \frac{1}{\sum_n \gamma(z_{nk})} \sum_n \gamma(z_{nk})(\mathbf{x_n} - \boldsymbol{\mu_k})(\mathbf{x_n} - \boldsymbol{\mu_k})^T$$

- For π_k: count the number of data points whose z_n is k and divide by the total number of data points
- For μ_k: the mean of all samples weighted by their responsibility (i.e., probability of belonging to mixture k)
- For Σ_k: the covariance matrix of all samples weighted by their responsibility (i.e., probability of belonging to mixture k)

Parameter estimation for GMMs: incomplete data

If we know the parameters of each Gaussian mixture $\{\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}, \pi_k\}$, we can find the probability of each data sample $\mathbf{x_n}$ belonging to Gaussian mixture k (Expectation Step)

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})}$$

Every data point $\mathbf{x_n}$ is assigned to a component fractionally according to $\gamma(z_{nk})$, also called responsibility

Parameter estimation for GMMs

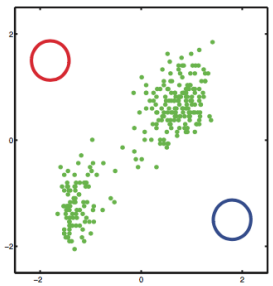
Since we do not know $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ to begin with, we cannot compute $\gamma(z_{nk})$ or π_k

But we can invoke an iterative procedure and alternate between estimating γ_{nk} using π_k, $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$, and vice-versa.

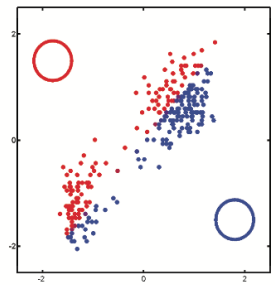
- Step 0: Guess π_k, $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ with initial values
- Step 1 (E-Step): Compute γ_{nk} using current π_k, $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$
- Step 2 (M-Step): Update π_k, $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ using computed γ_{nk}
- Step 3: Go back to Step 1

Parameter estimation for GMMs

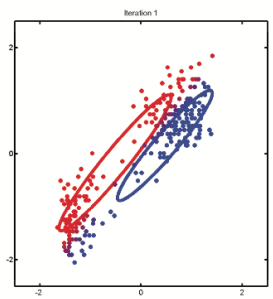
Example of GMM parameter estimation with EM



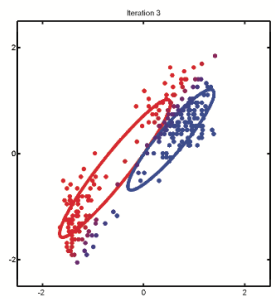
(a)



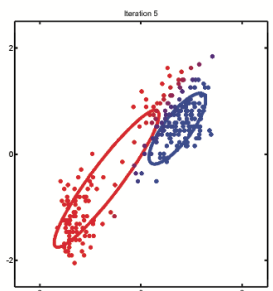
(b)

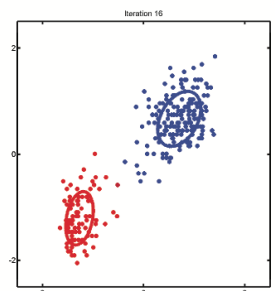


(c)



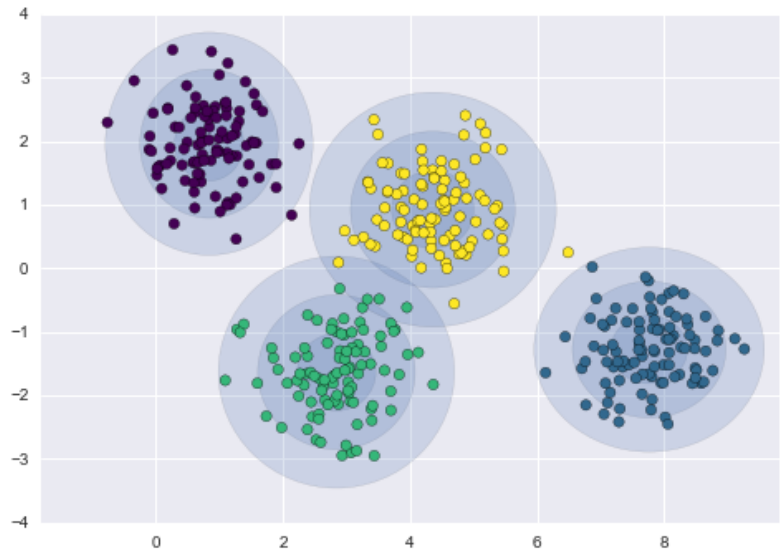
(d)





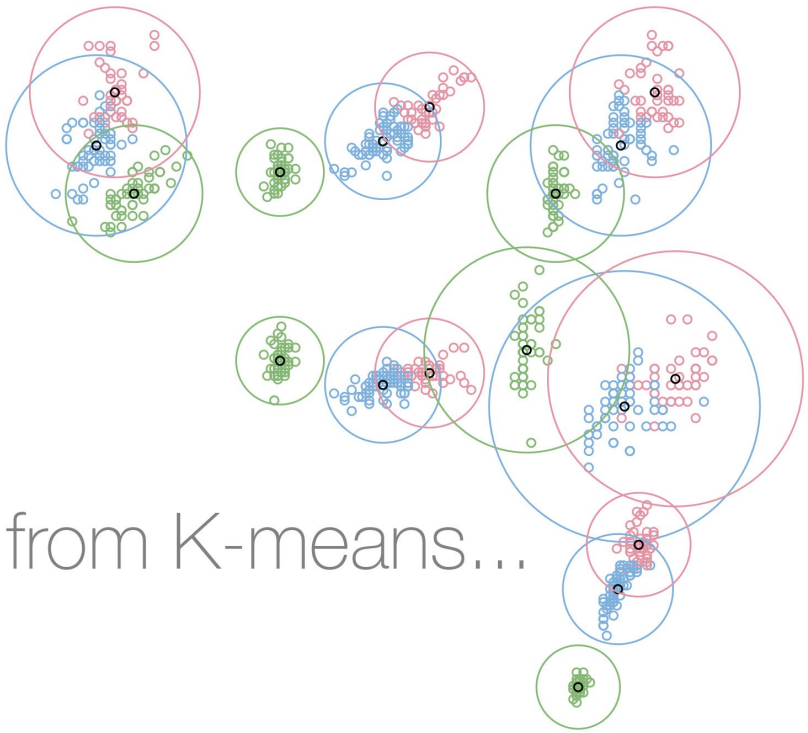
Gaussian Mixture Models

Example of GMM parameter estimation with EM



Gaussian Mixture Models

Comparison between K-Means and GMMs

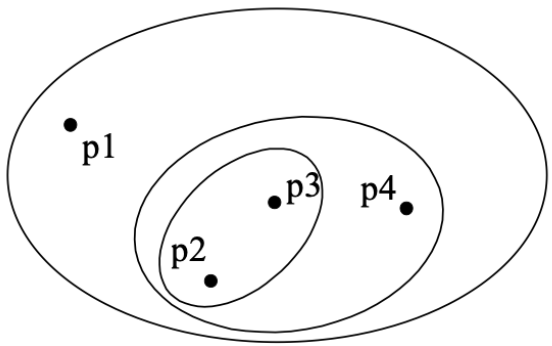


Overview

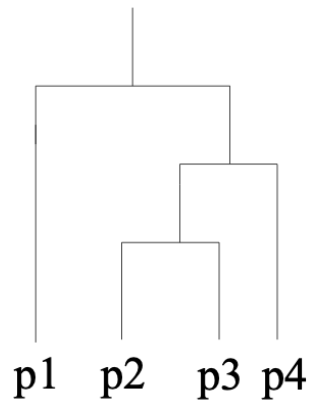
- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)
- Hierarchical clustering

Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits



Traditional Hierarchical Clustering



Traditional Dendrogram

Hierarchical clustering

Advantages of hierarchical clustering

- Do not have to pre-determine number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level
- Resulting clusters may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction)

Hierarchical clustering

Types of hierarchical clustering

- Agglomerative
  - Start with each sample as individual cluster
  - Merge the closest pair of clusters each time until only one cluster left
- Divisive
  - Start with one, all-inclusive cluster
  - Split a cluster each time until each cluster contains a point

Agglomerative hierarchical clustering

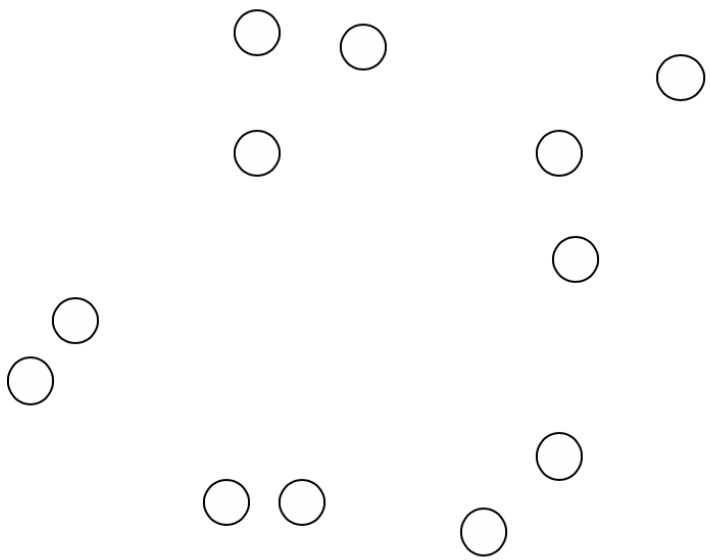
- Step 0: Compute the proximity matrix
- Step 1: Let each data sample be a cluster
- Step 2: **Repeat:**
  - Merge the two closest clusters
  - Update the proximity matrix

  Until only a single cluster remains

Key operation is the computation of the proximity of two clusters $\rightarrow$ different approaches for defining distance between clusters

Agglomerative hierarchical clustering

Initialization: Start with each sample being a cluster

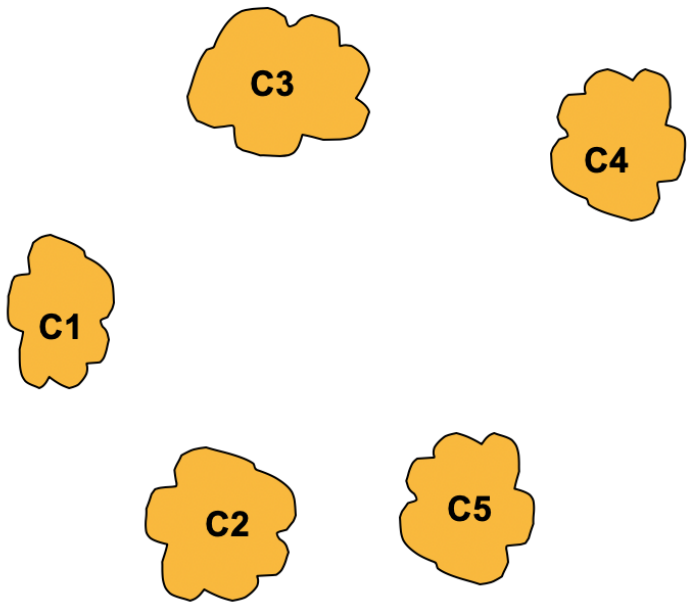


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

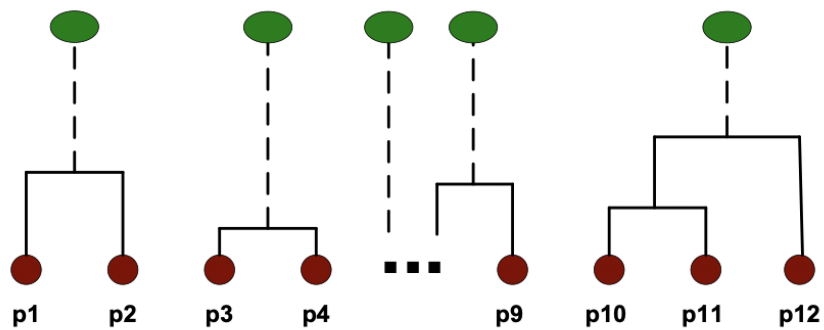
Agglomerative hierarchical clustering

After some steps: we have some clusters



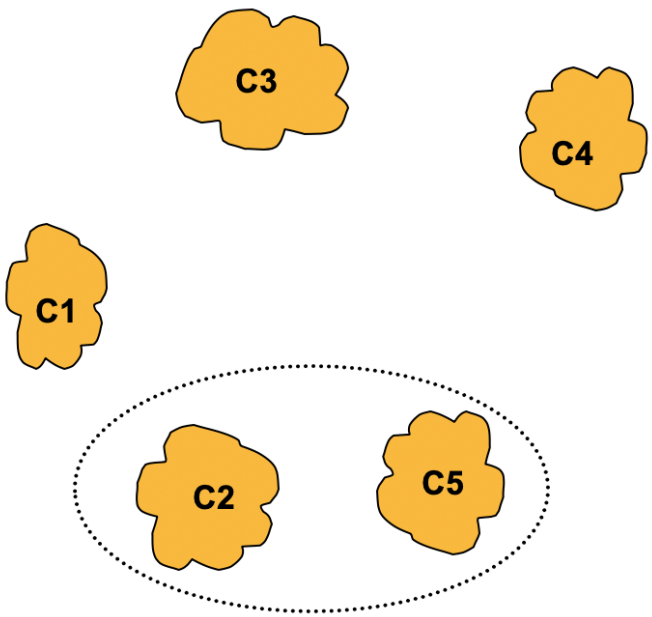
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

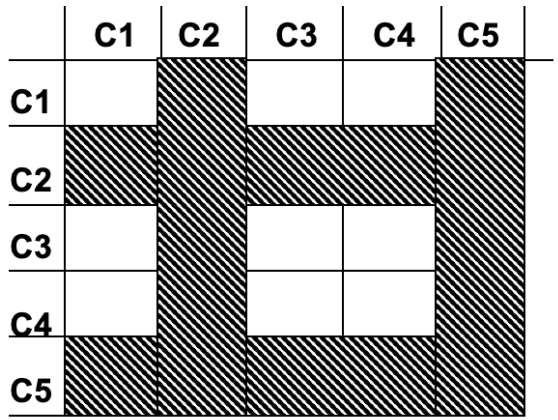
Proximity Matrix



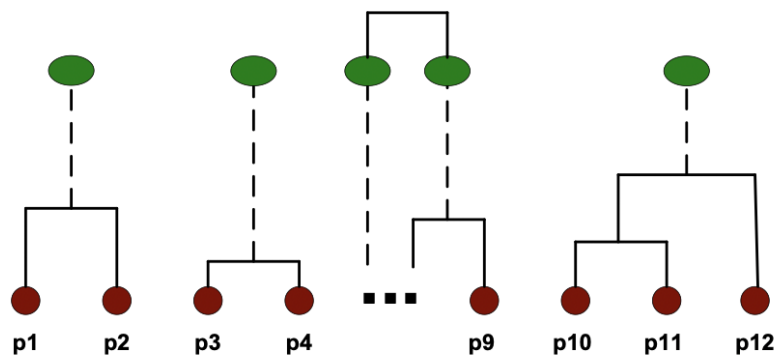
Agglomerative hierarchical clustering

We want to merge the two closest clusters (C2 and C5) and update the proximity matrix



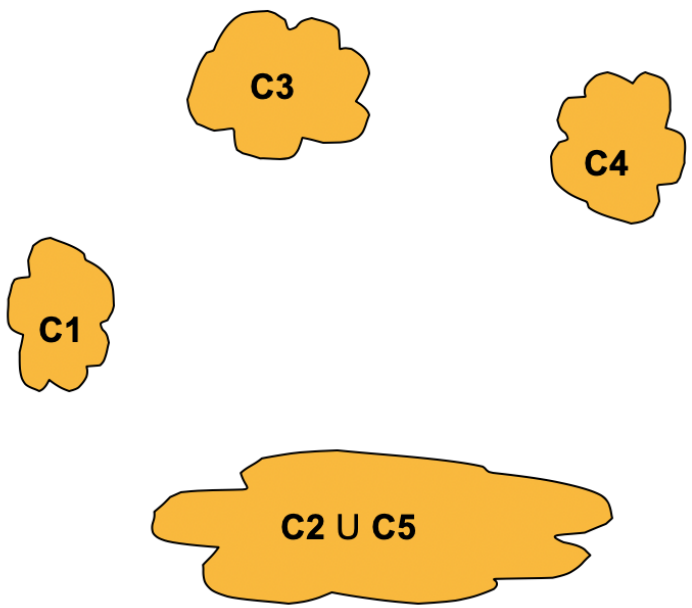


Proximity Matrix



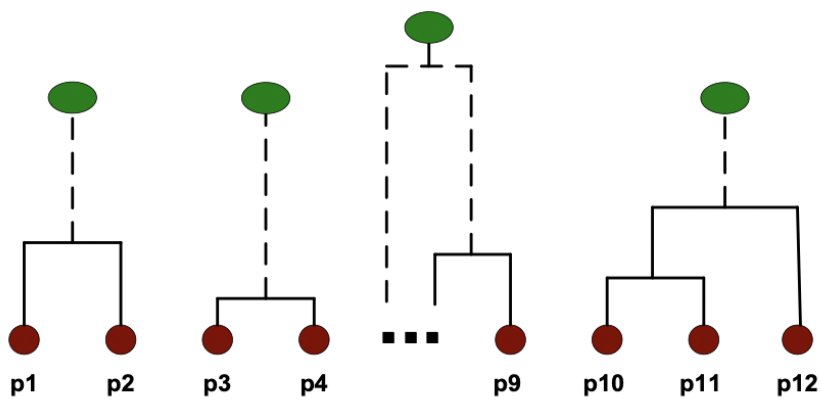
Agglomerative hierarchical clustering

How do we update the proximity matrix?



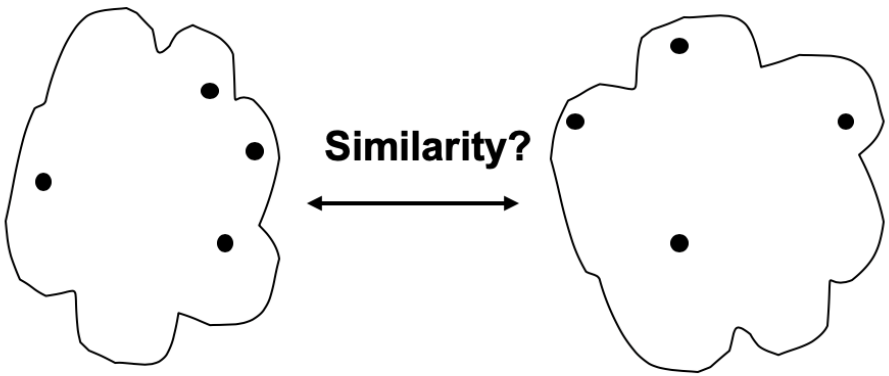
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Agglomerative hierarchical clustering

How to define inter-cluster similarity?

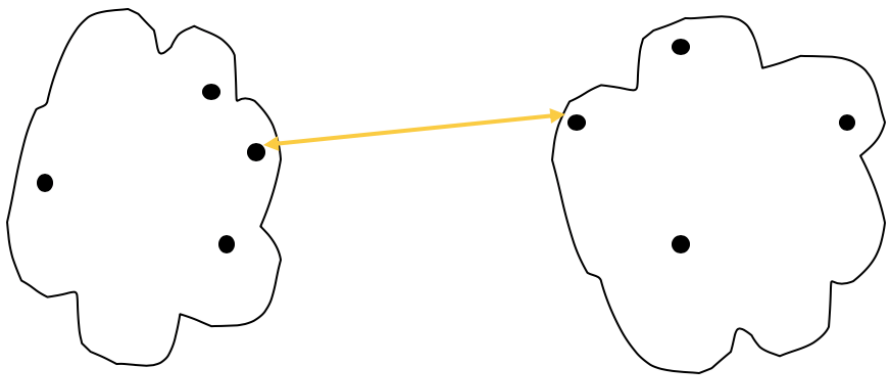


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

- min, max, group average, distance between centroids

Agglomerative hierarchical clustering

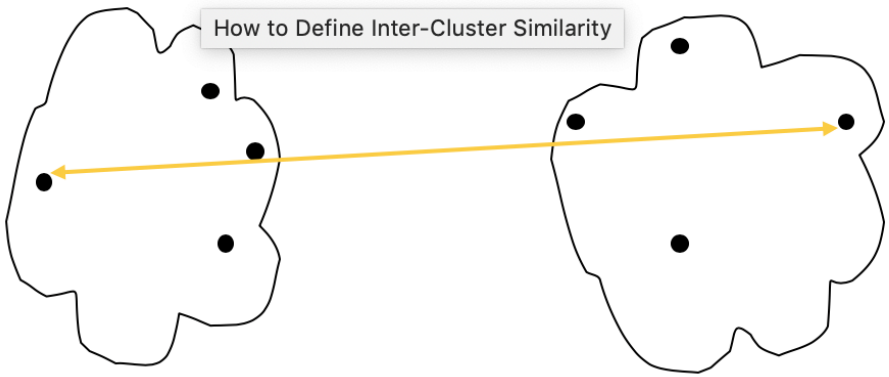
Distance between the closest samples (min)



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

Agglomerative hierarchical clustering

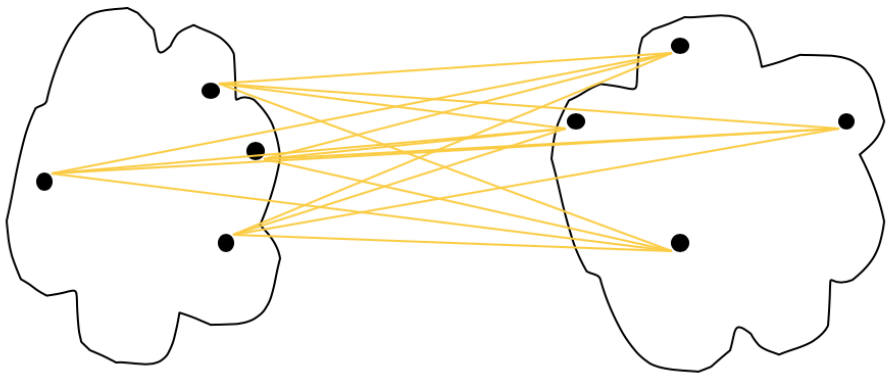
Distance between the furthest samples (max)



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

Agglomerative hierarchical clustering

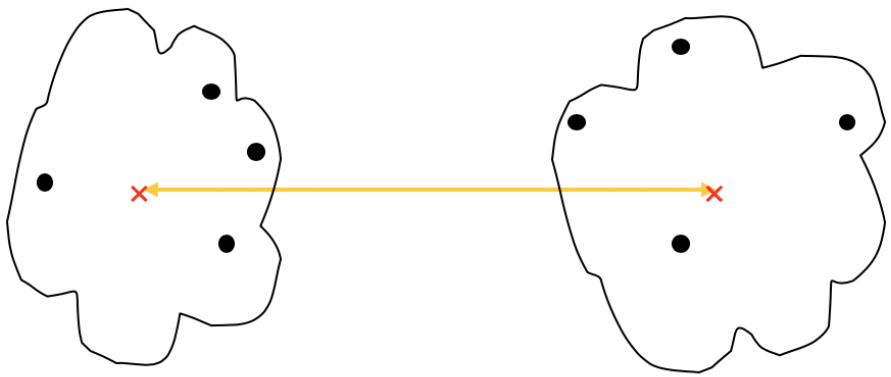
Average pairwise distance between samples (group average)



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						

Agglomerative hierarchical clustering

Distance between centroids



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

Overview

- Clustering tries to find patterns/hidden structures in data
- Partitional clustering
  - K-means: hard assignment of samples to one centroid
  - GMMs: soft assignment of samples to each Gaussian
- Hierarchical clustering: nested clusters organized as a hierarchical tree
- **Readings:** Alpaydin 7; Pang-Ning Tan 7 (uploaded on Piazza)