annotated version

Machine Learning Course - CS-433

Gaussian Mixture Models

Nov 9, 2017

minor changes by Martin Jaggi 2016

minor changes by Martin Jaggi 2017

Last updated: November 9, 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

K-means forces the clusters to be (1) *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, (2) 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.

for k-means: $\Sigma_k = I$

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

Soft-clustering

vs hard assignment

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
$\mathbb{R}^{D\times K}$	$\mu = (\mu_1, \ldots, \mu_K)$
$\mathbb{R}^{D\times D\times K}$	$\Sigma = (\Sigma_1, \ldots, \Sigma_K)$
$\mathbb{R}^{K}$	$\pi = (\pi_1, \ldots, \pi_K)$

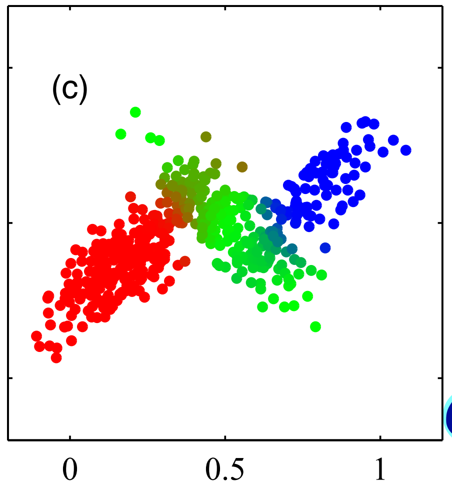
importance weight of group k

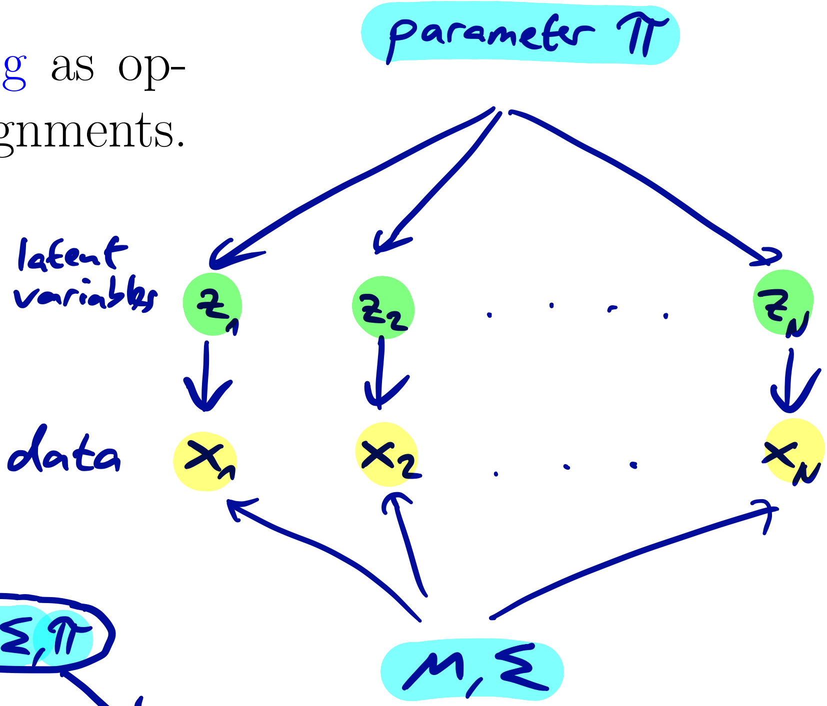
$$p(z_n = k) = \pi_k \text{ where } \pi_k > 0, \forall k \text{ and } \sum_{k=1}^{K} \pi_k = 1$$

$z_n := (0, \ldots, \overset{k}{1}, \ldots 0)$

with $\mathbb{P} = \pi_k$

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):

$$p(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_n p(x_n, z_n \ldots)$$

Bayes Rule

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

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

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

for each datapoint n

$$\pi \longrightarrow z_n$$

(0.3, 0.6, 0.1) (0, 1, 0)

k

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

$$z_n = \begin{cases} (1\ 0\ 0\ \ldots\ 0) & \text{if } k=1 \\ \ldots\ 1\ \ldots & \\ (0\ 0\ 0\ \ldots\ 1) & \text{if } k=K \end{cases}$$

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

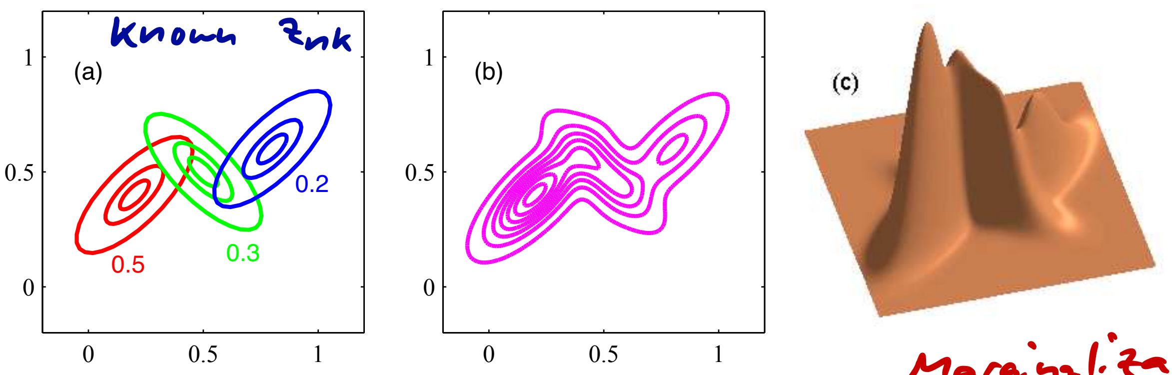
Specifically, we get the following marginal likelihood by marginalizing z_n out from the likelihood:

parameters $\theta = (\mu, \Sigma, \pi)$

$p(X, z | \theta)$

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

known z_{nk}



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 $D, K \ll N$).

Marginalization

~~z : N~~

π : K

μ : $K \cdot D$

Σ : $K \cdot D^2$

$$\log\left(p(\mathbf{x}_n^{\text{all}}|\theta)\right) = \prod_{n=1}^{N} \underbrace{p(\mathbf{x}_n|\theta)}_{\sum_{k=1}^{K} \cdots}$$

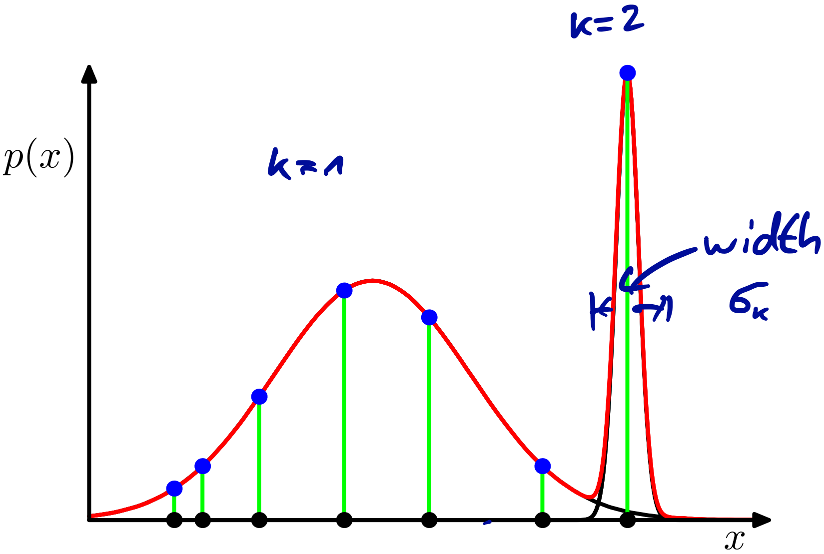
$$= \sum \log(\ldots)$$

Maximum likelihood

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

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

Is this cost convex? Identifiable? Bounded?



($\mathcal{L}$ not concave)

① non-convex in θ

② non-unique optimum θ

permutations of k

$k \to k'$

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

③ non-bounded

$\mathcal{L}(\theta) \longrightarrow \infty$

if $\Sigma_k := \sigma_k I$

in the limit

$\sigma_k \longrightarrow 0$

Exercises

1. Understand K-means extension to GMM. Why do we treat z_n as a random variable? Identify the joint, likelihood, prior, and marginal distributions, respectively.