annotated version

Machine Learning Course - CS-433

Gaussian Mixture Models

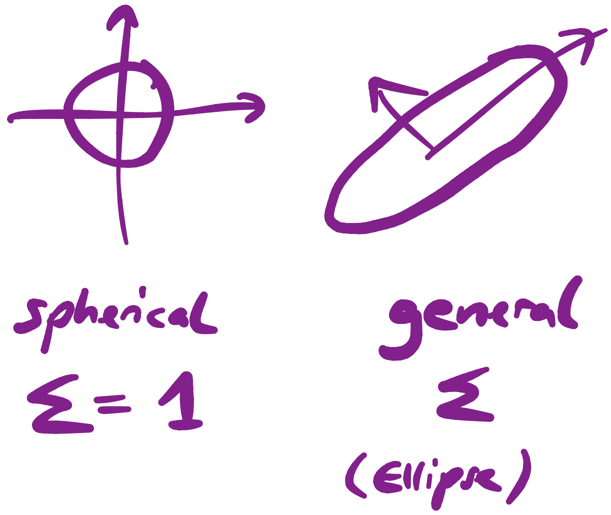
Nov 30, 2021

changes by Martin Jaggi 2021, 2020, 2019, changes by Rüdiger Urbanke 2018, changes by Martin Jaggi 2017, 2016

Last updated on: November 30, 2021

Motivation

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 cluster, 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

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

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

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

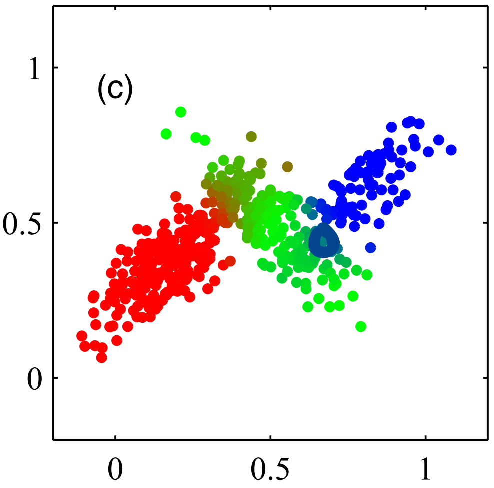
Soft-clustering

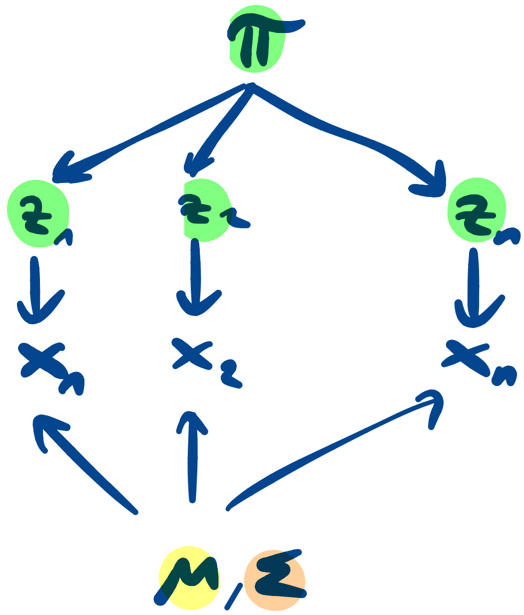
The second issue is resolved by defining z_n to be a random variable. Specifically, define $z_n \in \{1, 2, \ldots, K\}$ that follows a multinomial distribution.

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

importance of cluster k

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





Gaussian mixture model

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

joint

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

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

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

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

Bayes Rule

$p(a,b|\theta) = p(a|b,\theta) \cdot p(b|\theta)$

$z_n = (0, \ldots 1, \ldots 0)$

$z_{nk} = \{ 0 \ldots 1 \ldots$ $\uparrow$ nk

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

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

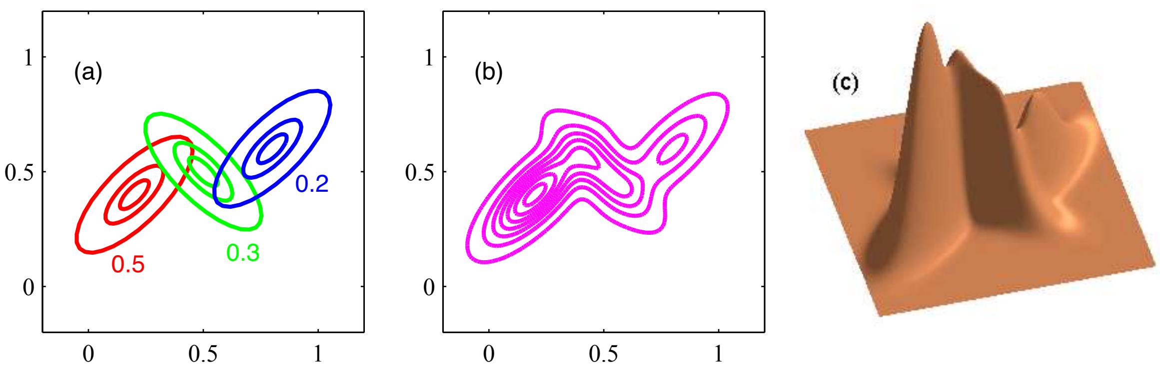
$= \sum p(x|z) \cdot \underbrace{p(z=k)}_{\pi_k}$

Bayes

orig. likelihood $p(x_n, z_n | \theta)$

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

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



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 : $\mathcal{O}(N)$

θ : μ : KD

Σ : KD^2

π : K

Maximum likelihood

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

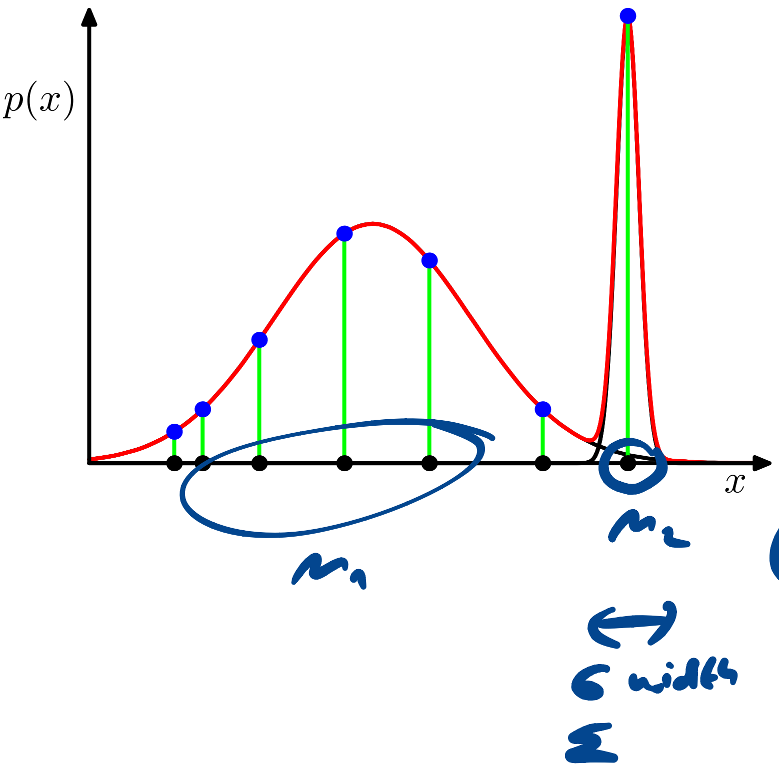
"(μ, Σ, π)

$-\mathcal{L} =$ marginal likelihood

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

$$\log\left(p(\text{"all" } x_n | \theta) = \prod_{n=1}^{N} \underbrace{p(x_n | \theta)}_{\sum_k \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)} \right)$$

Is this cost convex? no Identifiable? no Bounded? no



1. non-convex (see k-means)
2. non-unique optima

permutation of 1.. K

$k \to k'$ $\pi_k \to \pi_{k'}$ $\Sigma_k \to \Sigma_{k'}$ $\mu_k \to \mu_{k'}$

3. unbounded

$-\mathcal{L} \to \infty$

if $\Sigma = \sigma \cdot 1$ ($\uparrow$ width)

and $\sigma \to 0$