UNIVERSITÀ
di VERONA

Machine Learning

Parametric Unsupervised Learning

Cigdem Beyan

A. Y. 2024/2025

UNIVERSITÀ di VERONA

Unsupervised Learning

- **Non-Parametric**
  - **Clustering:** no assumptions are made about the underlying densities, instead we seek a partition of the data into clusters

- **Parametric approach**
  - Assume the parametric distribution of data
  - Model the underlying **class-conditional densities** with a mixture of parametric densities
  - The objective is to find the model parameters
    - Estimate parameters of the distribution assumed

Parametric Unsupervised Learning

UNIVERSITÀ di VERONA

- We assume the data was generated by a model with known shape but unknown parameters
- What is good about having a model? We can:
  - Adjust the parameters of the model to maximize the probability that our model produced for the observed data
  - Compute the likelihood of data
  - Compare multiple algorithms' performance: whichever gives the higher likelihood for the observed data

UNIVERSITÀ di VERONA

Recap: Parametric supervised Learning

- Given $\boldsymbol{m}$ classes
- Samples $\mathbf{x_1}, \ldots, \mathbf{x_n}$ each of class $\mathbf{1, 2, \ldots, m}$
- Suppose $\boldsymbol{D_i}$ has samples from class $\boldsymbol{i}$
- The probability distribution for class $\boldsymbol{i}$ is $p_i(x|\theta_i)$
  - For class 1: $p_1(x|\theta_1)$
  - For class 2: $p_2(x|\theta_2)$
- Use **Maximum likelihood** method to estimate parameters θ_i

Parametric Unsupervised Learning

UNIVERSITÀ di VERONA

- We do not have the true classes for samples. But we still know that
  - We have $\mathbf{m}$ classes
  - Have samples $\mathbf{x_1}, \ldots, \mathbf{x_n}$ each of unknown class
  - Probability distribution for class $\boldsymbol{i}$ is $\mathbf{p_i(x \mid \theta_i)}$

  - The task is in a way to determine the classes and parameters simultaneously

Example Application

- MRI brain image segmentation
  - Different brain tissues have different intensities
  - There are 6 major tissue types
  - Each type of tissue can be modelled by a Gaussian distribution
    - Parameters: **mean** and **standard deviation** ➔ **unknown**
  - Segmenting (classifying) the brain image into different tissue classes is our task. But we do not know:
    - which image pixel corresponds to which tissue (class)
    - parameters for **Gaussian distribution**

Gaussian Mixture Model (GMM)

- Gaussian mixtures of D-dimensional variables with *K* Gaussian components is written as:

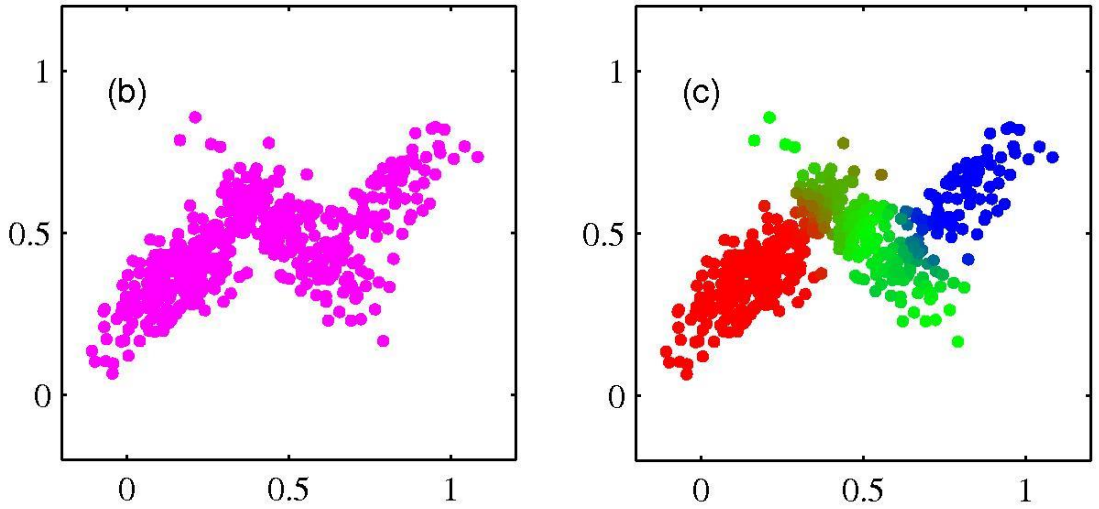
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

- GMM introduces a D-dimensional **binary** variable called **latent variable** shown as **z** (which is not observed, but inferred during modelling process) to indicate which Gaussian component a data point belongs to.
  - Let $z = z_1, \ldots z_k$ $z_k \in \{0, 1\}$, $\boldsymbol{\Sigma_k z_k = 1}$: Only one component of *z*z can be active (equal to 1) at a time
  - We have K possible states of **z** corresponding to **K components**
  - One of K representation is: [0…..0 1 0 0…..0] (one-hot vector)
    - E.g., z=[0,0,1,0,0] indicates that the data point belongs to the third Gaussian component.

Gaussian Mixture Model (GMM)

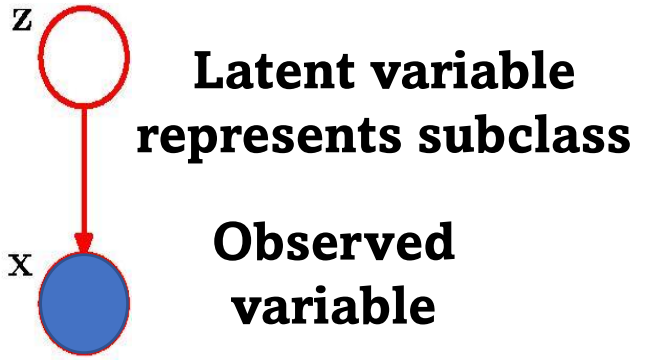
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

- The aim is to find maximum likelihood parameters: $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$
- Each data point is associated with a subclass (sub-group) ***k*** with probability π_k



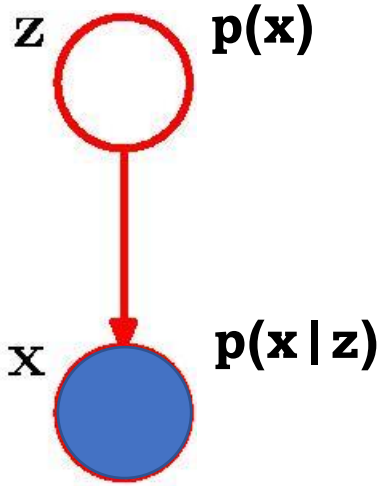
Gaussian Mixture Model (GMM)

- The joint probability of latent variable ***(z)*** and observed variable ***(x)***
  - $p(x,z) = p(x|z)\,p(z)$
    - ***x*** is the observed variable
    - ***z*** is the hidden or missing variable: latent variable
    - **Marginal distribution** or **prior probability *p(z)*:** Representing the likelihood of z independently of x.
    - **Conditional distribution *p(x|z):*** describes how the observed variable ***x*** depends on the state of the latent variable ***z***.
    - Graphical representation:



Gaussian Mixture Model (GMM)

- By using **p(z)** and **p(x | z)**
  - we specify **p(x)** in terms of **x** and **z**
- Each component $\mathbf{z_k}$ is assigned a probability
  - $p(z_k = 1) = \pi_k$, where parameters π_k satisfy
    $0 \leq \pi_k \leq 1$ and $\Sigma_k \pi_k = 1$
- Because **z** uses **1-of-K** it follows that $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
  since $\mathbf{z_k} \in \{0, 1\}$ and components of **z** are mutually exclusive and therefore independent.
  With one mixture component: $p(z_1) = \pi_1^{z_1}$
  With two components: $p(z_1, z_2) = \pi_1^{z_1} \pi_2^{z_2}$



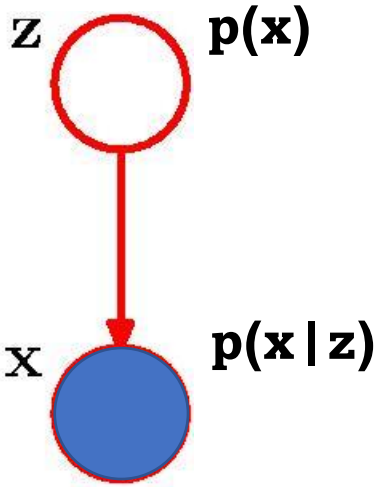
Gaussian Mixture Model (GMM)

- For a particular component (value of **z**)

$$p(\mathbf{x}|z_k = 1) = \mathrm{N}(x \mid \boldsymbol{\mu}_k, \Sigma_k)$$

- Thus **p(x | z)** can be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathrm{N}(x \mid \boldsymbol{\mu}_k, \Sigma_k)^{z_k}$$



- Using the one-hot property of **z**, only the term corresponding to the active component ($\mathbf{z_k} = 1$) contributes to the product.

Gaussian Mixture Model (GMM)

- The joint distribution **p(x,z)** is equal to **p(z)p(x | z)**
- Based on law of total probability, marginal distribution of **x** is obtained by summing over all possible states of **z** to give

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{z} \prod_{k=1}^{K} \pi_k^{z_k} N\left(\boldsymbol{x} \mid \mu_k, \Sigma_k\right)^{z_k} = \sum_{k=1}^{K} \pi_k N\left(\boldsymbol{x} \mid \mu_k, \Sigma_k\right)$$

  - Since $\mathbf{z_k} \in \mathbf{\{0, 1\}}$
- This is the standard form of a Gaussian Mixture

Gaussian Mixture Model (GMM)

- If we have observations $\mathbf{x_1}, \ldots \mathbf{x_N}$
- Because marginal distribution is in the form: $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$
  - Every observed data point $\mathbf{x_n}$ there is a corresponding latent vector $\mathbf{z_n}$, i.e., its sub-class
- Thus, we have found a formulation of GMM involving a latent variable meaning that we can directly incorporate the relationships between the $\mathbf{x_n}$ and $\mathbf{z_n}$ (i.e., component memberships).
  - This approach simplifies parameter estimation by working with $\boldsymbol{p(x,z)}$ rather than $\boldsymbol{p(x)}$

Gaussian Mixture Model (GMM)

- Another conditional probability is called **Responsibility**, which is is the **posterior probability $p(z|x)$** of a data point x being associated with a particular component k. $p(z_k = 1|\mathrm{x})$ is denoted $\gamma(z_k)$
- From Bayes theorem

$$p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)\, p(\mathbf{x}|z_k = 1)}{p(\mathbf{x})}.$$

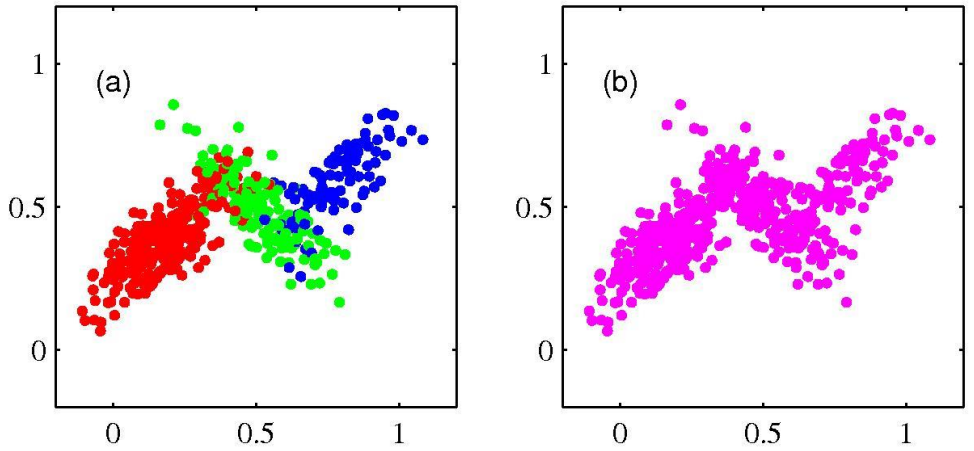
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) =$$

$$= \frac{p(z_k = 1)\, p(\mathbf{x}\,|\, z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)\, p(\mathbf{x}\,|\, z_j = 1)} = \frac{\pi_k \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)}$$

- π_k can be viewed as the **prior probability** of $z_k = 1$
- $\gamma(z_k)$ can be interpreted as the ***responsibility*** that component ***k*** takes for 'explaining' the observation **x.**

Getting data from GMM Synthetically

- By generating synthetic data, you can compare model predictions against the true data distribution, which is useful for evaluating the model's performance, diagnosing issues, and fine-tuning parameters.
- Two ways:
  - We can generate a value of $\mathbf{z}^*$ from $p(\mathbf{z})$, and then, generate a value for $\mathbf{x}$ from $p(\mathbf{x}|\mathbf{z})$ → plot (a) in the figure
  - Generate from $p(\mathbf{x})$ by ignoring the values of $\mathbf{z}$ → plot (b) in the figure



Maximum Likelihood For GMM

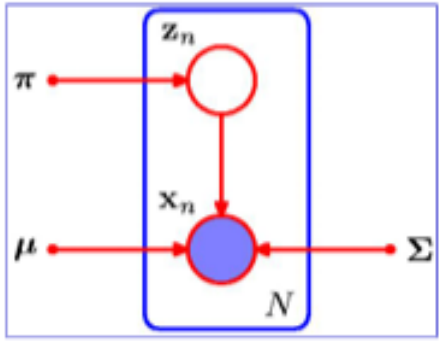
- We would like to model data set $\{x_1,.. x_N\}$ using a mixture of Gaussians (N items each of dimension D).
- **X** is represented by a N x D matrix
  - n^{th} row is given by x_n^T

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_N \end{bmatrix}$$

- **Z** represents N latent variables with N x K matrix
  - n^{th} row is given by z_n^T

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \\ z_N \end{bmatrix}$$

- The goal is to estimate **means**, **covariances**, and **mixing coefficients)** by maximizing the likelihood of observing the dataset X. Maximizing the likelihood involves adjusting the parameters so that the probability of observing the data under the model is as high as possible.



Likelihood Function For GMM

- Given mixture density function

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{z} \prod_{k=1}^{K} \pi_k^{z_k} N\left(\boldsymbol{x} \mid \mu_k, \Sigma_k\right)^{z_k} = \sum_{k=1}^{K} \pi_k N\left(\boldsymbol{x} \mid \mu_k, \Sigma_k\right)$$

- The joint likelihood for the dataset X is

$$p(X \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

The product is over the N Independent and identically distributed samples

Since the data points are independently drawn.

- Log-likelihood function is

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

We need to maximize this

Log-likelihood Function for GMM

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- The goal is to estimate the three sets of parameters:

$$\pi_k, \mu_k, \Sigma_k$$

  - We need to take derivatives w.r.t. each while keeping the others constant
  - However, there are no closed-form solutions (check the formulas to see why)
    - Task is not straightforward since summation appears in Gaussian and logarithm does not operate on Gaussian

- A gradient-based optimization is possible,
  - we consider the iterative EM algorithm

- Before proceeding with the Expectation Maximization (Maximum Likelihood Estimation) we need to mention two technical issues:

  - Problem of singularities with Gaussian mixtures
    - Singularities occur when one or more of the components of the mixture model become overly narrow, essentially collapsing to a single point or becoming degenerate (i.e., zero variance).

  - Problem of Identifiability of mixtures
    - A model is said to be identifiable if, for a given set of observed data, there is a unique set of model parameters that can explain the data.
    - If a model is non-identifiable, it means that there are multiple different sets of parameters that could explain the same data, leading to ambiguity in the model's interpretation.

Singularities Problem

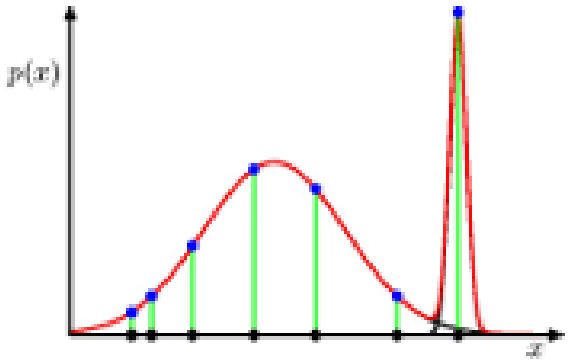
- Consider a Gaussian mixture
  - Having components with covariance matrices $\Sigma_k = \sigma_k{}^2 I$
- Data point that falls on a mean $\mu_j = x_n$ contribute to the likelihood function

$$N(\boldsymbol{x}_n \mid \boldsymbol{x}_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

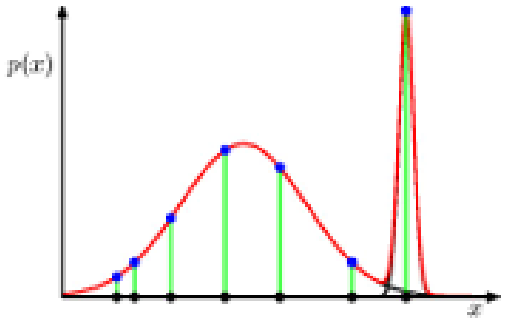
since $\exp(x_n - \mu_j)^2 = 1$ *(see Appendix – a for the proof)*

  - As σ_k -> 0 terms goes to infinity
  - Thus, maximization of log-likelihood is not well-posed

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$



Singularities Problem



- One component assigns finite values because its variance is almost zero.
- Instead, the other component has a large variance.
- The solution in the case of σ_k -> 0 is resetting the mean or covariance to a **small but nonzero value.**

Identifiability problem

- A density $p(x|\theta)$ is identifiable if $\theta \neq \theta'$ then there is an x for which $p(x|\theta) \neq p(x|\theta')$
- A K-component mixture will have a total of K! equivalent solutions
  - Corresponding to K! ways of assigning K sets of parameters to K components
    - E.g., for K=3 K!=6: 123, 132, 213, 231, 312, 321
  - For any given point in the space of parameter values there will be a further K!-1 additional points all giving the same distribution
- However, any of the equivalent solutions are as good as the other
- The EM algorithm can help address the problem of identifiability by iteratively refining the estimates of the parameters, although it does not necessarily solve the issue of multiple equivalent solutions.
  - To avoid non-identifiable solutions, GMM parameters can be initialized using K-means, providing a better starting point for EM.

Expectation-Maximization (EM) for GMM

UNIVERSITÀ di VERONA

- EM is a method for finding maximum likelihood solutions for models with latent variables
- We begin with Log-likelihood function

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- We want to find $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ that maximizes this function
- This task is not straightforward since summation appears in Gaussian and logarithm does not operate on Gaussian

Expectation-Maximization (EM) for GMM

- $$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- Take derivatives one by one
  - Means $\boldsymbol{\mu}_k$ and set to zero
  - Covariance matrices Σ_k and set to zero
  - mixing coefficients π_k, and set to zero

Expectation-Maximization (EM) for GMM

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- Take derivative w.r.t. the means $\boldsymbol{\mu}_k$ and set to zero
  - When you use the exponential form of Gaussian and then use the formulas $\frac{d}{dx} \ln u = \frac{u'}{u}$ and $\frac{d}{dx} e^u = e^u u'$
  - We obtain:

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(\boldsymbol{x}_n \mid \mu_j, \Sigma_j)} \Sigma_k^{-1} (\boldsymbol{x}_n - \mu_k)$$

Inverse of covariance matrix

$\gamma(z_{nk})$ the posterior probabilities

Expectation-Maximization (EM) for GMM

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(\boldsymbol{x}_n \mid \mu_j, \Sigma_j)} \Sigma_k^{-1} (\boldsymbol{x}_n - \mu_k)$$

Inverse of covariance matrix

$\gamma(z_{nk})$ the posterior probabilities

- Mean of k^{th} Gaussian component is the weighted mean of <u>all</u> the points in the data set:
  - where data point x_n is weighted by the posterior probability that component k was responsible for generating x_n
- Consequently, rearranging the above formula and multiplying Σ_k

Expectation-Maximization (EM) for GMM

UNIVERSITÀ di VERONA

- We obtain the maximum likelihood estimation for **means** as:

(refer to Appendix –B for mathematical derivations)

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

- The effective number of points assigned to *cluster k (mixture k)*:

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

Expectation-Maximization (EM) for GMM

UNIVERSITÀ di VERONA

- Maximum likelihood solution for **covariance**:
  - Take derivative w.r.t. the means Σ_k and set to zero

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^T$$

  - Notice that each data point is **weighted** by the corresponding **posterior probability**!!!!

  - The denominator is the same as in the **mean**, i.e., the number of data points in a component.

Expectation-Maximization (EM) for GMM

- Maximize $\ln p(X \mid \pi,\mu,\Sigma)$ w.r.t. π_k
  - Must consider that mixing coefficients sum to one
  - To solve this, we need to use Lagrange multiplier and maximizing

$$\ln p(X \mid \pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

  - And then set derivative wrt. Π_k to zero, giving us:

$$\pi_k = \frac{N_k}{N}$$

Expectation-Maximization (EM) for GMM

UNIVERSITÀ di VERONA

- GMM maximum likelihood parameter estimates:

Means:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

Covariance matrices:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^T$$

Mixing Coefficients:

$$\pi_k = \frac{N_k}{N}$$

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

- All three are in terms of responsibilities therefore we have not completely solved the problem.

EM formulation for GMM

- The results for $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ are not closed for solutions for the parameters since γ_{nk} (responsibilities) depend on those parameters

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- This results in building an iterative solution, which is an instance of EM algorithm for GMM.

EM formulation for GMM

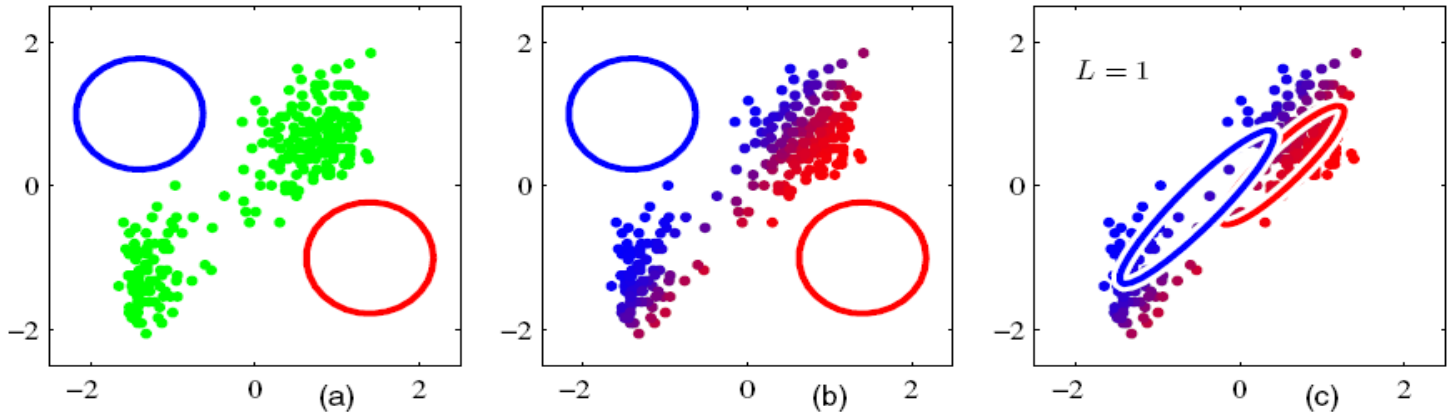
- Initialize the parameters of GMM: means, covariances and mixing coefficients
- Apply two updated: **E step** & **M step**
  - **E step:** use the current value of parameters to evaluate posterior probabilities or responsibilities
  - **M step:** use these posterior probabilities to re-estimate means, covariances, and mixing coefficients

EM formulation for GMM

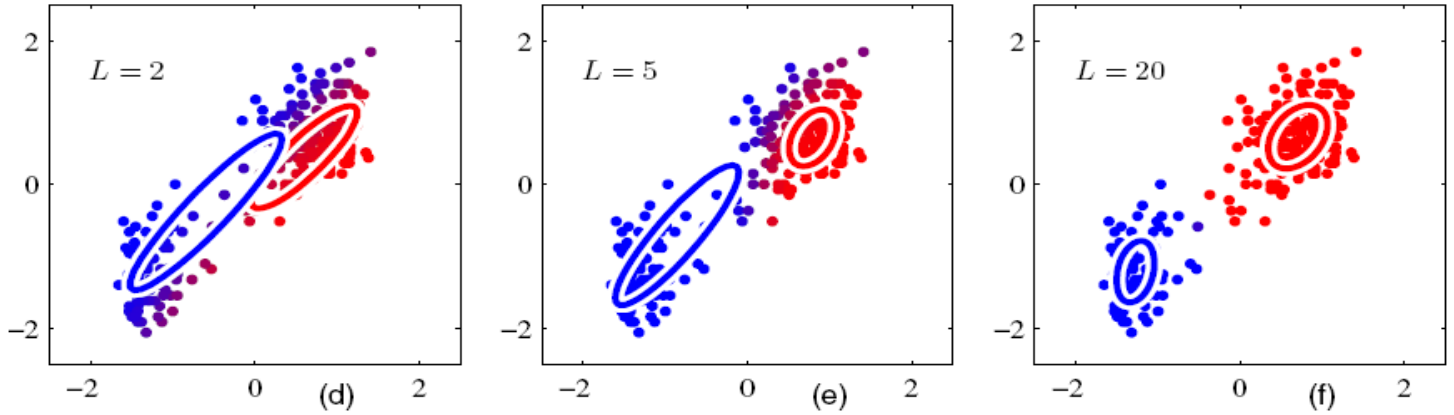
Data points and initial mixture model

Initial E step: determine responsibilities

After the first M step re-evaluate parameters



After 2, 5 and 20 cycles ➔



EM FORMULATION FOR GMM

- Given a Gaussian mixture model
- Goal is to maximize the likelihood function w.r.t. the parameters (means, covariances and mixing coefficients)
- **Step1:** Initialize the $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ and evaluate the initial value of log-likelihood

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- **Step 2: E-step:** Evaluate responsibilities using current parameter values

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM FORMULATION FOR GMM

- **Step 3: M-step:** Re-estimate parameters using current responsibilities

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N} \quad \text{where} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

EM formulation for GMM

- **Step 4:** Evaluate the **log-likelihood**

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\}$$

- And check the convergence of either **parameters** or **log-likelihood**
- If the convergence is not satisfied GOTO Step 2

Properties of EM for GMM

- Takes many **more iterations** than **K-means**
  - Each cycle requires significantly more comparison
- Common to run **K-means** first to find **suitable initialization**
- Covariance matrices can be initialized to **covariances** of **clusters** found by **K-means**
- **EM** is not guaranteed to find the **global maximum** of the log-likelihood function

Example

https://colab.research.google.com/drive/1hn0UuluAxRPzNAMZmEyLPsJQ6tT9vRHz?usp=sharing

GMM SUMMARY

UNIVERSITÀ di VERONA

- A probabilistic view of unsupervised learning (even clustering).
- Each cluster corresponds to a **different Gaussian**.
- Model using **latent variables**.
- General approach, can eventually replace Gaussian with other distributions (continuous or discrete) .
- More generally, mixture model are very powerful models, universal approximator.
- Optimization is done using the **EM algorithm**.

Appendix –A

- In a Gaussian Mixture Model (GMM), when a data point x_n falls exactly on the mean μ_j of a particular component j, it contributes to the likelihood function through the probability density function of the corresponding Gaussian distribution.
- The pdf of a Gaussian distribution at a specific point x is given by:

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

- For a GMM component j with mean μ_j and covariance matrix Σ_j, the pdf at x_n is:

$$\mathcal{N}(x_n|\mu_j,\Sigma_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}} e^{-\frac{1}{2}(x_n-\mu_j)^T\Sigma_j^{-1}(x_n-\mu_j)}$$

Appendix - A

$$\mathcal{N}(x_n|\mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}} e^{-\frac{1}{2}(x_n-\mu_j)^T\Sigma_j^{-1}(x_n-\mu_j)}$$

- where $|\Sigma_j|$ denotes the determinant of the covariance matrix Σ_j.

Σ_j^{-1} is the inverse of the covariance matrix.

d is the dimensionality of the data space.

- When x_n falls exactly on the mean μ_j, the term $(x_n-\mu_j)$ becomes zero in the exponent. This means that the exponent reduces to zero, and $e^0=1$. Consequently, the likelihood contribution from x_n to the Gaussian component j becomes:

$$\mathcal{N}(x_n|\mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}}$$

Appendix - B

- The log-likelihood function for a Gaussian Mixture Model (GMM) is given by:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \sum_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

- First, let's differentiate the log-likelihood function w.r.t $\boldsymbol{\mu}_k$:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \sum_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

$$= \frac{1}{\sum_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)} \sum_{n=1}^{N} \left(\frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right)$$

- Next, we differentiate the term $\pi_k \mathrm{N}(\mathbf{x}_n|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with respect to $\boldsymbol{\mu}_k$:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) = \pi_k \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Appendix - B

- The derivative of the Gaussian distribution with respect to the mean vector $\boldsymbol{\mu}_k$ is:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = -\tfrac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

- Substituting this back into our expression, we get:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)}\sum_{n=1}^{N}\left(-\pi_k\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)$$

- Now, we set this derivative equal to zero and solve for $\boldsymbol{\mu}_k$ to find the stationary points:

$$\frac{1}{\sum_{n=1}^{N}\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)}\sum_{n=1}^{N}\left(-\pi_k\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) = 0$$

Appendix - B

- Solving this equation for $\boldsymbol{\mu}_k$ generally requires numerical optimization methods due to the complexity of the expression and the non-linearity introduced by the Gaussian density. The resulting solution will be the MLE estimate for $\boldsymbol{\mu}_k$.
- The resulting solution for $\boldsymbol{\mu}_k$ in the context of the EM algorithm for GMMs involves a weighted average of the data points assigned to the ***k*-th** component, where the weights are determined by the **responsibilities** of each component for each data point.
- Specifically, the solution for $\boldsymbol{\mu}_k$ is given by:
- γ_{nk} represents the responsibility of the k-th component for the n-th data point.

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}$$

Cigdem Beyan
cigdem.beyan@univr.it
https://cbeyan.github.io/