annotated version

Machine Learning Course - CS-433

Gaussian Mixture Models

Nov 24, 2020

Last updated on: November 24, 2020

EPFL

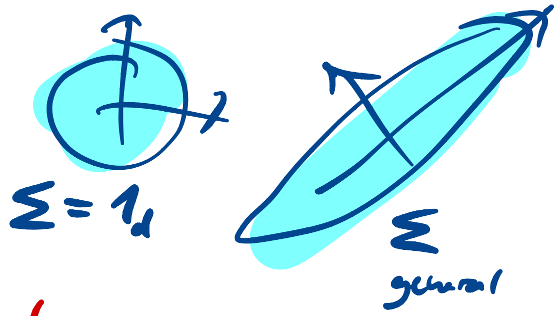
Motivation

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 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.

2

Clustering with Gaussians

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



more general

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

$\boxed{\text{K-means if } \Sigma_k = \mathbb{1}}$

$z_n = (0, \ldots, 0, 1, 0 \ldots 0) \in \mathbb{R}^K$

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.

parameters:

$\mu \in \mathbb{R}^{D \cdot K}$

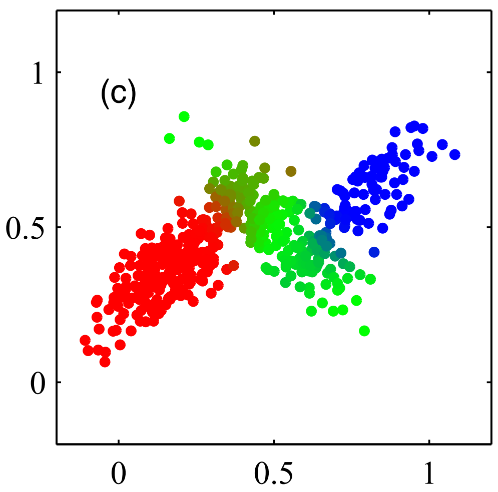
$\Sigma \in \mathbb{R}^{D^2 \cdot K}$

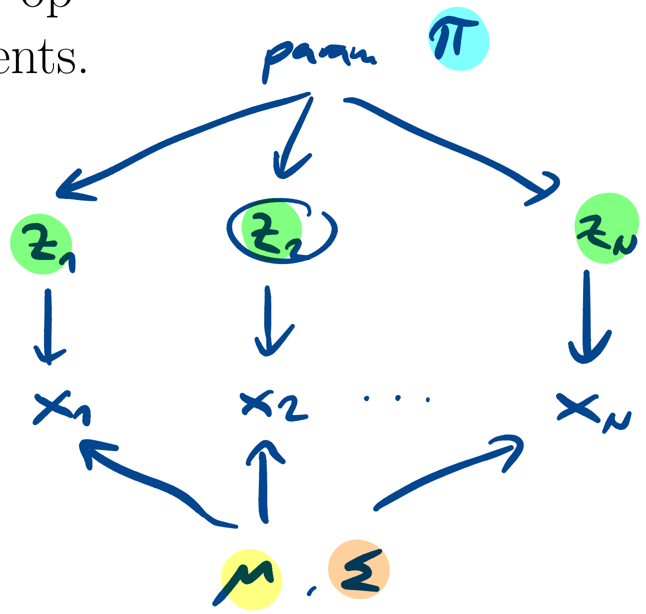
$\pi \in \mathbb{R}^K$

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

importance weight of group 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):

Bayes Rule

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

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

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

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

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

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

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.

joint

$p(x_n, z_n)$

marginal

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

$= \sum_{k=1}^{K} p(x_n | z_n = k) \cdot \underbrace{p(z_n = k)}_{\pi_k}$

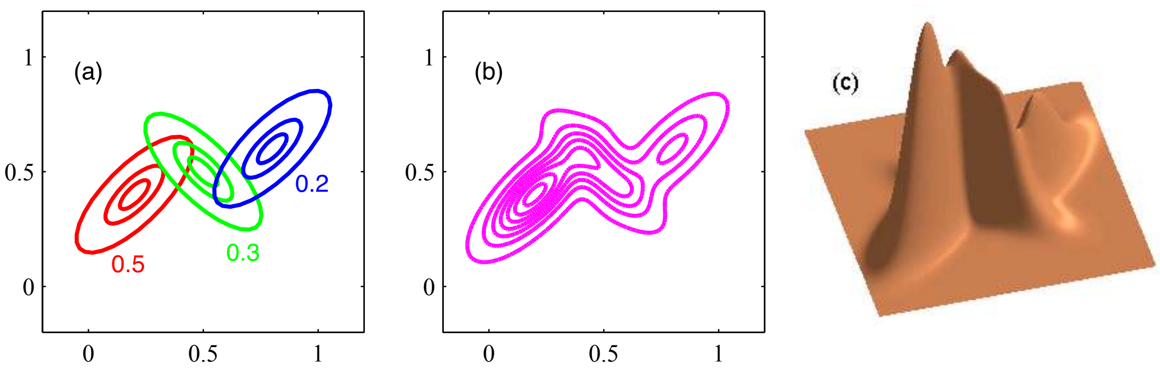
Bayes

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

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

marginal 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 grow at rate $O(N)$. After marginalization, the growth is reduced to $O(D^2K)$ (assuming $D, K \ll N$).

z : N

θ : μ_k : KD

Σ_k : KD^2

π_k : K

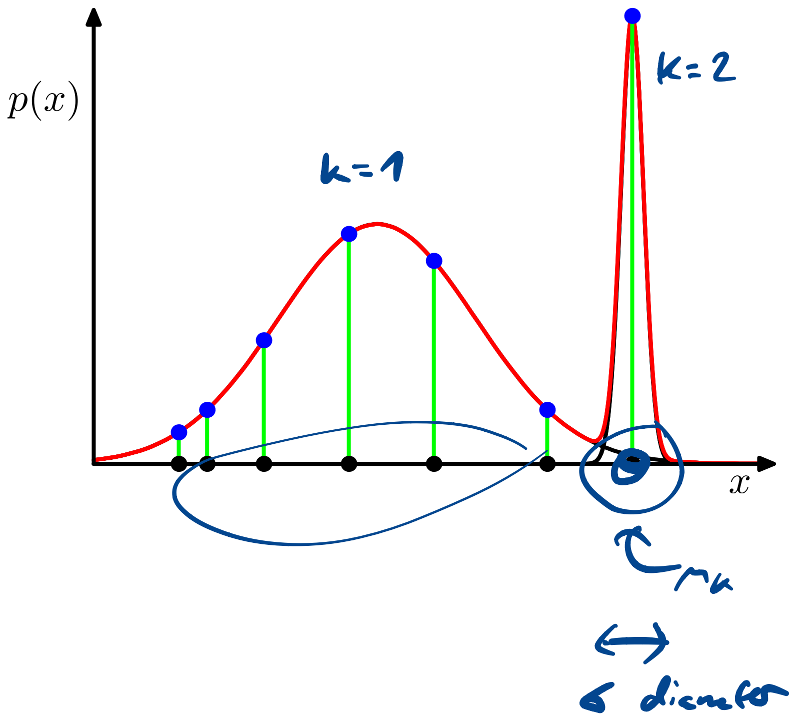
Maximum likelihood

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

$$\log\Big(p(\text{"all" } x_n \mid \theta)\Big) = \log \underbrace{\prod_{n=1}^{N} p(x_n \mid \theta)}_{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)} = \sum_{n=1}^{N} \log \sum_k \mathcal{N}(\cdot\cdot \mid \cdot\cdot)$$

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

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



① non-convex

② non-unique optima

permutation of [K]

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

③ un-bounded

$\mathcal{L} \to \infty$

if $\Sigma_k = \sigma I$ (σ: width)

when $\sigma \to 0$