

# Mixtures

---

Course of Machine Learning  
Master Degree in Computer Science  
University of Rome "Tor Vergata"

Giorgio Gambosi

a.a. 2017-2018

## Linear combinations of probability distributions $q(x|\theta)$

- Same type of distributions
- Differ by parameter values

$$p(x|\boldsymbol{\psi}) = p(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k q(x|\theta_k)$$

where

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \qquad \boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \qquad \boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\pi})$$

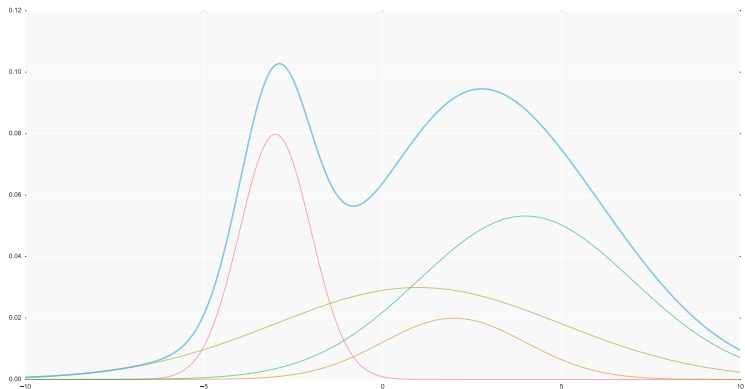
## Mixing coefficients

$$0 \leq \pi_k \leq 1 \quad k = 1, \dots, K \qquad \sum_{k=1}^K \pi_k = 1$$

Terms  $\pi_k$  have the properties of probability values

# Mixtures of distributions

Provide extensive capabilities to model complex distributions. For example, almost all continuous distributions can be modeled by the linear combination of a suitable number of gaussians.



## Mixture parameters estimation

Given a dataset  $\mathbf{X} = (x_1, \dots, x_n)$ , the parameters  $\boldsymbol{\pi}, \boldsymbol{\theta}$  of a mixture can be estimated by maximum likelihood.

$$L(\boldsymbol{\psi}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\psi}) = \prod_{i=1}^n p(x_i|\boldsymbol{\psi}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k q(x|\theta_k)$$

or maximum log-likelihood

$$l(\boldsymbol{\psi}|\mathbf{X}) = \log p(\mathbf{X}|\boldsymbol{\psi}) = \sum_{i=1}^n \log p(x_i|\boldsymbol{\psi}) = \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right)$$

## Mixture parameters estimation

Let us derive the set of derivatives for  $j = 1, \dots, K$  and set them to 0

$$\frac{\partial l(\psi|\mathbf{X})}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left[ \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] = 0$$
$$\frac{\partial l(\psi|\mathbf{X})}{\partial \pi_j} = \frac{\partial}{\partial \pi_j} \left[ \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] = 0$$

which itself results, for  $k = 1, \dots, K$ , into

$$\pi_k = \frac{1}{n} \sum_{i=1}^n \gamma_k(x_i) \quad \sum_{i=1}^n \gamma_k(x_i) \frac{\partial \log q(x_i|\theta_k)}{\partial \theta_k} = 0$$

where

$$\gamma_k(x) = \frac{\pi_k q(x|\theta_k)}{\sum_{j=1}^K \pi_j q(x|\theta_j)}$$

## Mixture parameters estimation

The constraint  $\sum_{i=1}^K \pi_i = 1$  can be taken into account by introducing a Lagrange multiplier  $\lambda$  and considering the Lagrangian

$$L(\boldsymbol{\psi}, \lambda) = l(\boldsymbol{\psi}|\mathbf{X}) + \lambda(1 - \sum_{i=1}^K \pi_i)$$

Setting the derivative wrt  $\pi_j$  to 0 turns out to be equivalent to

$$\begin{aligned}\lambda &= \frac{\partial l(\boldsymbol{\psi}|\mathbf{X})}{\partial \pi_j} = \frac{\partial}{\partial \pi_j} \left[ \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] = \sum_{i=1}^n \frac{\partial}{\partial \pi_j} \left[ \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] \\&= \sum_{i=1}^n \frac{1}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \frac{\partial}{\partial \pi_j} \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \\&= \sum_{i=1}^n \frac{1}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \sum_{k=1}^K \frac{\partial}{\partial \pi_j} (\pi_k q(x_i|\theta_k)) \\&= \sum_{i=1}^n \frac{q(x_i|\theta_j)}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} = \sum_{i=1}^n \frac{\gamma_j(x_i)}{\pi_j} = \frac{1}{\pi_j} \sum_{i=1}^n \gamma_j(x_i)\end{aligned}$$

## Mixture parameters estimation

Setting the derivative wrt  $\lambda$  to 0

$$\frac{\partial}{\partial \lambda} \left( l(\boldsymbol{\psi} | \mathbf{X}) + \lambda \left( 1 - \sum_{i=1}^K \pi_i \right) \right) = 0$$

is equivalent to

$$\sum_{i=1}^K \pi_i = 1$$

Moreover, since, as shown above,

$$\pi_j = \frac{1}{\lambda} \sum_{i=1}^n \gamma_j(x_i)$$

it results

$$\sum_{j=1}^K \pi_j = \frac{1}{\lambda} \sum_{j=1}^K \sum_{i=1}^n \gamma_j(x_i) = 1$$

and

$$\lambda = \sum_{j=1}^K \sum_{i=1}^n \gamma_j(x_i) = \sum_{i=1}^n \sum_{j=1}^K \gamma_j(x_i) = \sum_{i=1}^n \sum_{j=1}^K \frac{\pi_j q(x_i | \theta_j)}{\sum_{k=1}^K \pi_k q(x_i | \theta_k)} = \sum_{i=1}^n 1 = n$$

## Mixture parameters estimation

Finally,

$$\begin{aligned}\frac{\partial l(\psi|\mathbf{X})}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \left[ \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] = \sum_{i=1}^n \frac{\partial}{\partial \theta_j} \left[ \log \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \right] \\&= \sum_{i=1}^n \frac{1}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \frac{\partial}{\partial \theta_j} \left( \sum_{k=1}^K \pi_k q(x_i|\theta_k) \right) \\&= \sum_{i=1}^n \frac{1}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \sum_{k=1}^K \frac{\partial}{\partial \theta_j} (\pi_k q(x_i|\theta_k)) \\&= \sum_{i=1}^n \frac{\pi_j}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \frac{\partial}{\partial \theta_j} q(x_i|\theta_j) \\&= \sum_{i=1}^n \frac{\pi_j q(x_i|\theta_j)}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \frac{1}{q(x_i|\theta_j)} \frac{\partial}{\partial \theta_j} q(x_i|\theta_j) \\&= \sum_{i=1}^n \frac{\pi_j q(x_i|\theta_j)}{\sum_{k=1}^K \pi_k q(x_i|\theta_k)} \frac{\partial \log q(x_i|\theta_j)}{\partial \theta_j} = \sum_{i=1}^n \gamma_j(x_i) \frac{\partial \log q(x_i|\theta_j)}{\partial \theta_j} = 0\end{aligned}$$



Log likelihood maximization is intractable analytically: its solution cannot be given in closed form.

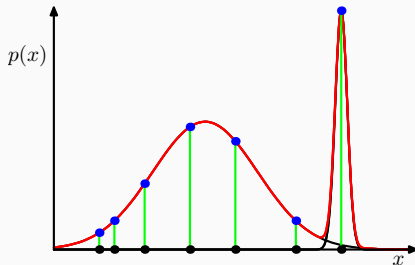
- $\boldsymbol{\pi}$  and  $\boldsymbol{\theta}$  can be derived from  $\gamma_k(x_i)$
- Also,  $\gamma_k(x_i)$  can be derived from  $\boldsymbol{\pi}$  e  $\boldsymbol{\theta}$

## Iterative techniques

- Given an estimation for  $\boldsymbol{\pi}$  e  $\boldsymbol{\theta}$ ...
- derive an estimation for  $\gamma_k(x_i)$ , from which ...
- derive a new estimation for  $\boldsymbol{\pi}$  e  $\boldsymbol{\theta}$ , from which ...
- derive a new estimation for  $\gamma_k(x_i)$  ...

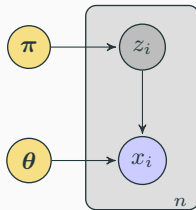
# Issues in ML for mixtures

- Identifiability: for each solution (assignment of parameters to component distributions), there exist  $K! - 1$  equivalent solutions
- Singularity: risk of severe overfitting. A mixture collapses to a single point.



# Mixtures as generative processes

Graphical model representation of a mixture of distributions.



## Latent variables

- Terms  $z_i$  are **latent** random variable with domain  $z \in \{1, \dots, K\}$
- While  $x_i$  is observed, the value of  $z_i$  cannot be observed
- $z_i$  denotes the component distribution  $q(x|\theta)$  responsible for the generation of  $x_i$

## Generation process

1. Starting from the distribution  $\pi_1, \dots, \pi_K$ , the component distribution to apply to sample the value of  $x_i$  is sampled: its index is given by  $z_i$ : hence  $z_i$  is dependent from  $\boldsymbol{\pi}$
2. Let  $z_i = k$ : then,  $x_i$  is sampled from distribution  $q(x|\theta_k)$ . That is,  $x_i$  is dependent from both  $z_i$  and  $\boldsymbol{\theta}$

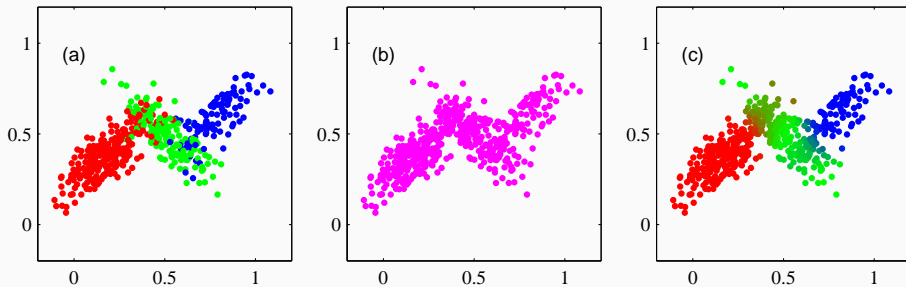
## Latent variables coding

Indeed,  $z_i$  can be seen as components of a single latent  $K$ -dimensional variable  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_K)$

1-to- $K$  coding:  $K$  possible values  $\zeta_i \in \{0, 1\}$ ,  $\sum_{i=1}^K \zeta_i$ .

# Mixtures as generative processes

Example of generation of dataset from mixture of 3 gaussians



## Distributions with latent variables

$$p(x|z = k, \boldsymbol{\psi}) = p(x|z = k, \boldsymbol{\theta}) = q(x|\theta_k)$$

Marginalizing wrt  $z$ ,

$$\begin{aligned} p(x|\boldsymbol{\psi}) &= \sum_{k=1}^K p(x, z = k|\boldsymbol{\psi}) = \sum_{k=1}^K p(x|z = k, \boldsymbol{\theta})p(z = k|\boldsymbol{\pi}) \\ &= \sum_{k=1}^K q(x|\theta_k)p(z = k|\boldsymbol{\pi}) \end{aligned}$$

Since, by definition,

$$p(x|\boldsymbol{\psi}) = \sum_{k=1}^K \pi_k q(x_i|\theta_k)$$

it results

$$p(z = k|\boldsymbol{\psi}) = p(z = k|\boldsymbol{\pi}) = \pi_k$$

## Responsibilities

An interpretation for  $\gamma_k(x)$  can be derived as follows

$$\begin{aligned}\gamma_k(x) &= \frac{\pi_k q(x|\theta_k)}{\sum_{j=1}^K \pi_j q(x|\theta_j)} \\ &= \frac{p(z=k)p(x|z=k)}{\sum_{j=1}^K p(z=j)p(x|z=j)} = p(z=k|x)\end{aligned}$$

## Mixing coefficients and responsibilities

- A mixing coefficient  $\pi_k = p(z=k)$  can be seen as the prior (wrt to the observation of the point) probability that the next point is generated by sampling the  $k$ -th component distribution
- A responsibility  $\gamma_k(x) = p(z=k|x)$  can be seen as the posterior (wrt to the observation of the point) probability that a point has been generated by sampling the  $k$ -th component distribution

## Expectation maximization for gaussian mixtures



## Data set

- Let  $\mathbf{X} = (x_1, \dots, x_n)$  be the set of values of observed variables and let  $\mathbf{Z} = (z_1, \dots, z_n)$  be the set of values of the latent variables. Then  $(\mathbf{X}, \mathbf{Z})$  is the **complete dataset**: it includes the values of all variables in the model
- $\mathbf{X}$  is the **observed dataset** (incomplete). It only includes “real” data, that is observed data.

Indeed,  $\mathbf{Z}$  is unknown. If values have been assigned to model parameters, the only possible knowledge about  $\mathbf{Z}$  is given by the posterior distribution  $p(\mathbf{Z}|\mathbf{X}, \psi)$ .

# Inferring parameters for gaussian mixtures

- If we assume that the complete dataset  $(\mathbf{X}, \mathbf{Z})$  is known (that is the observed points **together with their corresponding components**) a maximum likelihood estimation of  $\boldsymbol{\pi}$  and  $\boldsymbol{\theta}$  would be easy. In particular,
- For the mixing coefficients  $\pi_k$  it would result, as usual

$$\pi_k = \frac{n_k}{n}$$

where  $n_k$  is the number of elements of the set  $C_k$  such that  $z = k$

- For component parameters  $\theta_k = (\mu_k, \Sigma_k)$  the usual estimations for gaussians would provide

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} \mathbf{x} \\ \boldsymbol{\Sigma}_k &= \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T\end{aligned}$$

The above results derive from the maximization, wrt  $\pi_k, \mu_k, \Sigma_k$ , ( $k = 1, \dots, K$ ) of the log likelihood

$$\begin{aligned} l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) &= \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi}) = \log \prod_{i=1}^n \prod_{k=1}^K \pi_k^{z_{ik}} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}} \\ &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)) \end{aligned}$$

## Dealing with latent variables

Unfortunately, since  $\mathbf{Z}$  is unknown, the log-likelihood of the complete dataset cannot be defined (the sets  $C_k$  are not known).

Our approach will be to consider for maximization, instead of:

- the log-likelihood where each  $z_i$  is specified

its expectation wrt to the conditional distribution  $p(\mathbf{Z}|\mathbf{X})$ , that is

- the expectation of the log-likelihood with respect to probabilities  $p(z|x_i)$  is specified

$$\begin{aligned} E[l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z})] &= \sum_{i=1}^n \sum_{k=1}^K p(z_i = k|\mathbf{x}_i)(\log \pi_k + \log q(\mathbf{x}_i|\theta_k)) \\ &= \sum_{i=1}^n \sum_{k=1}^K \gamma_k(\mathbf{x}_i)(\log \pi_k + \log q(\mathbf{x}_i|\theta_k)) \end{aligned}$$

Observe that this expectation can be derived if  $p(\mathbf{Z}|\mathbf{X})$ , that is the set of all values  $\gamma_k(\mathbf{x}_i)$ , is known.

## Maximization of expected log-likelihood

The maximization of  $E[l(\psi|\mathbf{X}, \mathbf{Z})]$  wrt to  $\pi_k, \mu_k, \Sigma_k$  results easily into

$$\pi_k = \frac{1}{n} \sum_{i=1}^n \gamma_k(\mathbf{x}_i)$$

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^n \gamma_k(\mathbf{x}_i) \mathbf{x}_i$$

$$\Sigma_k = \frac{1}{n_k} \sum_{i=1}^n \gamma_k(\mathbf{x}_i) (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T$$

this is named **M-step** (from “Maximization”)

The computed values for the parameters result into new, different values for  $\gamma_k(x_i) = p(z_i = k|x_i)$ , and a different expectation  $E[l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z})]$ .

$$p(z_i = k|x_i) = \frac{p(z_i = k)p(x_i|z_i = k)}{\sum_{j=1}^K p(z_i = j)p(x_i|z_i = j)} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

this is named **E-step** (from “Expectation”)

# ML and mixtures of gaussians: iterative approach

1. Assign an initial estimate to  $\mu_j, \Sigma_j, \pi_j, j = 1, \dots, K$
2. Repeat
  - 2.1 Compute

$$\gamma_j(x_i) = \frac{1}{\gamma_i} \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j) \quad \text{con} \quad \gamma_i = \sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)$$

- 2.2 Compute

$$\pi_j = \frac{n_j}{n} \quad \text{con} \quad n_j = \sum_{i=1}^n \gamma_j(x_i)$$

- 2.3 Compute

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^n \gamma_j(x_i) x_i$$

- 2.4 Compute

$$\Sigma_j = \frac{1}{n_j} \sum_{i=1}^n \gamma_j(x_i) (x_i - \mu_j)(x_i - \mu_j)^T$$

3. until some convergence property is verified

The convergence test may refer to the the increase of log-likelihood in the last iteration