Lecture 10: Clustering

Andreas Wichert
Department of Computer Science and Engineering
Técnico Lisboa

1

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

2

- We approach first using a nonprobabilistic technique called the K-means algorithm (Lloyd, 1982).
- Then we introduce the latent variable view of mixture distributions in which the discrete latent variables can be interpreted as defining assignments of data points to specific components of the mixture
- A general technique for finding maximum likelihood estimators in latent variable models is the expectation-maximization (EM) algorithm.

3

- Training set consists on N observations (sample)

$$X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\eta, \cdots, \mathbf{x}_N)$$

- Our goal is to partition the data set into some number *K* of clusters, where we shall suppose for the moment that the value of *K* is given.
- Clustering is a useful tool for data compression.
- Instead of reducing the dimensionality of a data set, clustering reduces the number of data points.

4

- Custer as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster
- It groups the data points into clusters according to a distance function.
- The points are similar to one another within the same cluster and dissimilar to the objects in other clusters

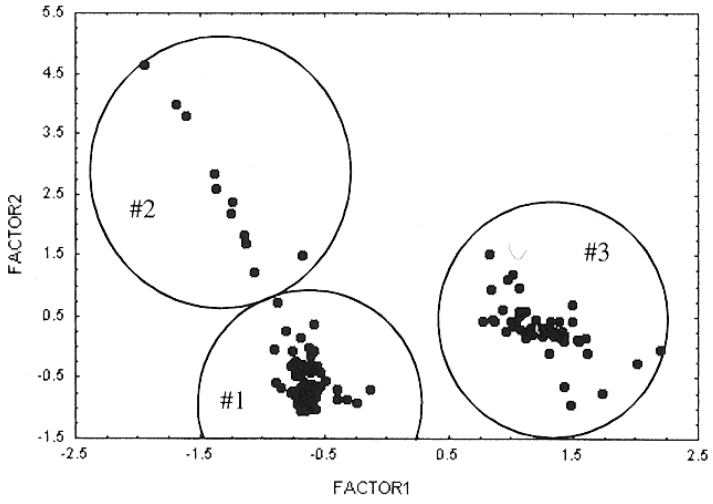
5

- The cluster centers (also called centroids) represent the compressed data set
- The most popular clustering method is *K*-means clustering. We map *N* data points, represented by vectors of dimension *D*, into *K* centroids with

$$K \ll N$$

6

K-Means Clustering



7

K-means Clustering

$$d_2(\mathbf{x}, \mathbf{y}) = \|x - y\| = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}.$$

In K-means clustering with K vectors called centroids w

$$\mathbf{c}_1, \mathbf{c}_2 \cdots, \mathbf{c}_K$$

and K sets called clusters

$$C_1, C_2 \cdots, C_K.$$

each cluster set is defined as the set of points with where

$$k = 1, ..., K$$

$$C_k = \{\mathbf{x} | d_2(\mathbf{x}, \mathbf{c}_k) = \min_j d_2(\mathbf{x}, \mathbf{c}_j)\}.$$

8

K-means Clustering

$$k = 1, ..., K$$

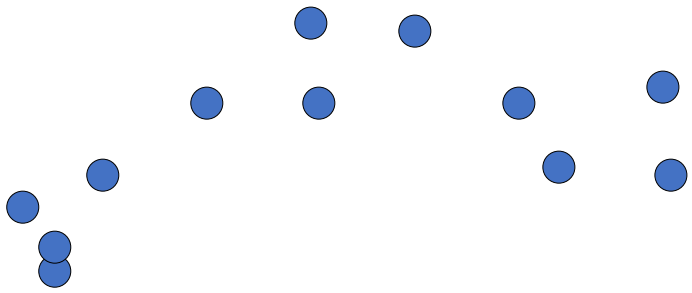
$$C_k = \{\mathbf{x} | d_2(\mathbf{x}, \mathbf{c}_k) = \min_j d_2(\mathbf{x}, \mathbf{c}_j)\}.$$

Each cluster C_k contains the points that are closest to the centroid $\mathbf{c}_k$.
centroid $\mathbf{c}_k$ is represented by the mean value of all the points of C_k

$$\mathbf{c}_k = \frac{1}{|C_k|} \cdot \sum_{x \in C_k} \mathbf{x}.$$

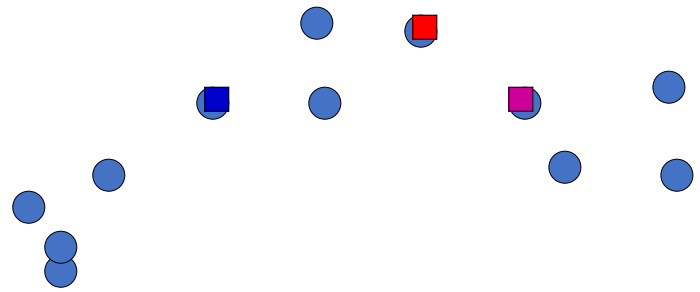
9

K-means: an example



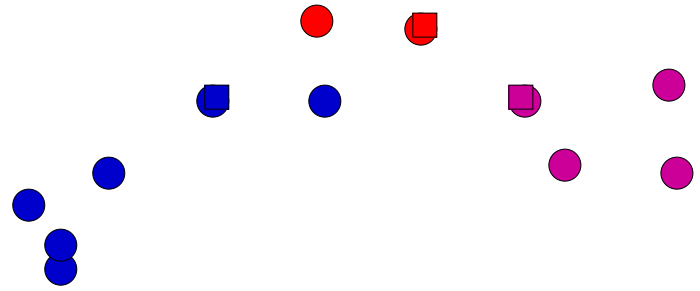
10

K-means: Initialize centers randomly



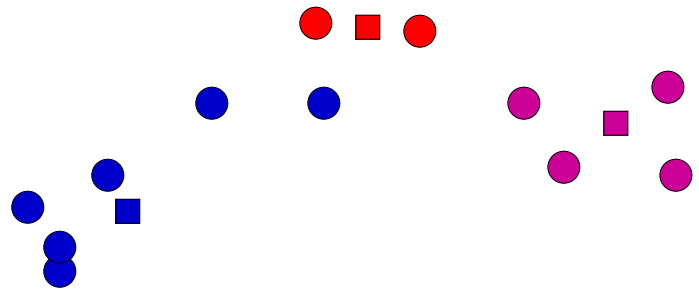
11

K-means: assign points to nearest center



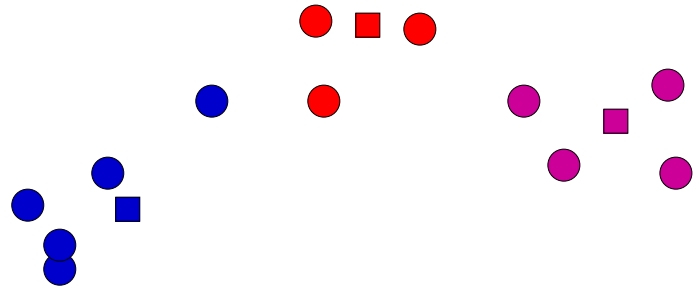
12

K-means: readjust centers



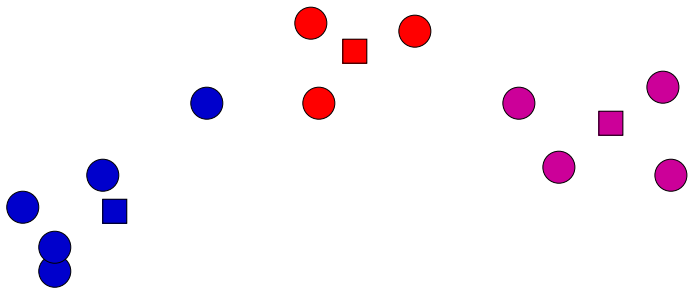
13

K-means: assign points to nearest center



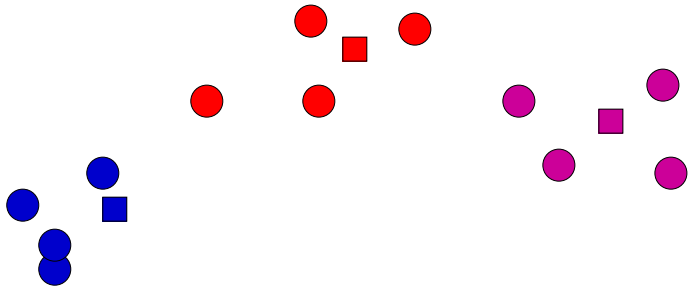
14

K-means: readjust centers



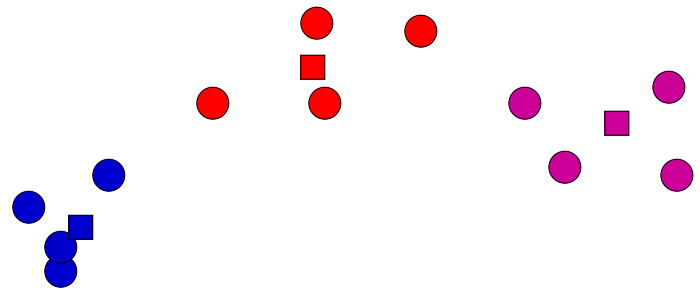
15

K-means: assign points to nearest center



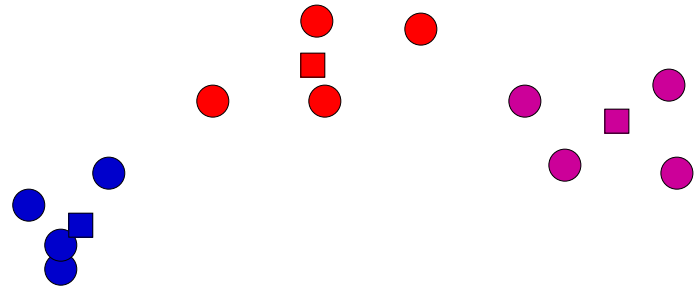
16

K-means: readjust centers



17

K-means: assign points to nearest center



No changes: Done

18

For each data point $\mathbf{x}_\eta$, we introduce a corresponding set of binary indicator variables

$$r_{\eta k} \in \{0, 1\}$$

with

$$k = 1, ..., K$$

describing which of the K clusters the data point $\mathbf{x}_\eta$, is assigned to, so that if data point $\mathbf{x}_\eta$, is assigned to cluster k then $r_{\eta k} = 1$, and $r_{\eta j} = 0$ for $j \neq k$

This is known as the 1-of-K coding scheme.

$$r_{\eta k} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_\eta - \mathbf{c}_j\|^2 \\ 0 & \text{otrwise} \end{cases}$$

19

We can then define an objective function, sometimes called a distortion measure, given by

$$J = E = \sum_{\eta=1}^{N} \sum_{k=1}^{K} r_{\eta k} \cdot \|\mathbf{x}_\eta - \mathbf{c}_k\|^2 = \sum_{k=1}^{K} \sum_{x \in C_k} (d_2(\mathbf{x}, \mathbf{c}_k))^2$$

we want to minimize it

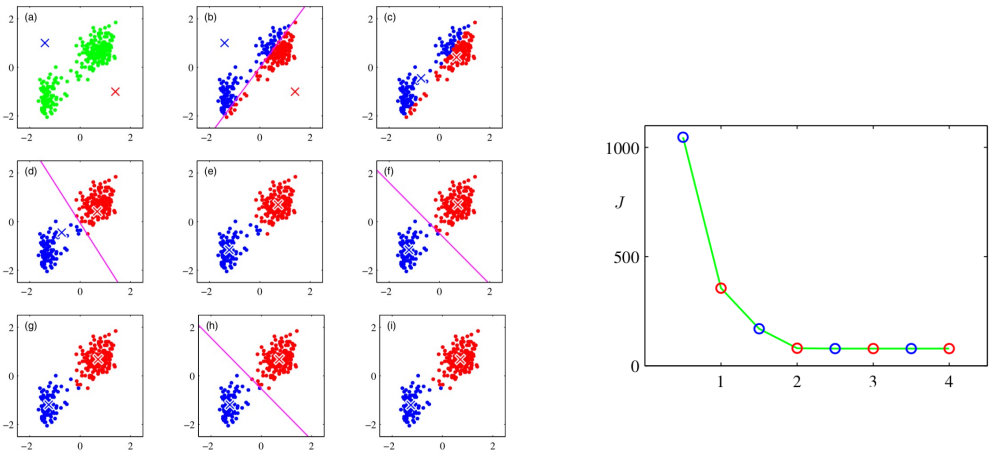
$$\frac{\partial J}{\partial \mathbf{c}_k} = -2 \cdot \sum_{\eta=1}^{N} r_{\eta k} \cdot (\mathbf{x} - \mathbf{c}_k)$$

$$\sum_{\eta=1}^{N} r_{\eta k} \cdot (\mathbf{x} - \mathbf{c}_k) = 0$$

$$\sum_{\eta=1}^{N} r_{\eta k} \cdot \mathbf{x} - \sum_{\eta=1}^{N} r_{\eta k} \cdot \mathbf{c}_k = 0$$

$$\mathbf{c}_k = \frac{\sum_{\eta=1}^{N} r_{\eta k} \cdot \mathbf{x}}{\sum_{\eta=1}^{N} r_{\eta k}} = \frac{1}{|C_k|} \cdot \sum_{x \in C_k} \mathbf{x}$$

20



21

Standard K-means

Random initialisation of K centroids;
do
{
assign to each $\mathbf{x}_\eta$ in the dataset the nearest centroid $\mathbf{c}_k$ according to d_2;
compute all new centroids $\mathbf{c}_k = \frac{1}{|C_k|} \cdot \sum_{x \in C_k} \mathbf{x}$;

}
until ($|E_{new} - E_{old}| < \epsilon$ or number of iterations max iterations).

22

Sequential K-means

For large data sets, the adaptive K-means learning algorithm is given by with sequential update in which, for each data point

$$\mathbf{c}_k^{new} = \mathbf{c}_k^{old} + \eta_\eta \cdot (\mathbf{x}_\eta - \mathbf{c}_k^{old}) = \mathbf{c}_k^{old} + \frac{1}{|C_k^{old}| + 1} \cdot (\mathbf{x}_\eta - \mathbf{c}_k^{old})$$

where η_η is the learning rate parameter, which is typically made to decrease monotonically as more data points are considered and can be represented by $\frac{1}{|C_k^{old}|+1}$.

23

Sequential K-means

Random initialisation of K centroids;
do
{
chose $\mathbf{x}_\eta$ from the dataset;
determine the nearest centroid $\mathbf{c}_k$ according to d_2;
compute the new centroid $\mathbf{c}_k^{new} = \mathbf{c}_k^{old} + \frac{1}{|C_k^{old}|+1} \cdot (\mathbf{x}_\eta - \mathbf{c}_k^{old})$;
}
until ($|E_{new} - E_{old}| < \epsilon$ or number of iterations max iterations).

24

- K-means represents an unsupervised learning; it is an unsupervised classification because no predefined classes are present.
- One notable feature of the K-means algorithm is that at each iteration, every data point is assigned uniquely to one, and only one, of the clusters.

25

Color reduction

- K-means can be used for color reduction for RGB images.
- K would indicate the number of the reduced colors, and the dimension would be there for *R, G, B.* $x_i = R_i, G_i, B_i$ would correspond to the pixel at position i in a one dimensional array.
- Color segmentation represents a weak segmentation, in which the segmented parts (the same color) do not correspond necessarily to objects.

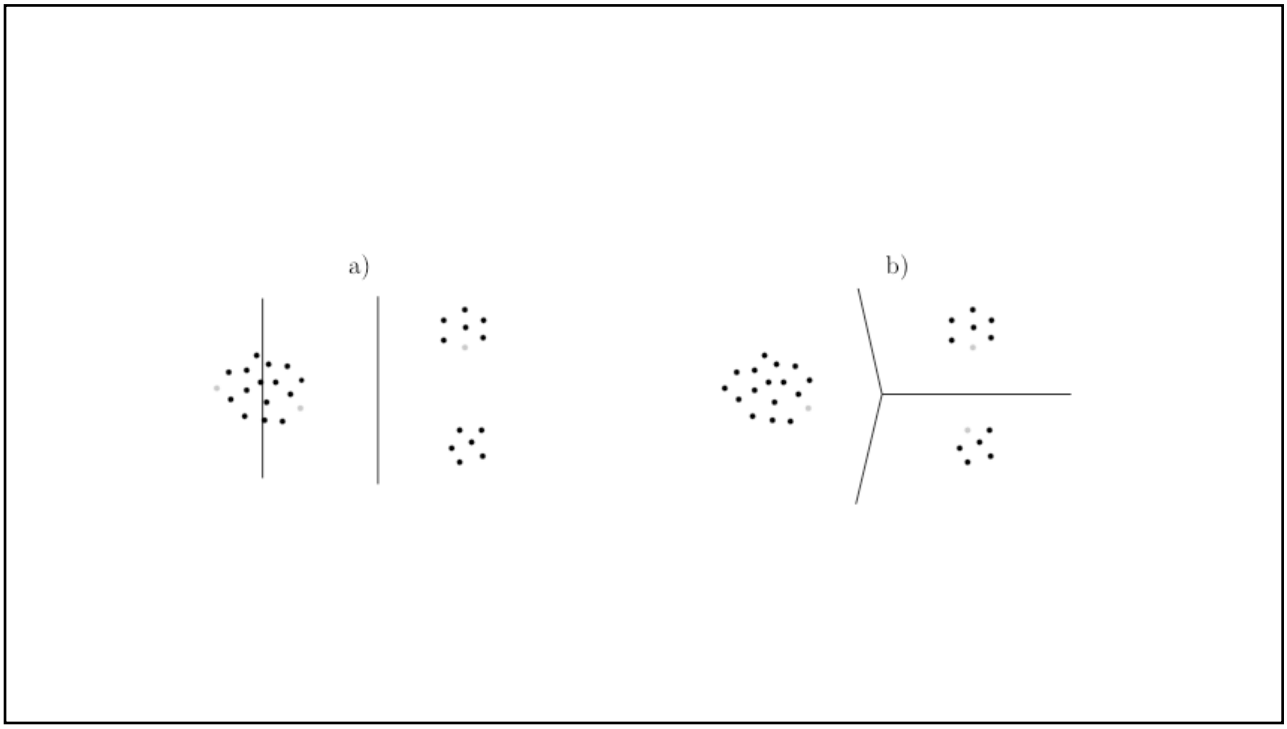
26

Color reduction

27

- How to chose *K* ?
  - You have to know your data!
- Repeated runs of *K*-means clustering on the same data can lead to quite different partition results
  - Why? Because we use random initialization

28



29

Adaptive Initialization

- Choose a maximum *radius* within every data point should have a cluster seed after completion of the initialization phase
- In a single sweep go through the data and assigns the cluster seeds according to the chosen r*adius*
  - A data point becomes a new cluster seed, if it is not covered by the spheres with the chosen *radius* of the other already assigned seeds
  - K-MAI clustering (Wichert et al. 2003)

30

Mixture of Gaussians

Gaussian distribution or normal is defined over D dimensional space

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

where

- $\boldsymbol{\mu}$ is the D dimensional mean vector
- Σ is a $D \times D$ covariance matrix
- $|\Sigma|$ is the determinant of Σ

31

Gaussian Mixture Distribution

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

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \cdot \frac{1}{(2 \cdot \pi)^{D/2}} \cdot \frac{1}{|\Sigma_k|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

with

$$0 \leq \pi_k \leq 1$$

and

$$\sum_{k=1}^{K} \pi_k = 1$$

32

We will now use $c_k = 1$ to denote the cluster k and $p(c_k = 1|\mathbf{x})$

$$p(c_k = 1|\mathbf{x}) = \frac{p(\mathbf{x}|c_k = 1) \cdot p(c_k = 1)}{p(\mathbf{x})}$$

$$p(c_k = 1|\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \cdot p(c_k = 1)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \cdot p(c_k = 1)}$$

We define $\gamma(c_k)$ to be equivalent to $p(c_k = 1|\mathbf{x})$

$$\gamma(c_k) \equiv p(c_k = 1|\mathbf{x}) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)}$$

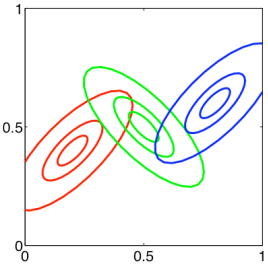
We define $\gamma(c_{\eta k})$ to be equivalent to $p(c_k = 1|\mathbf{x})_\eta$ for a certain pattern $\mathbf{x}_\eta$ with the index η

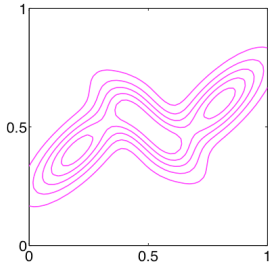
$$\gamma(c_{\eta k}) \equiv p(c_k = 1|\mathbf{x}_\eta) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)}$$

33

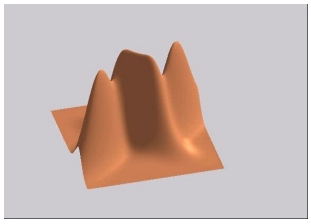
Example: Mixture of 3 Gaussians *k*

$p(\mathbf{x}|k)$



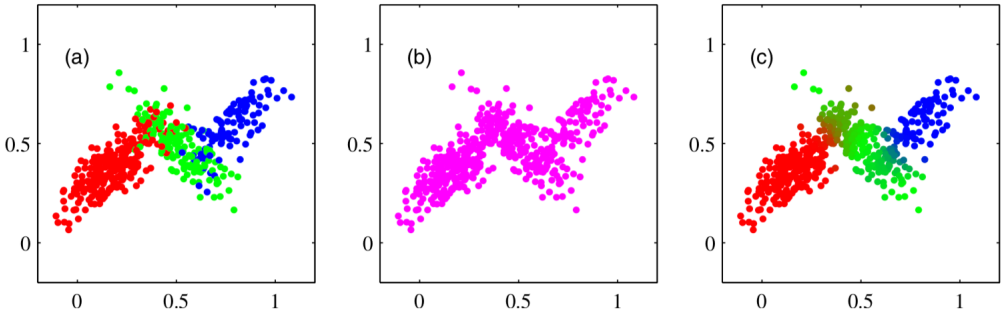


$p(\mathbf{x})$



$p(\mathbf{x})$

34



- (a) corresponding to the three components of the mixture, are depicted in red, green, and blue,
- (b) the corresponding samples from the marginal distribution ***p(x)***
- (c) The same samples in which the colours represent the value of the responsibilities

$$p(c_k = 1|\mathbf{x}) = \frac{p(\mathbf{x}|c_k = 1) \cdot p(c_k = 1)}{p(\mathbf{x})}$$

35

Maximum likelihood

Training set consists on N observations (sample)

$$X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\eta, \cdots, \mathbf{x}_N)$$

We can represent the dataset as a design matrix X of the dimension $N \times D$ as before

The log of the likelihood function is given by

$$\log p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{\eta=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

Significant problem associated with the maximum likelihood framework applied to Gaussian mixture models

36

Significant problem

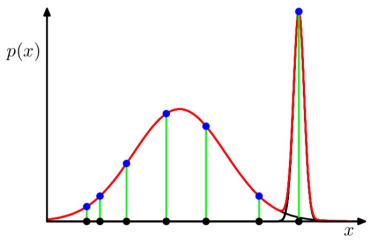
Consider a Gaussian mixture whose components have covariance matrices given by $\Sigma_k = \sigma_k^2 \cdot I$ with I being the identity matrix.

Suppose that one of the components of the mixture model, let us say the j^th component, has its mean $|\boldsymbol{\mu}_k$ exactly equal to one of the data points

$$\boldsymbol{\mu}_k = \mathbf{x}_\eta$$

$$\mathcal{N}(\mathbf{x}_\eta | \mathbf{x}_\eta, \sigma_j^2 \cdot I) = \frac{1}{(2 \cdot \pi)^{1/2}} \cdot \frac{1}{\sigma_j}$$

For $\sigma_j \to 0$ the term goes to infinity



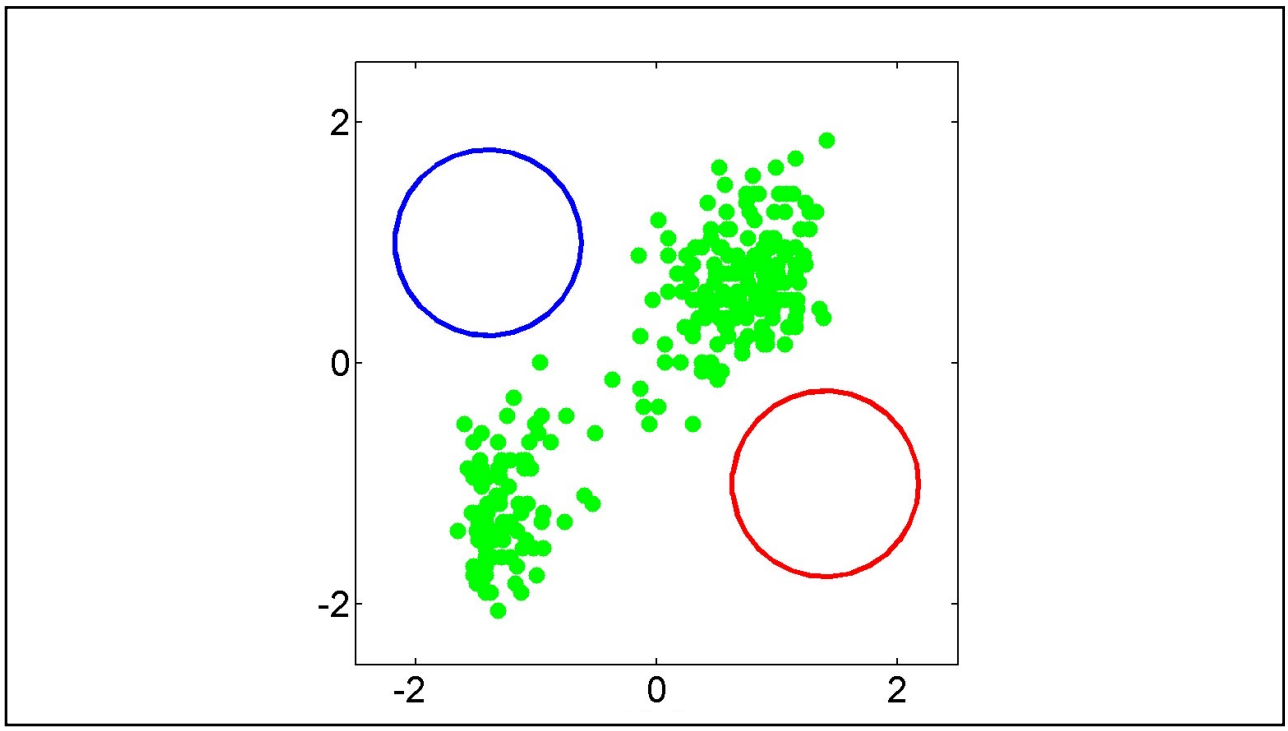
37

Algorithm: EM for Gaussian mixtures

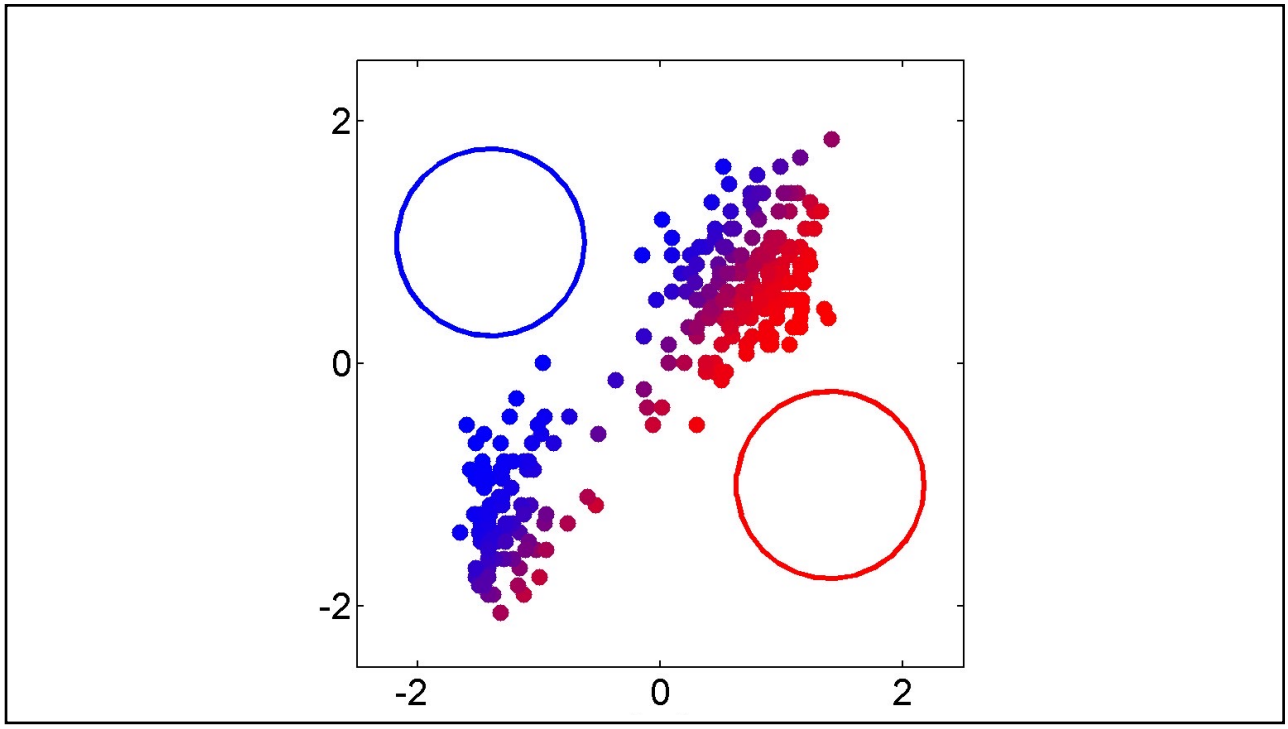
Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).
Training set consists on N observations (sample)

$$X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\eta, \cdots, \mathbf{x}_N)$$

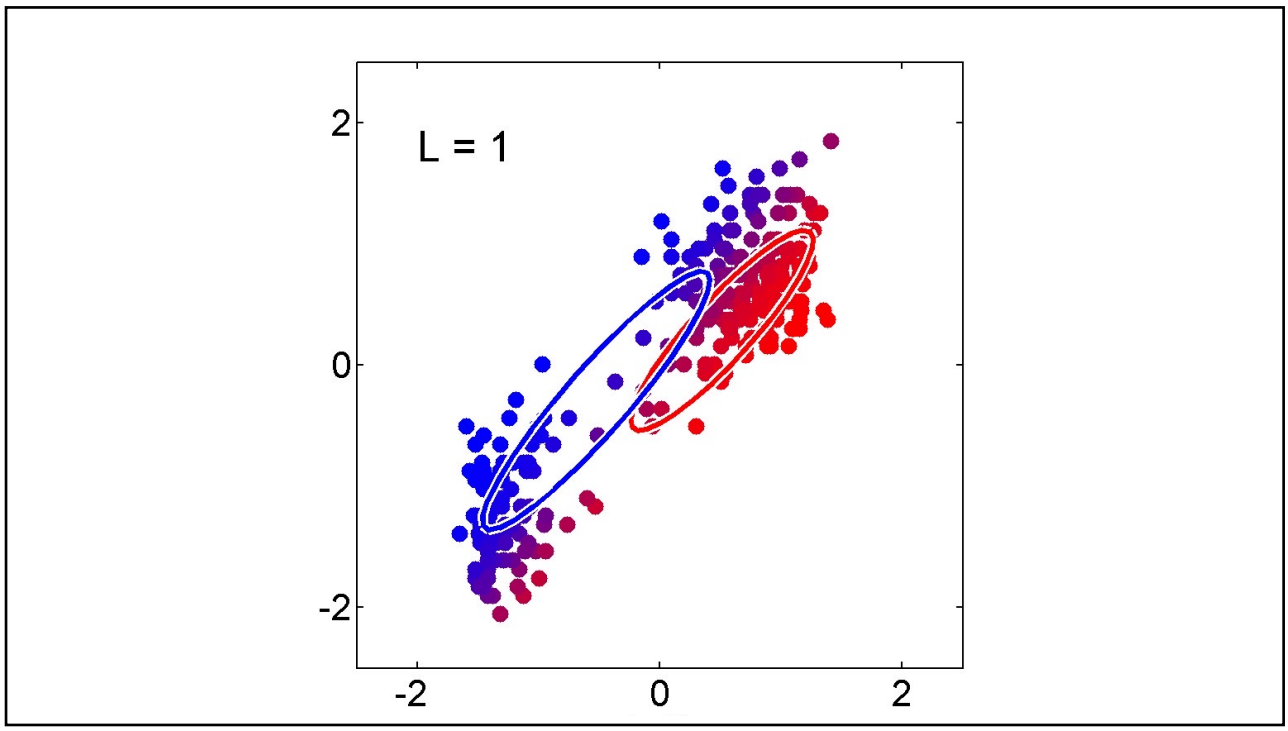
38



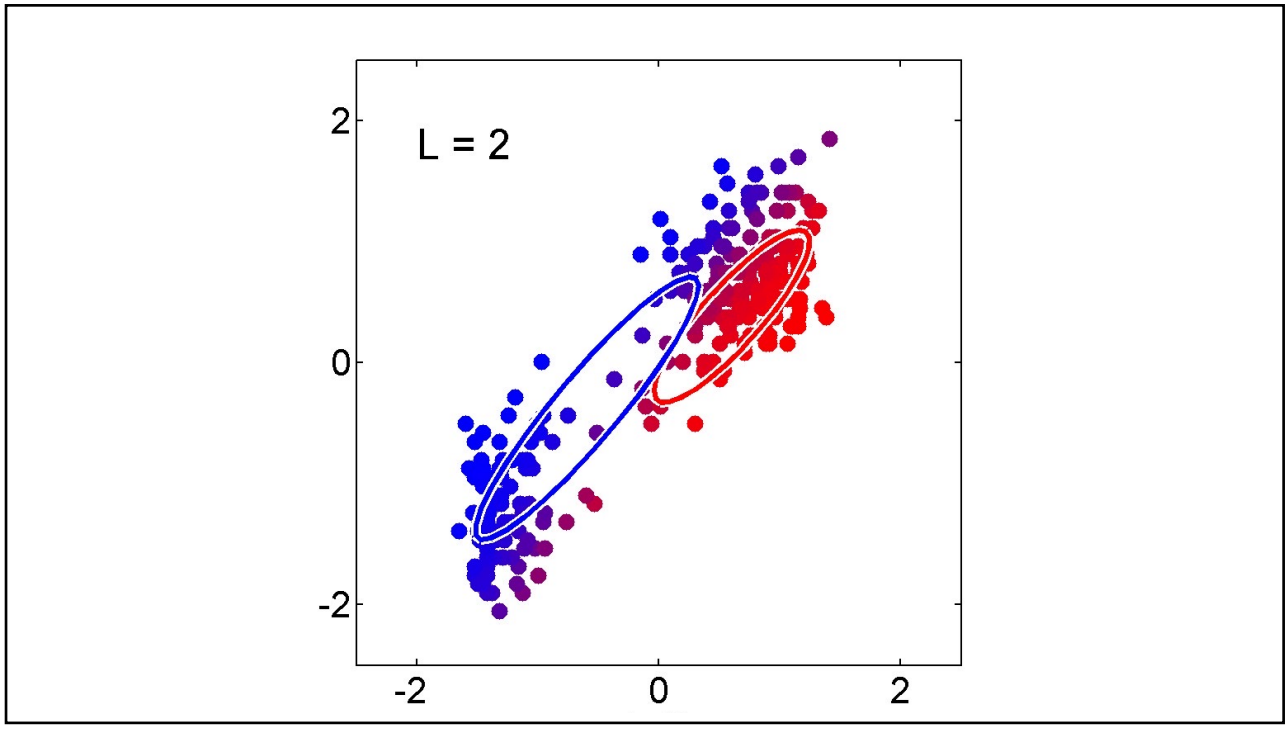
39



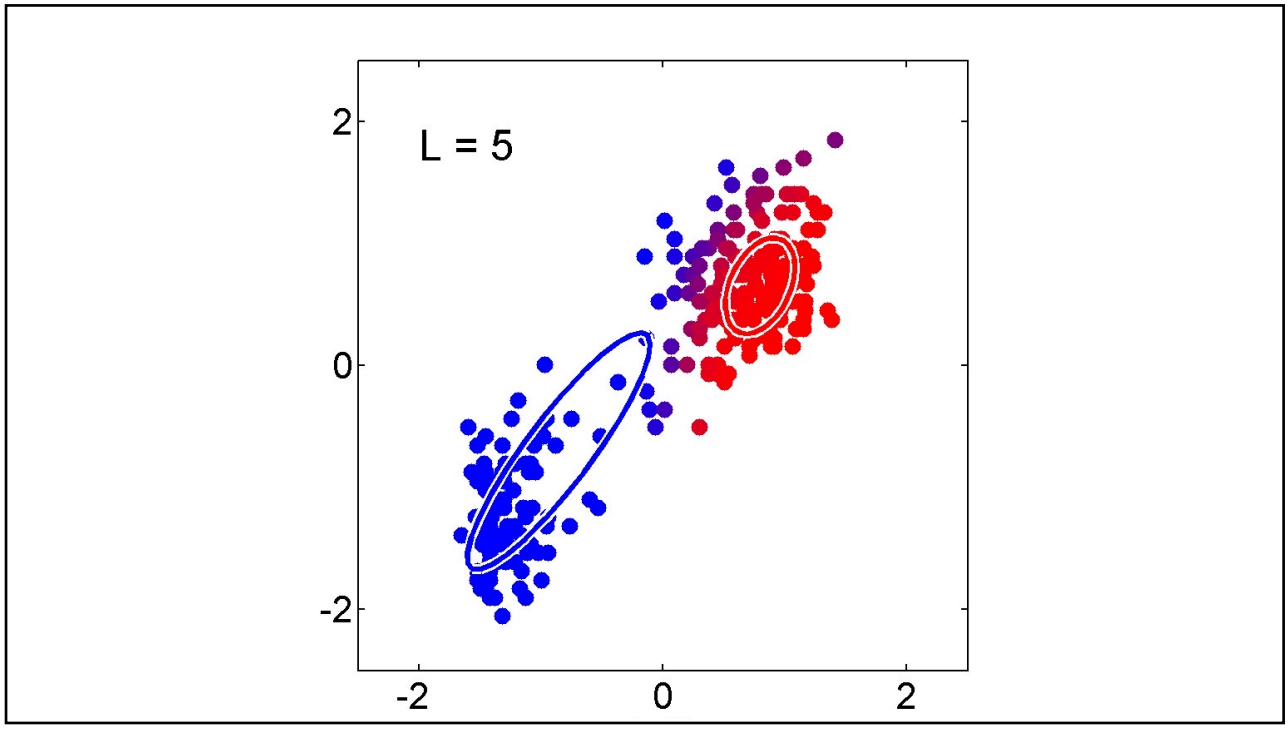
40



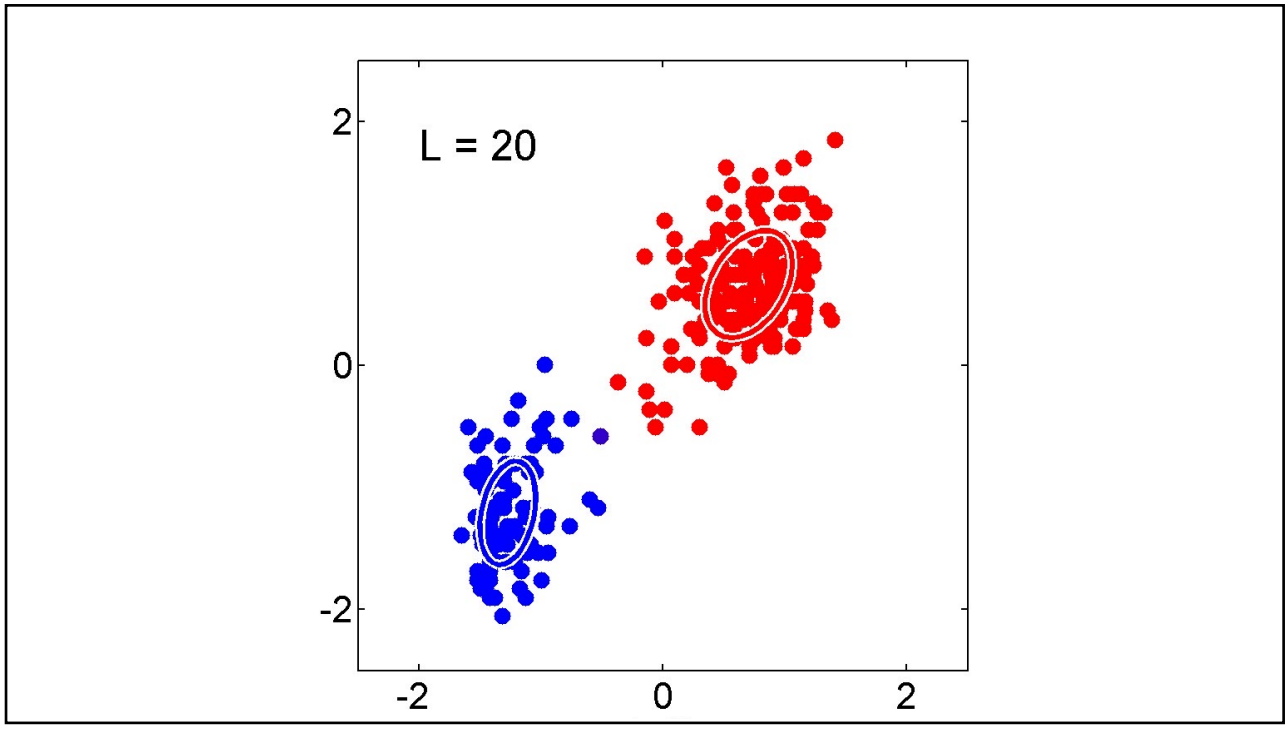
41



42



43



44

Algorithm: EM for Gaussian mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).
Training set consists on N observations (sample)

$$X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\eta, \cdots, \mathbf{x}_N)$$

45

EM for Gaussian mixtures

1. Initialisation:

   Chose K, number of clusters. Then initialise

   - the means $\boldsymbol{\mu}_k$ (centres of clusters, random data point or random value)
   - Σ_k covariances (shape of clusters, usually we can start with identity matrix I)
   - $\pi_k = p(c_k = 1)$ mixing coefficients (prior, importance of clusters, usually we can start with the value $\frac{1}{K}$, each cluster has the same importance

46

2. **E-Step** (Expectation):

Compute for each data point η and each cluster k

$$\gamma(c_{\eta k}) = p(c_k = 1|\mathbf{x}_\eta) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)}$$

π_k is the mixture, cluster k

$$p(c_k = 1|\mathbf{x}_\eta) = \frac{p(c_k = 1) \cdot p(\mathbf{x}_\eta|c_k = 1)}{p(\mathbf{x}_\eta)}$$

Usually we can compute the likelihood

$$p(c_k = 1, \mathbf{x}_\eta) = \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)$$

with

$$p(\mathbf{x}_\eta|c_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2 \cdot \pi)^{D/2}} \cdot \frac{1}{|\Sigma_k|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)\right)$$

47

EM for Gaussian mixtures

$$p(c_k = 1, \mathbf{x}_\eta) = \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)$$

with

$$p(\mathbf{x}_\eta|c_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2 \cdot \pi)^{D/2}} \cdot \frac{1}{|\Sigma_k|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)\right)$$

and after it

$$p(\mathbf{x}_\eta) = \sum_{k=1}^{K} p(c_k = 1, \mathbf{x}_\eta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)$$

and normalize

$$\gamma(c_{\eta k}) = p(c_k = 1|\mathbf{x}_\eta) = \frac{p(c_k = 1, \mathbf{x}_\eta)}{p(\mathbf{x}_\eta)}$$

48

3. **M-Step** (Maximization):

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \cdot \sum_{\eta=1}^{N} \gamma(c_{\eta k}) \cdot \mathbf{x}_\eta$$

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{\eta=1}^{N} \gamma(c_{\eta k}) \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k) \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)^T$$

$$\pi_k = p(c_k = 1) = \frac{N_k}{N}$$

with

$$N_k = \sum_{\eta=1}^{N} \gamma(c_{\eta k})$$

49

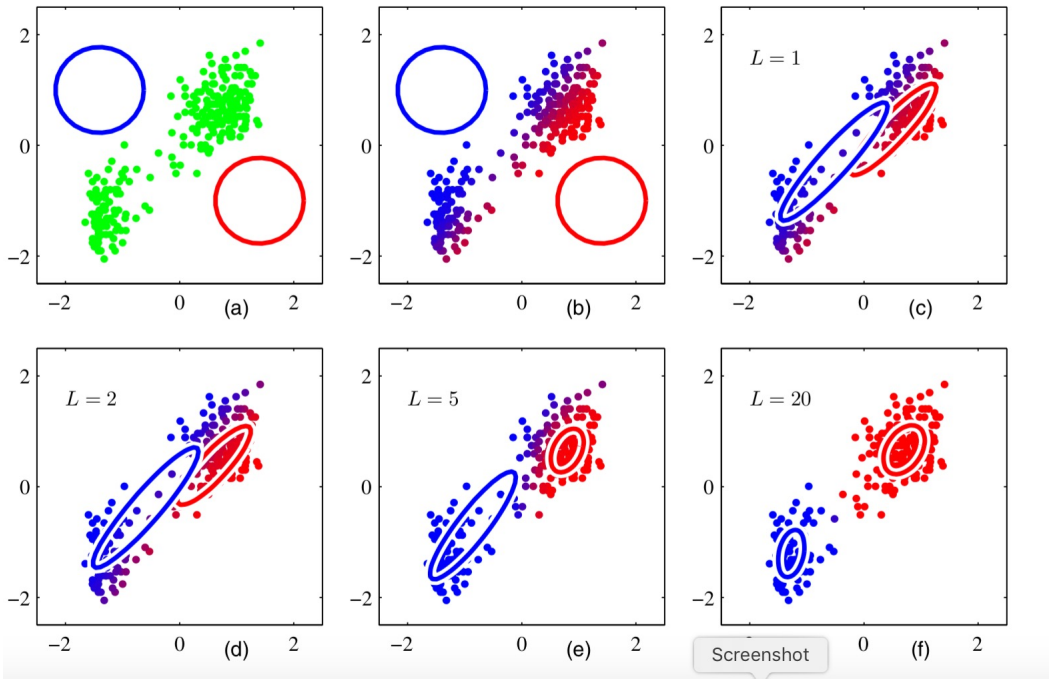
EM for Gaussian mixtures

4. Evaluate the log likelihood

$$\log p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{\eta=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k) \right)$$

and check for convergence of either the parameters or the log likelihood. the convergence criterion is not satisfied or below number of iterations *max* iterations, return to step 2.

50



51

Given the data

$$X = \left(\mathbf{x}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)$$

with $K = 2$ and

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

and

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\pi_1 = p(c_1 = 1) = 0.6, \quad \pi_2 = p(c_2 = 1) = 0.4,$$

we will perform a step of the EM clustering algorithm.

52

E-Step (Expectation)

Compute for each data point n and each cluster k

$$p(\mathbf{x}_n|c_k=1)=\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)=$$

$$\frac{1}{(2\cdot\pi)^{D/2}}\cdot\frac{1}{|\Sigma_k|^{1/2}}\cdot\exp\left(-\frac{1}{2}\cdot(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}\cdot(\mathbf{x}_n-\boldsymbol{\mu}_k)\right).$$

$$p(\mathbf{x}_1|c_1=1)=\frac{1}{(2\cdot\pi)^{2/2}}\cdot\frac{1}{1^{1/2}}\cdot\exp\left(-\frac{1}{2}\cdot(2-2,2-2)\cdot\begin{pmatrix}1&0\\0&1\end{pmatrix}\cdot\begin{pmatrix}2-2\\2-2\end{pmatrix}\right),$$

$$p(\mathbf{x}_1|c_1=1)=0.159155,\ \ p(\mathbf{x}_2|c_1=1)=0.0215393,\ \ p(\mathbf{x}_3|c_1=1)=0.00291502,$$

$$p(\mathbf{x}_1|c_2=1)=0.00291502,\ \ p(\mathbf{x}_2|c_2=1)=0.0215393,\ \ p(\mathbf{x}_3|c_2=1)=0.159155,$$

Then

$$p(c_k=1,\mathbf{x}_n)=\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)$$

$$p(\mathbf{x}_1,c_1=1)=0.095493,\ \ p(\mathbf{x}_2,c_1=1)=0.0129236,\ \ p(\mathbf{x}_3,c_1=1)=0.00174901,$$

$$p(\mathbf{x}_1,c_2=1)=0.00116601,\ \ p(\mathbf{x}_2,c_2=1)=0.00861571,\ \ p(\mathbf{x}_3,c_2=1)=0.063662,$$

53

$$p(\mathbf{x}_n)=\sum_{k=1}^{K}p(c_k=1,\mathbf{x}_n)=\sum_{k=1}^{K}\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)$$

$$p(\mathbf{x}_1)=0.096659,\quad p(\mathbf{x}_2)=0.0215393,\quad p(\mathbf{x}_3)=0.065411,$$

using Bayes (normalizing)

$$\gamma(c_{nk})=p(c_k=1|\mathbf{x}_n)=\frac{p(c_k=1,\mathbf{x}_n)}{p(\mathbf{x}_n)}.$$

$$\gamma(c_{11})=p(c_1=1|\mathbf{x}_1)=0.987937,\quad \gamma(c_{21})=0.6,\quad \gamma(c_{31})=0.0267388,$$

$$\gamma(c_{12})=p(c_2=1|\mathbf{x}_1)=0.0120631,\quad \gamma(c_{22})=0.4,\quad \gamma(c_{32})=0.973261.$$

54

M-Step (Maximization)

We evaluate

$$N_k = \sum_{n=1}^{N} \gamma(c_{nk})$$

$$N_1 = 1.61468, \quad N_2 = 1.38532.$$

we determine the mean values

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot \mathbf{x}_n$$

$$\boldsymbol{\mu}_1 = \frac{1}{1.61468} \cdot \left(0.987937 \cdot \begin{pmatrix} 2 \\ 2 \end{pmatrix} + 0.6 \cdot \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.0267388 \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right),$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1.2237 \\ 1.96688 \end{pmatrix}.$$

and

$$\boldsymbol{\mu}_2 = \frac{1}{1.38532} \cdot \left(0.0120631 \cdot \begin{pmatrix} 2 \\ 2 \end{pmatrix} + 0.4 \cdot \begin{pmatrix} 0 \\ 2 \end{pmatrix} + 0.973261 \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right),$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 0.0174156 \\ 0.594898 \end{pmatrix}.$$

55

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(c_{nk}) \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k) \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

$$\Sigma_1 = \frac{1}{1.61468} \cdot \left(0.987937 \cdot \begin{pmatrix} 2 - 1.2237 \\ 2 - 1.96688 \end{pmatrix} \cdot (2 - 1.2237, 2 - 1.96688) + \right.$$

$$0.6 \cdot \begin{pmatrix} 0 - 1.2237 \\ 2 - 1.96688 \end{pmatrix} \cdot (0 - 1.2237, 2 - 1.96688) +$$

$$\left. 0.0267388 \cdot \begin{pmatrix} 0 - 1.2237 \\ 0 - 1.96688 \end{pmatrix} \cdot (0 - 1.2237, 0 - 1.96688) \right),$$

$$\Sigma_1 = \begin{pmatrix} 0.94996 & 0.0405286 \\ 0.0405286 & 0.0651426 \end{pmatrix},$$

and

$$\Sigma_2 = \begin{pmatrix} 0.0345279 & 0.0244707 \\ 0.0244707 & 0.835892 \end{pmatrix}.$$

56

and the new mixing parameter is

$$\pi_k = p(c_k = 1) = \frac{N_k}{N}$$

$$\pi_1 = p(c_1 = 1) = \frac{1.61468}{3} = 0.538227, \;\; \pi_2 = p(c_2 = 1) = \frac{1.38532}{3} = 0.461775.$$

57

Cluster validation

- The procedure of evaluating the results of a clustering algorithm is known under the term cluster validity
- In general terms, there are **three approaches** to investigate cluster validity

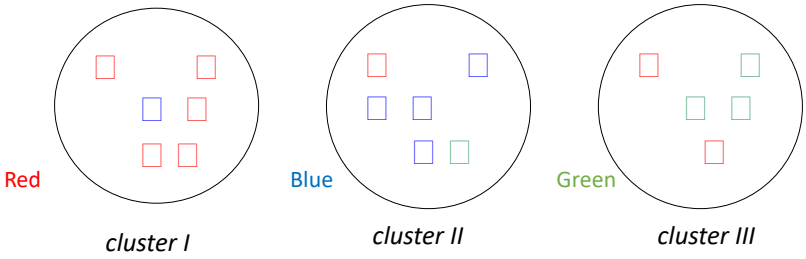
58

I) External Criteria

- The first is based on **external criteria**
- This implies that we evaluate the results of a clustering algorithm based on a **pre-specified structure**, which is imposed on a data set and reflects our intuition about the clustering structure of the data set

59

Purity: We assign a label to each cluster based on the most frequent class in it. Then the purity becomes the number of correctly matched class and cluster labels divided by the number of total data points.



$$purity = \frac{1}{17}(max(1,0,5) + max(4,1,1) + max(0,3,2)) = \frac{12}{17}$$

$$purity(\boldsymbol{c}_1) = \frac{5}{6}, \quad purity(\boldsymbol{c}_2) = \frac{4}{6}, \quad purity(\boldsymbol{c}_3) = \frac{3}{5}$$

60

External measures: purity

- $C = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ is the set of clusters
- $L = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_G\}$ is the set of **reference classes**

$$purity(C, L) = \frac{1}{n}\sum_{k=1}^{K} \max_j(|\boldsymbol{c}_k \cap \boldsymbol{l}_j|)$$

61

II) Internal Criteria

- The second approach is based on **internal criteria**
- We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves

- Distance (how close two objects are to each other)
- Similarity (how similar/distinct two objects are)

62

- The two criteria were proposed for clustering evaluation and selection of an optimal clustering scheme (Berry and Linoff, 1996)
- Compactness, the members of each cluster **should be as close to each other as possible.** A common measure of compactness is the variance, which should be minimized
- Separation, the **clusters themselves should be widely spaced**
- *We present three indexes*

63

i) Validity Dunn Index

- Dunn index, a cluster validity index for *K*-means clustering proposed in Dunn (1974)
- Attempts to identify "compact and well separated clusters"
  - *Notation: k the number of clusters*

64

i) Dunn index

$$d(C_i, C_j) = \min_{\vec{x} \in C_i, \vec{y} \in C_j} d(\vec{x}, \vec{y})$$

$$diam(C_i) = \max_{\vec{x}, \vec{y} \in C_i} d(\vec{x}, \vec{y})$$

$$D_k = \min_{1 \le i \le k} \left\{ \min_{\substack{1 \le j \le k \\ i \ne j}} \left\{ \frac{d(C_i, C_j)}{\max_{1 \le l \le k} \{diam(C_l)\}} \right\} \right\}$$

65

The implications of the Dunn index

- Considerable amount of time required for its computation
- **Sensitive to the presence of noise** in datasets, since these are likely to increase the values of *diam(c)*
- If the dataset contains compact and well-separated clusters, the *distance* between the clusters is expected to be *large* and the *diameter* of the clusters is expected to be *small*
- **Large values** of the index indicate the presence of compact and well-separated clusters

66

ii) Validity Davies-Bouldin Index

- The Davies-Bouldin (**DB**) index (1979)

$$d(C_i, C_j) = \min_{\vec{x} \in C_i, \vec{y} \in C_j} d(\vec{x}, \vec{y})$$

$$diam(C_i) = \max_{\vec{x}, \vec{y} \in C_i} d(\vec{x}, \vec{y})$$

$$DB_k = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)} \right\}$$

67

The implications of the Davies-Bouldin Index

- **Small indexes** correspond to good clusters, clusters are compact and their centers are far away
- Considerable amount of time required for its computation
- **Not sensitive to the presence of noise** since we take into account all clusters

68

Different Methods: Diameter - Distance

- Different methods to calculate the **diameter of a cluster**
- There are three common approaches measuring the **distance between two different clusters**
  - Single linkage: It measures the distance between the closest members of the clusters
  - Complete linkage: It measures the distance between the most distant members
  - Comparison of centroids: It measures the distance between the centers of the clusters
- **Silhouette "**overcomes" this problems

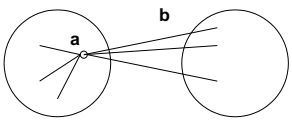
69

iii) Validity index Silhouette

- Silhouette combines both **cohesion** and **separation**
- Calculated for a specific object $\mathbf{x}_i$
  - ***a*** **= average distance** of $\mathbf{x}_i$ to the points in its cluster
  - ***b*** = min (average distance of $\mathbf{x}_i$ to points in another cluster)
  - the silhouette coefficient for a point is then given by

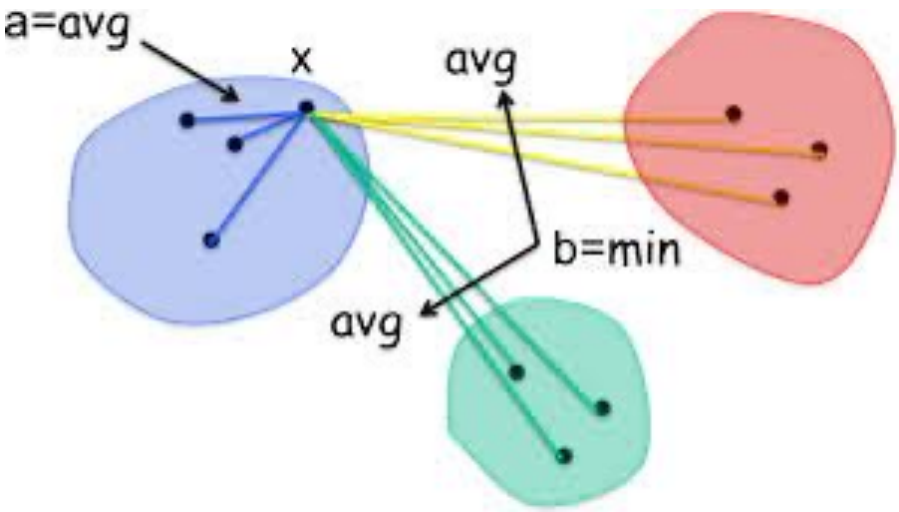
    s = 1 – a/b if a < b, (*or s = b/a - 1 if a ≥ b, not the usual case*)

    between -1 and 1 (the closer to 1 the better)

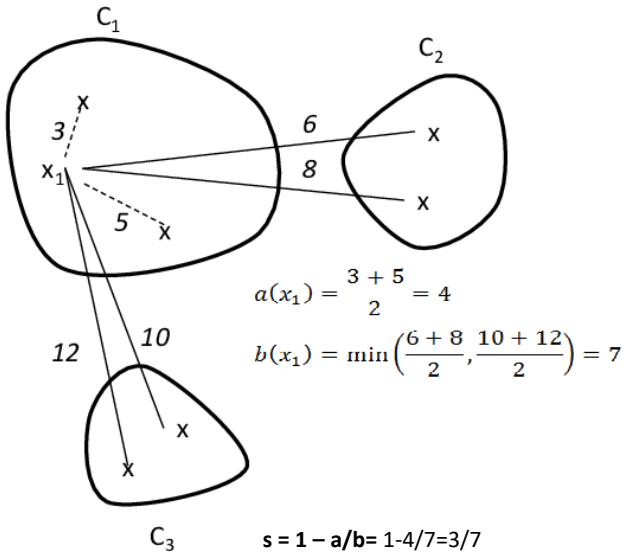


- Silhouette of **cluster**: average of observation silhouettes
- Silhouette of **clustering solution**: average of cluster silhouettes

70



71



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

s = 1 – a/b= 1-4/7=3/7

72

The implications of the Silhouette Index

- **Considerable amount of time** required for its computation
  - *We have to compute silhouette for each point in a cluster for each cluster…..*
- Diameter of a cluster and Distance between clusters well defined
- **Not sensitive to the presence of noise** since we take into account all clusters

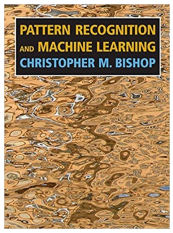
73

III) Relative Criteria

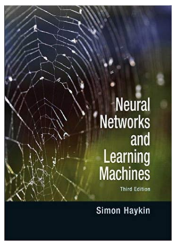
- Here the basic idea is the evaluation of a clustering structure by **comparing** it to other **clustering schemes**, resulting by the *same* algorithm but with *different parameter values*
- Compare different cluster structures, different parameters or algorithms
  - *Make many experiments with different initialization of cluster centers and number of clusters*
  - *Chose the best clustering according to the index*
  - *EM, vs k-Means with different parameters*

74

Literature



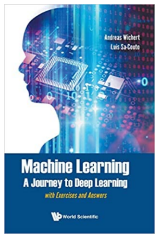
- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - Chapter 9



- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008
  - Chapter 9

75

Literature



- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021
  - Chapter 9

76