Machine Learning

Gaussian Mixture Models

Wojciech Czech

January 21, 2025

Institute of Computer Science

Table of contents

1. Mixture Models

2. Expectation Maximization (EM) for Gaussian Mixture Models (GMMs)

3. Mixtures of Bernoulli distributions

Mixture Models

Fitting data by Maximum Likelihood

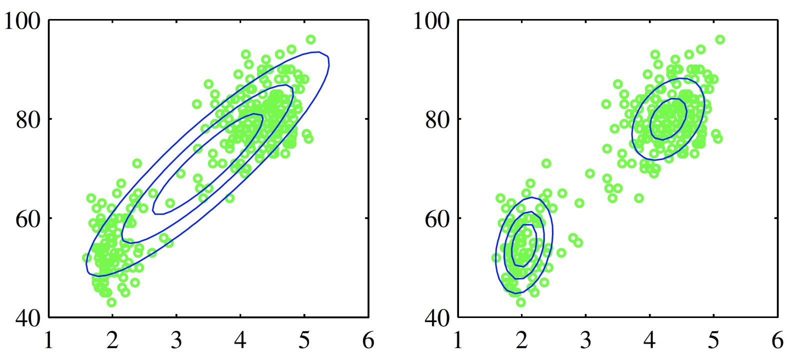


Figure 1: Old Faithful geyser eruption data [1].

Superposition of Gaussians

Superposition/Mixture of Gaussians

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

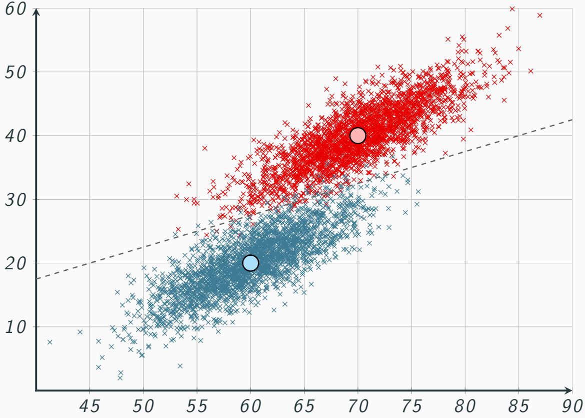
$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ - component of the mixture

π_k - mixing coefficients ($0 \leqslant \pi_k \leqslant 1$)

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

Mixing coefficients can be interpreted as probabilities

Mixture of two 2D Gaussian distributions



By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy.

Log likelihood function

- Parameters of Gaussian mixture distribution: $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$, $\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K]$
- Observed data: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

Responsibilities

- From product and sum rules:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}|k) \tag{4}$$

- From Eq. (34) and Eq. (4):

  $p(k) = \pi_k$ (prior probability of picking k-th component)
  $p(\boldsymbol{x}|k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (probability of $\boldsymbol{x}$ conditioned on k)
  $p(k|\boldsymbol{x})$ (posterior probabilities - *responsibilities*)

Responsibility

$$\gamma_k(\boldsymbol{x}) = p(k|\boldsymbol{x}) = \frac{p(k)p(\boldsymbol{x}|k)}{\sum_l p(l)p(\boldsymbol{x}|l)} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{5}$$

K-dimensional binary random variable

- Let $\boldsymbol{z}$ be K-dimensional binary random variable, having 1-of-K representation, in which a particular z_i is equal to 1 and all other elements are equal 0.

  $z_i \in \{0, 1\}$, $\sum_i z_i = 1$

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{6}$$

$$p(z_k = 1) = \pi_k \tag{7}$$

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{8}$$

$$p(\boldsymbol{x}|z_k = 1) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{10}$$

Marginal distribution and mixture model

- For every observed point $\boldsymbol{x}_i$ there is a corresponding latent variable $\boldsymbol{z}_i$

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11}$$

Responsibility as a posterior probability

$$\gamma(z_k) \equiv p(z_k = 1|\boldsymbol{x}) = \frac{p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\boldsymbol{x}|z_j = 1)} \tag{12}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{13}$$

- π_k is interpreted as prior probability of $z_k = 1$
- $\gamma(z_k)$ is interpreted as the corresponding posterior probability once we have observed $\boldsymbol{x}$ (*responsibility* that component k takes for explaining the observation $\boldsymbol{x}$)

Fitting data by Maximum Likelihood

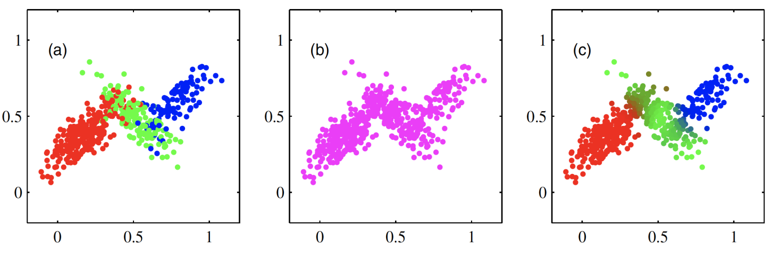


Figure 2: Samples from Gaussian mixture with three components [1]. Scatterplot (c) shows colors assigned based on *responsibility* $\gamma(z_{ik})$.

Gaussian Mixtures - Maximum Likelihood

- Dataset of observations: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_i \in \mathbb{R}^{D \times 1}$
- $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ - matrix of observations, rows of $\boldsymbol{X}$ are given by $\boldsymbol{x}_i^T$
- $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$ - matrix of latent variables, rows of $\boldsymbol{Z}$ are given by $\boldsymbol{z}_i^T$
- Data points are drawn independently from the distribution (i.i.d.)

$$p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{\boldsymbol{k}}) \tag{14}$$

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{\boldsymbol{k}}) \right\} \tag{15}$$

Pitfalls of maximizing GMM log likelihood function (1)

Singularities

- Let $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$
- j-th component has its mean $\boldsymbol{\mu}_j$ equal to one of data points: $\boldsymbol{\mu}_j = \boldsymbol{x}_n$ for some value n
- Single data point contribution to log likelihood: $\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{x}_n, \sigma_k^2 \boldsymbol{I}) = \frac{1}{(2\pi)^{\mathbf{1/2}}} \frac{1}{\sigma_j}$
- If $\sigma_j \longrightarrow 0$, then log likelihood goes to ∞
- Component shrinks onto one specific data point and contributes an ever increasing additive value to the log likelihood (severe over-fitting).

Pitfalls of maximizing GMM log likelihood function (2)

Identifiability

The parameters of the mixture components can be permuted without changing the overall likelihood or probability distribution generated by the model. As a result it is impossible to uniquely assign a specific label to a specific Gaussian component.

- permutation ambiguity - the order of the components is not not identifiable (depend e.g. on `random_state`)
- parameter space overlaps

No closed form solution

Equation (31) contains summation inside logarithm - setting derivatives to 0 does not help in maximization. Gradient-based optimization techniques have to be used.

Expectation Maximization (EM) for Gaussian Mixture Models (GMMs)

Expectation Maximization

The method for finding maximum likelihood solutions for models with latent variables

$$U = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma_k}) \right\} \tag{16}$$

Conditions satisfied by LLF at maximum: μ_k

$$\frac{\partial U}{\partial \boldsymbol{\mu}_k} = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{17}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \tag{18}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{19}$$

- N_k - effective number of points assigned to cluster k
- $\boldsymbol{\mu}_k$ - weighted mean of all of the points from the data set, with weights given by posterior probability $\gamma(z_{nk})$ (component k was responsible for generating $\boldsymbol{x}_n$)

Conditions satisfied by LLF at maximum: Σ_k

$$\frac{\partial U}{\partial \boldsymbol{\Sigma}_k} = 0 \tag{20}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \tag{21}$$

- $\boldsymbol{\Sigma}_k$ -weighted covariance of all of the points from the data set with each outer product weighted by the corresponding posterior probability and with the denominator given by the effective number of points associated with the corresponding component

Conditions satisfied by LLF at maximum: π_k

$$\sum_{k=1}^{K} \pi_k = 1 \tag{22}$$

$$U' = U + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{23}$$

$$\frac{\partial U'}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda = 0 \tag{24}$$

$$\pi_k = \frac{N_k}{N} \tag{25}$$

- π_k - mixing coefficient for k-th component is given by average responsibility, which that component takes for explaining the data points

Iterative solution (Expectation Maximization for GMM) i

1. Configure initial values for the means ($\boldsymbol{\mu}_k$), covariances ($\boldsymbol{\Sigma}_k$), and mixing coefficients π_k. Assign random or k-means based cluster assignments. Evaluate the initial value of log likelihood.
2. Iterate steps:
   2.1 **E**xpectation: Evaluate the responsibilities using the current parameter values:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{26}$$

Iterative solution (Expectation Maximization for GMM) ii

2.2 Maximization: Re-estimate the parameters using the current responsibilities:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \tag{27}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T \tag{28}$$

$$\pi_k = \frac{N_k}{N} \tag{29}$$

where:

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}) \tag{30}$$

2.3 Evaluate the log likelihood and check convergence criterion:

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma_k})\right\} \tag{31}$$

EM

- **E**xpectation: estimating the cluster assignments
- **M**aximization: updating the mixture parameters

K-means for initialization of EM

- The covariance matrices initialized to the sample covariances of the clusters found by the K-means algorithm
- Mixing coefficients can be set to the fractions of data points assigned to the respective clusters

Example: EM for Old Faithful dataset

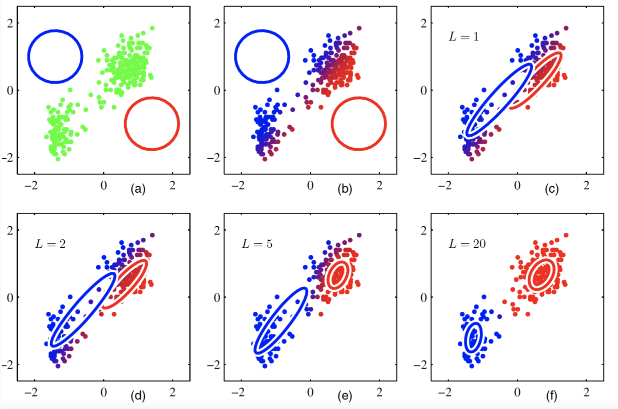


Figure 3: Old Faithful geyser eruption data [1].

EM for mixture of 1D Gaussians (1)

- Input: $\{x_1, \ldots, x_N\}$
- Model:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

- Expectation (responsibilities):

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left(-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right)}{\sum_{k=1}^{K} \pi_k \exp\left(-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right)}$$

- Maximization:

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

EM for mixture of 1D Gaussians (2)

- Maximization:
  - Update means μ_k:

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

  - Update variances σ_k:

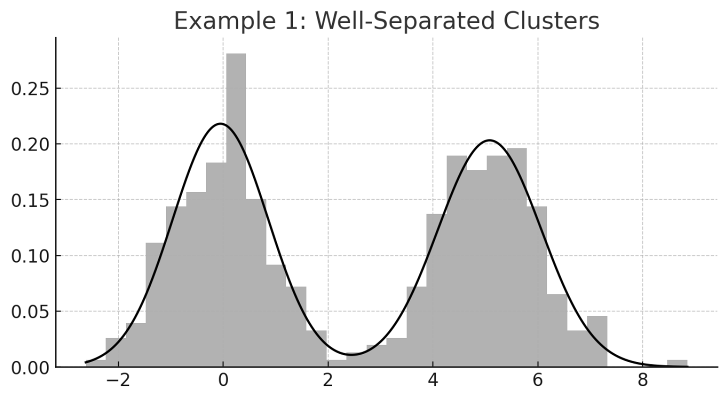
$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)^2$$

  - Update mixing coefficients π_k:

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

Example 1: well-separated clusters

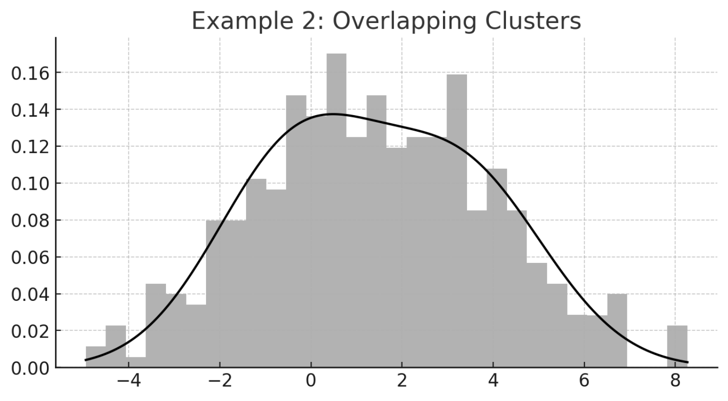
Cluster 1: $\mathcal{N}(0, 1)$ Cluster 2: $\mathcal{N}(5, 1)$



$\mu_1 = 0.05$, $\mu_2 = 5.08$, $\sigma_1 = 0.83$, $\sigma_2 = 0.97$,$\pi_1 = 0.5$,$\pi_2 = 0.5$

Example 2: overlapping clusters

Cluster 1: $\mathcal{N}(0, 2)$ Cluster 2: $\mathcal{N}(3, 2)$

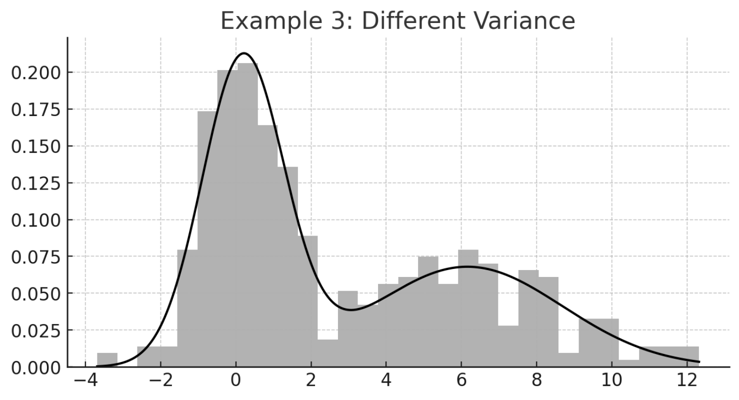


$\mu_1 = -0.31$, $\mu_2 = 3.32$, $\sigma_1 = 3.19$, $\sigma_2 = 3.40$,$\pi_1 = 0.52$,$\pi_2 = 0.48$

Example 3: different variance

Cluster 1: $\mathcal{N}(0,1)$

Cluster 2: $\mathcal{N}(5,3)$

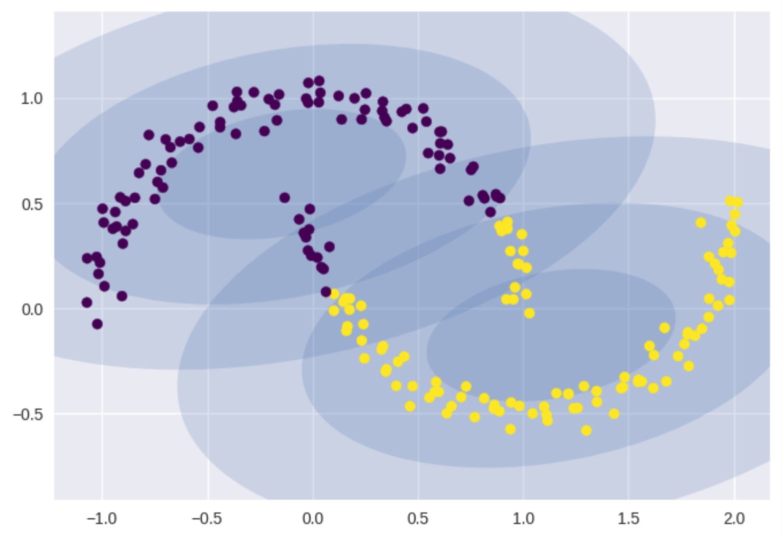


$\mu_1 = 0.19$, $\mu_2 = 6.16$, $\sigma_1 = 1.18$, $\sigma_2 = 6.40$,$\pi_1 = 0.57$,$\pi_2 = 0.43$

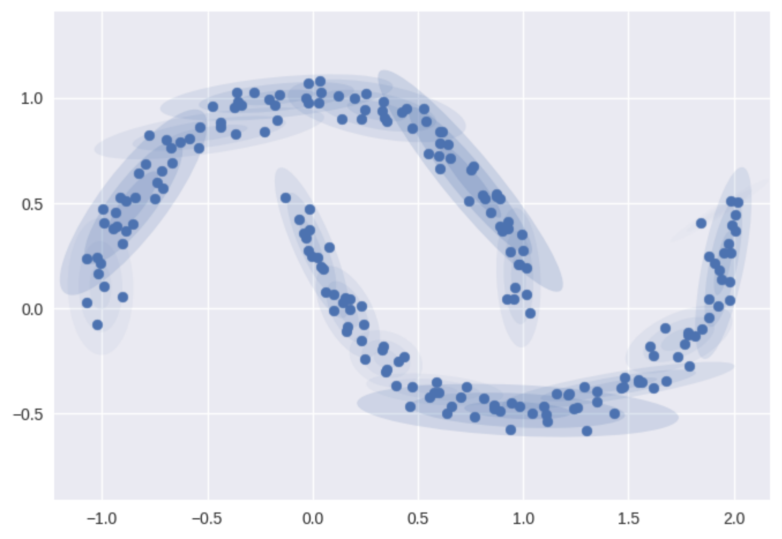
Example 4: Fitting GMM using EM

- Data: $X = \{5, 15, 25, 30, 40\}$
- Mixture of two 1D Gaussians
- Initial conditions: $\mu_1 = 10$, $\sigma_1 = 10$, $\mu_2 = 35$, $\sigma_2 = 10$, $\pi_1 = 0.5$, $\pi_2 = 0.5$

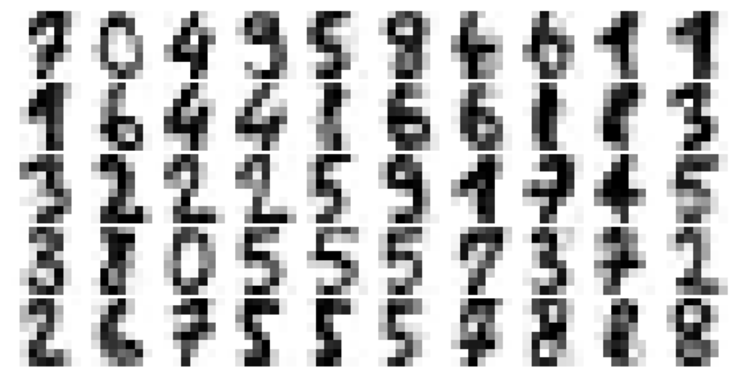
Example 5: Density estimation: 2D, 2 components



Example 6: Density estimation: 2D, 16 components



Example 7: Generating new data: 41D, 140 components



Akaike Information Criterion (AIC)

- AIC is a measure for model selection among a set of models.
- It is based on the concept of information entropy, providing a relative estimate of the information lost when a given model is used to represent the process that generates the data.

$$\mathrm{AIC} = 2m - 2\ln(\hat{L}) \tag{32}$$

- m represents the number of parameters in the statistical model, and $\hat{L}$ is the maximum value of the likelihood function for the model

AIC for GMM

- $m = K\left(D + \frac{D(D+1)}{2} + 1\right)$
- Fit GMMs with different numbers of components
- Calculate the AIC for each model
- Select the model with the lowest AIC value

GMM vs. k-means

	GMM	k-means
Model type	Probabilistic	Non-probabilistic (data partitioning)
Assumptions	Mixture of k Gaussians	Clusters are spherical and of similar size
Flexibility	Clusters of different size and covariance structure	Struggles with non-spherical clusters
Algorithm	EM	Iterative re-assignment based on centers
Type of clustering	Soft	Hard
Sensitivity	High	High

Extensions of GMMs

- Bayesian Gaussian Mixture Models
- Infinite Gaussian Mixture Models
- Dirichlet Process Mixture Models
- Robust Gaussian Mixture Models (models outliers or noise)
- Regularized Gaussian Mixture Models
- Spatial Gaussian Mixture Models
- Deep Gaussian Mixture Models (transform data into a latent space where GMM assumptions hold better)

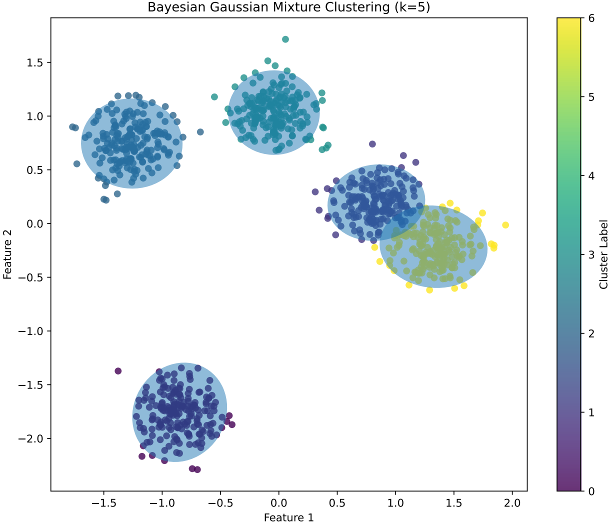
Bayesian Gaussian Mixture Model

- Automatically determines the optimal number of components
- Uses a Dirichlet Process (DP) or a finite Dirichlet prior to model the mixing weights, allowing the number of components to be flexible
- In BGMM, the parameters π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are treated as random variables with prior distributions:
    - Mixing coefficients follow Dirichlet distribution: $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$
    - Means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ typically follow conjugate priors: $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $\boldsymbol{\Sigma}_k^{-1} \sim \mathrm{Wishart}(\boldsymbol{W}_0, \nu_0)$

GMM vs. BGMM

	GMM	BGMM
Number of components	Fixed (K)	Flexible (determined by data)
Overfitting	May overfit if K is too large	Reduced risk due to regularization
Parameter estimation	Maximum Likelihood (MLE)	Bayesian Inference (MAP)
Mixing weights	Fixed proportions from MLE	Probabilistic, follows a Dirichlet prior.
Complexity	Simpler to implement and interpret	More complex due to Bayesian treatment
Applications	Well-suited for known K	Useful when K is unknown or variable

Example 8: BGMM



Dirichlet Process Mixture Models

- Bayesian approach to clustering
- Infinite mixture models model the number of components directly
- Estimate the optimal number of components from the data

Dirichlet distribution

- K-dimensional distribution over $K-1$ simplex
- A sample from a Dirichlet is a K-dimensional nonnegative vector $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ that sums to one $\sum_{k=1}^{K} x_k = 1$
- $\boldsymbol{\alpha}$ controls the concentration of density around each index
- PDF for Dirichlet distribution:

$$p(\boldsymbol{x}|\alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \tag{33}$$

- The Dirichlet is a popular prior for categorical variables with K categories.

Dirichlet-based clustering models

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{34}$$

$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ - component of the mixture

- Introducing prior $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_2] \sim \mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$
- Solution: Markov Chain Monte Carlo (MCMC)

Mixtures of Bernoulli distributions

Mixture of discrete binary variables (1)

Mixture of Bernoulli distributions

The set of D independent binary variables x_i, each of which is governed by Bernoulli distribution with parameter μ_i

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{(1-x_i)} \tag{35}$$

where $\boldsymbol{x} = [x_1, \ldots, x_D]^T$ and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_D]^T$

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu} \tag{36}$$

$$\mathrm{cov}[\boldsymbol{x}] = \mathrm{diag}\{\mu_i(1 - \mu_i)\} \tag{37}$$

Finite mixture:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) \tag{38}$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$

Mixture of discrete binary variables (2)

Mixture of Bernoulli distributions

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)} \tag{39}$$

$$\mathbb{E}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \tag{40}$$

$$\mathrm{cov}[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\} - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T \tag{41}$$

where $\boldsymbol{\Sigma}_k = \mathrm{diag}\{\mu_{ki}(1-\mu_{ki})\}$

Mixture of Bernoulli distributions - log likelihood

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$

$$\ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k) \tag{42}$$

Introducing latent variable

$\boldsymbol{z} = [z_1, \ldots, z_K]^T$ - associated with each instance of $\boldsymbol{x}$, single elements equal to 1, with all other elements equal to 0

$$p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k} \tag{43}$$

Prior distribution for the latent variables

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k-1}^{K} \pi_k^{z_k} \tag{44}$$

EM algorithm for mixture of Bernoulli distributions i

$\boldsymbol{X} = \{\boldsymbol{x}_n\}$, $\boldsymbol{Z} = \{\boldsymbol{z}_n\}$

1. **E**xpectation: Evaluate the responsibilities using the current parameter values:

$$\gamma(z_{nk}) = \frac{\sum_{z_{nk}} z_{nk}(\pi_k p(\boldsymbol{z}_n|\boldsymbol{\mu_k}))^{z_{nk}}}{\sum_{z_{nj}}(\pi_j p(\boldsymbol{z}_n|\boldsymbol{\mu_j}))^{z_{nj}}} = \frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\boldsymbol{x}_n|\boldsymbol{\mu}_j)} \tag{45}$$

2. **M**aximization: Re-estimate the parameters using the current responsibilities:

$$\boldsymbol{\mu}_k^{new} = \overline{\boldsymbol{x}}_k \tag{46}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{47}$$

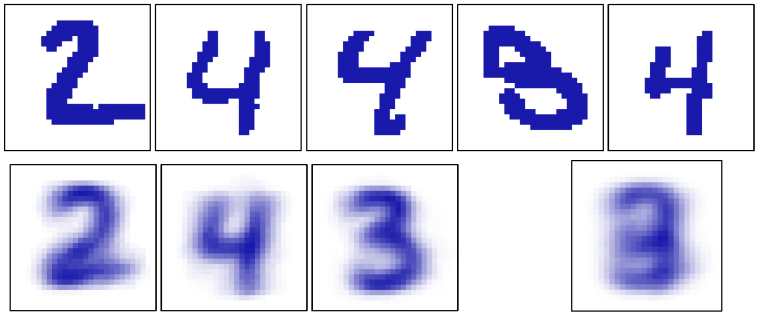
$$\overline{\boldsymbol{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x_n} \tag{48}$$

EM algorithm for mixture of Bernoulli distributions ii

$$\pi_k = \frac{N_k}{N} \tag{49}$$

3. Evaluate the log likelihood and check convergence criterion

Example 9: Bernoulli Mixture Model with 3 and 1 components



Zadanie 1

Dokładna Dokładna długość kabla wynosi θ. Masz do dyspozycji linijkę, dla której błąd pomiarowy podlega rozkładowi normalnemu z średnią 0 i odchyleniem standardowym 10^{-4}. Korzystając z linijki, mierzysz długość kabla, uzyskując rezultat pomiaru x podlegający rozkładowi normalnemu $\mathcal{N}(\theta, 10^{-4})$.

a) Przypuśćmy, że twoja wiedza wstępna mówi, że: $\theta \sim \mathcal{N}(9, 1)$. Jeśli pierwszy pomiar wynosi $x = 10$, to jaka jest gęstość prawdopodobieństwa rozkładu *a posteriori*?

b) Zakładając taką samą wiedzę wstępną, jak w punkcie a) oblicz liczbę pomiarów potrzebną do tego, żeby wariancja *posterior* dla θ była mniejsza niż 10^{-6}.

Zadanie 2

Piotr ma dwie monety: uczciwą (*fair*) oraz stronniczą (*biased*), dla której prawdopodobieństwo otrzymania orła w pojedynczym rzucie wynosi $\frac{3}{4}$. Piotr losuje jedną z monet (50-50), a następnie wykonuje serię rzutów wybraną monetą, aż do otrzymania pierwszej reszki. Biorąc pod uwagę, że zaobserwował on 3 orły zanim wypadła pierwsza reszka, znajdź prawdopodobieństwa *a posteriori*, że w pierwszym kroku wybrał monetę:

a) *fair*

b) *biased*

Wskazówka: Rozkład geometryczny to dyskretny rozkład prawdopodobieństwa opisujący prawdopodobieństwo zdarzenia, że proces Bernoulliego odniesie pierwszy sukces dokładnie w k-tej próbie:
$P(X = k) = (1 - \theta)^{k-1}\theta$

Zadanie 3

Zamierzamy zbudować model mikstur Gaussowskich jednowymiarowych przy użyciu $K = 2$ komponentów. Mamy $N = 5$ przypadków uczących, w których wartości x są następujące: $x_1 = 5$, $x_2 = 15$, $x_3 = 25$, $x_4 = 30$, $x_5 = 40$. Używamy algorytmu EM to znalezienia estymat *maximum likelihood* dla parametrów tego modelu. Są nimi współczynniki mieszania π_1 oraz π_2 oraz średnie μ_1 i μ_2. Zakładamy, że wartość odchylenia standardowego dla obu komponentów jest stała i wynosi 10. Przyjmując warunki początkowe: $\pi_1 = 0.5$, $\pi_2 = 0.5$, $\mu_1 = 10$, $\mu_2 = 35$ przedstaw pierwszą iterację algorytmu EM (faza E) obliczając wartości odpowiedzialności $\gamma(z_{nk})$ (potrzebne do wypełnienia tabeli poniżej).

Zadanie 4

Czas oczekiwania klientów (w godzinach) w popularnej restauracji można modelować jako **wykładniczą zmienną losową** z parametrem λ. Załóżmy, że *a priori* wiemy, że λ może przyjmować dowolną wartość w przedziale $(0, \infty)$ i ma funkcję gęstości prawdopodobieństwa:

$$f(\lambda) = \frac{1}{4!}\lambda^4 e^{-\lambda}$$

Załóżmy, że obserwujemy 5 klientów z czasami oczekiwania: $x_1 = 0.23$, $x_2 = 0.80$, $x_3 = 0.12$, $x_4 = 0.35$, $x_5 = 0.5$. Oblicz funkcję gęstości prawdopodobieństwa *a posteriori* dla λ. *Wskazówka:*

$$\int_0^\infty y^{a-1} e^{-by} dy = \frac{(a-1)!}{b^a}$$

References

C. M. Bishop.
***Pattern Recognition and Machine Learning*.**
Springer, 2006.