annotated version

Machine Learning Course - CS-433

Gaussian Mixture Models

Nov 28, 2023

Martin Jaggi
Last updated on: November 28, 2023

credits to Mohammad Emtiyaz Khan & Rüdiger Urbanke

EPFL

Motivation

$\Sigma = 1$ $\quad$ Σ general

1. K-means forces the clusters to be *spherical*, but sometimes it is desirable to have *elliptical* clusters. Another issue is that, in K-means, each example can only belong to one clus-
2. ter, but this may not always be a good choice, e.g. for data points that are near the "border". Both of these problems are solved by using Gaussian Mixture Models.

Clustering with Gaussians

parameters:
$\mu \in \mathbb{R}^{D \cdot K}$
$\Sigma \in \mathbb{R}^{D^2 \cdot K}$
$\pi \in \mathbb{R}^{K}$

The first issue is resolved by using full covariance matrices $\mathbf{\Sigma}_k$ instead of *isotropic* covariances.

1.

$$p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)]^{z_{nk}}$$

Soft-clustering

$\boxed{\begin{array}{l} z_n = \text{vector } 0 \ldots 1 \ldots 0 \\ z_{nk} = \begin{cases} 1 & \text{if assigned to } k \\ 0 \end{cases} \end{array}}$

The second issue is resolved by defining z_n to be a random variable. Specifically, define $z_n \in \{1, 2, \ldots, K\}$ that follows a multinomial distribution.

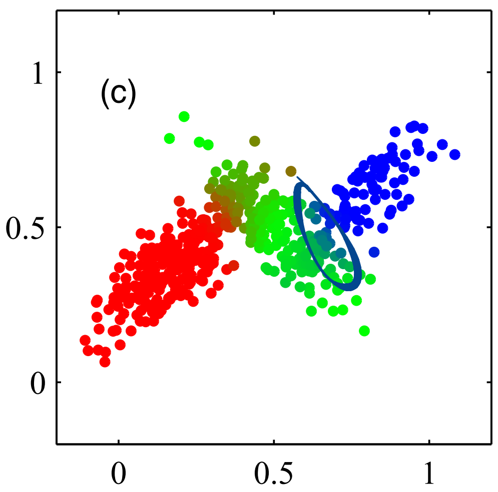
random variable : vector $0 \ldots 1 . \ldots 0$ $\quad$ cluster k

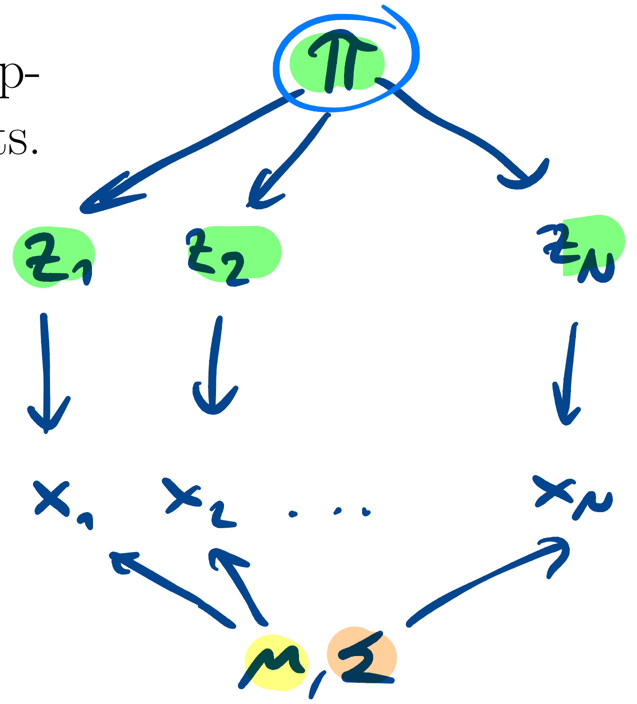
2.

$$\boxed{p(z_n = k) \overset{\text{def}}{=} \pi_k} \text{ where } \pi_k > 0, \forall k \text{ and } \sum_{k=1}^{K} \pi_k = 1$$

$z_{nk} = 1$

This leads to soft-clustering as opposed to having "hard" assignments.





Bayes Rule
$p(a,b) = p(a|b) \cdot p(b)$

Gaussian mixture model

Together, the likelihood and the prior define the joint distribution of Gaussian mixture model (GMM):

$$p(\mathbf{X}, \mathbf{z} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$

$$= \prod_{n=1}^{N} p(\mathbf{x}_n | z_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(z_n | \boldsymbol{\pi})$$

π $\quad \Sigma, \mu |$

$p(z|\pi)$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} [\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} \prod_{k=1}^{K} [\pi_k]^{z_{nk}}$$

$= \pi_{k'}$ $\quad$ k' is where $z_{k'n} = 1$

Here, $\mathbf{x}_n$ are observed data vectors, z_n are *latent* unobserved variables, and the unknown *parameters* are given by $\boldsymbol{\theta} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K, \boldsymbol{\pi}\}$.

Marginal likelihood

joint $p(x_n, z_n)$

GMM is a latent variable model with z_n being the unobserved (latent) variables. An advantage of treating z_n as latent variables instead of *parameters* is that we can *marginalize* them out to get a cost function that does not depend on z_n, i.e. as if z_n never existed.

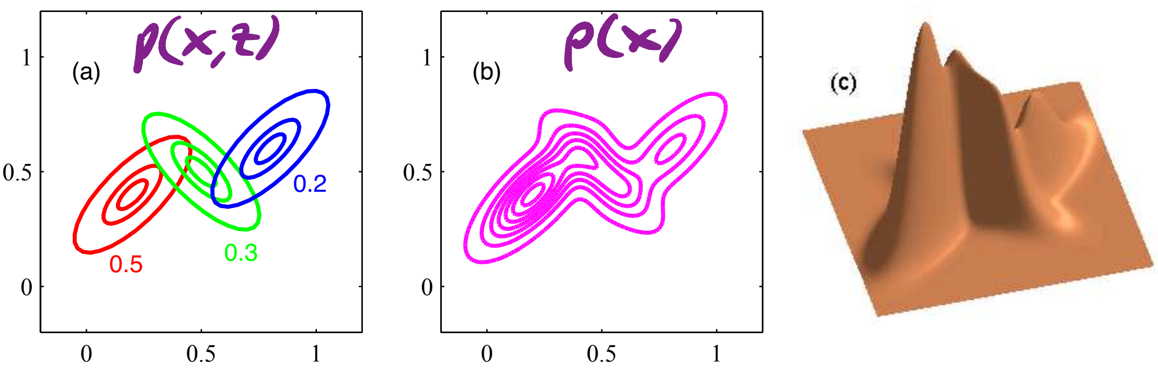
marginal $p(x_n) = \sum_{k=1}^{K} p(x_n, z_n = k)$

$= \sum_k p(x_n|z_n)\, p(z_n)$

Specifically, we get the following marginal likelihood by marginalizing z_n out from the likelihood:

μ, Σ, π $=$

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



Deriving cost functions this way is good for *statistical efficiency*. Without a latent variable model, the number of parameters grows at rate $\mathcal{O}(N)$. After marginalization, the growth is reduced to $\mathcal{O}(D^2K)$ (assuming $D, K \ll N$).

$z \in \{0,1\}^{N \cdot K}$

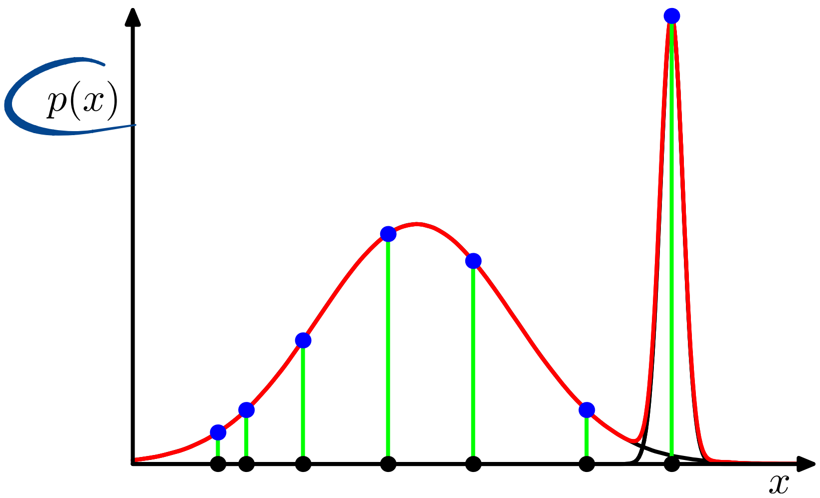
Maximum likelihood

To get a maximum (marginal) likelihood estimate of $\boldsymbol{\theta}$, we maximize the following: $\log p(X, \theta)$

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\log \left(p(X^{all} | \theta) = \prod_{n=1}^{N} \underbrace{p(x_n | \theta)}_{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)} \right)$$

Is this cost convex? Identifiable? Bounded?



width $\sigma_k \to 0$

1. non-convex

   (see k-means)

2. non-unique optima

   θ: $\pi_k \leftrightarrow \pi_{k'}$, $\mu_k \leftrightarrow \mu_{k'}$, $\Sigma_k \leftrightarrow \Sigma_{k'}$

   permutation of K

3. unbounded

   $\Sigma_k = \sigma_k \mathbb{1}$

   scalar