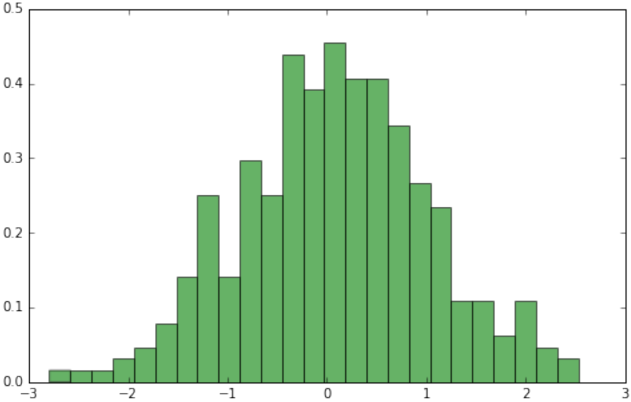
Mixture density models

Victor Kitov

Table of Contents

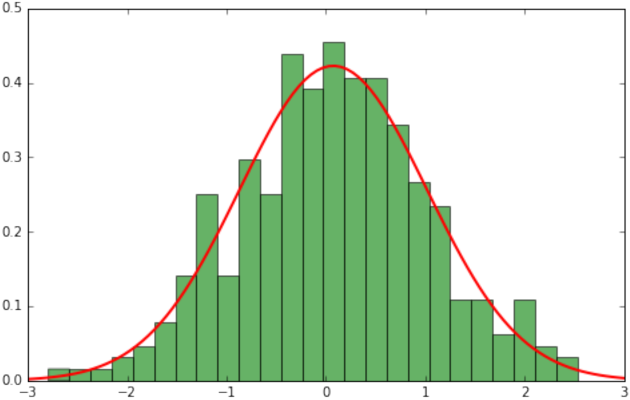
Sample density

Consider sample density:



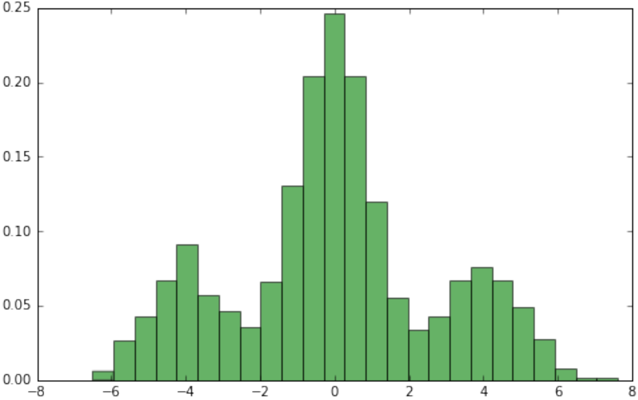
Parametric density approximation

It can be accurately modelled with existing parametric family - Normal



Non-standard sample distribution

What to do if no parametric model fits well?



Mixture models

$$p(x) = \sum_{z=1}^{Z} \phi_z p(x;\theta_z)$$

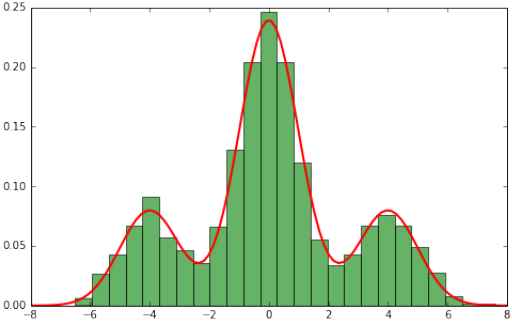
- Z - number of components
- ϕ_z, $z = 1, 2, ...Z$ - mixture component probabilities, $\phi_z \geq 0$, $\sum_{z=1}^{Z} \phi_z = 1$
- $p(x;\theta_z)$ - component density functions
- Parameters of mixture model $\Theta = \{\phi_z, \theta_z,\ z = 1, 2, ...Z\}$

$p(x,\theta_z)$ may be of single or different parametric families.

Mixture of Gaussians

Gaussians model continious r.v. on $(-\infty, +\infty)$.
$p(x, \theta_z) = N(x, \mu_z, \Sigma_z)$, $\theta_z = \{\mu_z, \Sigma_z\}$.

$$p(x) = \sum_{z=1}^{Z} \phi_z N(x, \mu_z, \Sigma_z) \quad (1)$$



Mixtures of other distributions

Mixture of random variables:

- continious, distributed on $(-\infty, +\infty)$
- continious, distributed on $[a, \infty)$
- continious, distributed on $[a, b]$
- discrete, distributed on $[a, \infty)$
- discrete, distributed on $[a, b]$

Mixtures of other distributions

Mixture of random variables:

- continious, distributed on $(-\infty, +\infty)$
  - Normal, Laplace, Student
- continious, distributed on $[a, \infty)$
  - Gamma
- continious, distributed on $[a, b]$
  - Beta
- discrete, distributed on $[a, \infty)$
  - Poisson
- discrete, distributed on $[a, b]$
  - Binomial

Sampling from mixture

1. Sample mixture component z with random probabilities $\phi_1, \phi_2, ...\phi_Z$

Sampling from mixture

1. Sample mixture component z with random probabilities $\phi_1, \phi_2, ...\phi_Z$
   - to do that we sample $u \sim Uniform[0,1]$ and select component z if $\sum_{k=1}^{z-1} \phi_k < u \leq \sum_{k=1}^{z} \phi_z$
2. Sample observation $x \sim p(x|\theta_z)$

Classification using mixtures

Model within class probability with mixtures:

$$p(x|y) = \sum_{z=1}^{Z_y} \phi_{y,z} p(x; \theta_{y,z})$$

where Z_y, $\pi_{y,z}$ and $p(x; \theta_{y,z})$ are specific for each class y.

Bayes decision rule:

$$\widehat{y} = \arg\max_y \lambda_y p(y) p(x|y)$$

λ_y - cost for misclassifying class y
$p(y)$ - prior for class y
$p(x|y)$ - within class probability

EM-algorithm for normal mixtures

Initialize ϕ_j, μ_j and Σ_j, $j = 1, 2, ...g$.

repeat while stopping condition not satisfied:

E-step. Calculate correspondences of x_n to component z:

for $n = 1, 2, ...N$:

for $z = 1, 2, ...Z$:

$$w_{nz} = \frac{\phi_z N(x_n; \mu_z, \Sigma_z)}{\sum_k \phi_k N(x_n; \mu_k, \Sigma_k)} \quad \text{\# =p(z|x(n))}$$

M-step. Update component parameters:

for $z = 1, 2, ...Z$:

$$\widehat{\phi}_z = \frac{1}{N} \sum_{n=1}^{N} w_{nz}$$

$$\widehat{\mu}_z = \frac{\sum_{n=1}^{N} w_{nz} x_n}{\sum_{n=1}^{N} w_{nz}}$$

$$\widehat{\Sigma}_z = \frac{1}{\sum_{n=1}^{N} w_{nz}} \sum_{n=1}^{N} w_{nz} \left(x_n - \widehat{\mu}_z\right)\left(x_n - \widehat{\mu}_z\right)^T$$

Interpretation

$$\begin{aligned} w_{nz} &= P(z|x_n) = \frac{P(z,x_n)}{P(x_n)} = \frac{P(z,x_n)}{\sum_k P(k,x_n)} = \\ &= \frac{P(z)P(x_n|z)}{\sum_k P(k)P(x_n|k)} = \frac{\widehat{\phi}_z N(x_n;\widehat{\mu}_z,\widehat{\Sigma}_z)}{\sum_k \widehat{\phi}_k N(x_n;\widehat{\mu}_k,\widehat{\Sigma}_k)} \end{aligned}$$

$\widehat{\phi}_z, \widehat{\mu}_z, \widehat{\Sigma}_z$ are weighted averages, weighted with $w_{nz} = P(z|x_n)$:

$$\widehat{\phi}_z = \frac{1}{N}\sum_{n=1}^{N} w_{nz} \qquad \widehat{\mu}_z = \frac{\sum_{n=1}^{N} w_{nz}x_n}{\sum_{n=1}^{N} w_{nz}}$$

$$\widehat{\Sigma}_z = \frac{1}{\sum_{n=1}^{N} w_{nz}}\sum_{n=1}^{N} w_{nz}\left(x_n - \widehat{\mu}_z\right)\left(x_n - \widehat{\mu}_z\right)^T \tag{2}$$

Table of Contents

K-means algorithm

- Suppose we want to cluster our data into g clusters.
- Cluster i has a center μ_i, i=1,2,...g.
- Consider the task of minimizing

$$\sum_{n=1}^{N} \rho(x_n, \mu_{z_n})^2 \to \min_{z_1,...z_N,\mu_1,...\mu_g} \tag{3}$$

  where $z_i \in \{1, 2, ...g\}$ is cluster assignment for x_i and $\mu_1, ...\mu_g$ are cluster centers.
- Direct optimization requires full search and is impractical.
- K-means is a suboptimal algorithm for optimizing (3).

K-means algorithm

```
Initialize μ_j, j = 1,2,...g   # usually by setting them
                               # to randomly chosen x(n)
repeat while stop condition not satisfied:
    for i = 1,2,...N:   # cluster assignements
        z_i = arg min_{j∈{1,2,...g}} ||x_i − μ_j||
    for j = 1,2,...g:   # means recalculation
        μ_j = 1/(Σ_{n=1}^N I[z_n=j]) Σ_{n=1}^N I[z_n = j] x_i
```

$$z_i = \arg\min_{j\in\{1,2,...g\}} ||x_i - \mu_j||$$

$$\mu_j = \frac{1}{\sum_{n=1}^{N}\mathbb{I}[z_n=j]}\sum_{n=1}^{N}\mathbb{I}[z_n=j]x_i$$

Possible stop conditions:

- cluster assignments $z_1, ...z_N$ stop to change (typical)
- maximum number of iterations reached
- cluster means $\{\mu_i,\ i = 1, 2, ...g\}$ stop changing significantly

K-means versus EM clustering

K-means versus EM clustering

- For each x_n EM algorithm gives $w_{nz} = p(z|x_n)$.
- This is soft or probabilistic clustering into Z clusters, having priors $\phi_1, ...\phi_Z$ and probability distributions $p(x;\theta_1), ...p(x;\theta_Z)$.
- We can make it hard clustering using $z_n = \arg\max_z w_{nz}$.

- EM clustering becomes K-means clustering when:
  - applied to Gaussians
  - with equal priors
  - with unity covariance matrices
  - with hard clustering

Initialization for Gaussian mixture EM

1. Fit K-means to $x_1, x_2, ...x_N$, obtain cluster centers μ_z, $z = 1, 2, ...Z$ and cluster assignments $z_1, z_2, ...z_N$.
2. Initialize mixture probabilities

$$\widehat{\phi}_z \propto \sum_{n=1}^{N} \mathbb{I}[z_n = z]$$

3. Initialize Gaussian means with cluster centers μ_z, $z = 1, 2, ...Z$.
4. Initialize Gaussian covariance matrices with

$$\widehat{\Sigma}_z = \frac{1}{\sum_{n=1}^{N} \mathbb{I}[z_n = z]} \sum_{n=1}^{N} \mathbb{I}[z_n = z] \left(x_n - \mu_z\right) \left(x_n - \mu_z\right)^T$$

Properties of EM

- Many local optima exist
    - in particular likelihood$\to \infty$ when $\mu_z = x_i$ and $\sigma_z \to 0$
- Only local optimum is found with EM
- Results depends on initialization
    - It is common to run algorithm multiple times with different initializations and then select the result maximizing the likelihood function.
- Number of components may be selected with:

Properties of EM

- Many local optima exist
  - in particular likelihood$\to \infty$ when $\mu_z = x_i$ and $\sigma_z \to 0$
- Only local optimum is found with EM
- Results depends on initialization
  - It is common to run algorithm multiple times with different initializations and then select the result maximizing the likelihood function.
- Number of components may be selected with:
  - cross-validation on the final task
  - out-of-sample maximum likelihood
  - statistical tests, heuristics, such as AIC/BIC information criteria

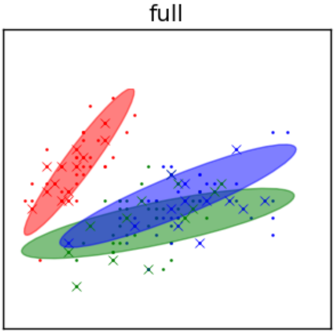
Table of Contents

Simplifications of Gaussian mixtures

- $\Sigma_z \in \mathbb{R}^{DxD}$ requires $\frac{D(D+1)}{2}$ parameters.
- Covariance matrices for Z components require $Z\frac{D(D+1)}{2}$ parameters.
- Components can be poorly identified when
    - $Z\frac{D(D+1)}{2}$ is large compared to N
    - when components are not well separated
- In these cases we can impose restrictions on covariance matrices.

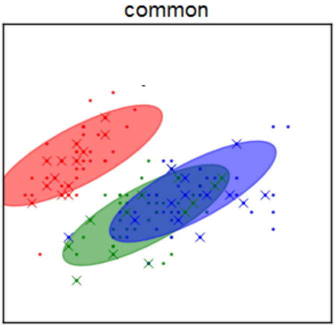
Unrestricted covariance matrices

- full covariance matrices Σ_z, $z = 1, 2, ...Z$.



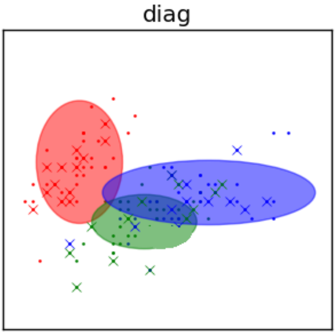
Common covariance matrix

- $\Sigma_1 = \Sigma_2 = ... = \Sigma_Z$



Diagonal covariance matrices

- $\Sigma_z = \text{diag}\{\sigma^2_{z,1}, \sigma^2_{z,2} ... \sigma^2_{z,D}\}$



Spherical matrices

- $\Sigma_z = \sigma_z^2 I$, $I \in \mathbb{R}^{DxD}$ - identity matrix