Learning Probabilities and Causality

Chapter II - Gaussian mixture models (GMM) and the expectation-maximisation (EM) algorithm

Xavier Alameda-Pineda and Thomas Hueber

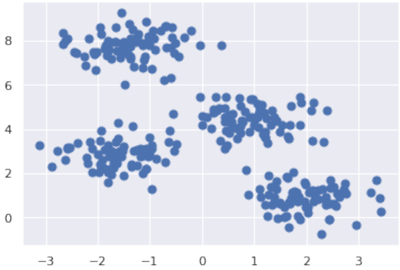
Ensimag/Inria/CNRS/Univ. Grenoble-Alpes

Contents

Model-based clustering and GMM

Clustering

Definition: find groups of data points without labels.



Very intuitive algorithm

What would you do?

Very intuitive algorithm

What would you do?

1. Initialise randomly K centroids.
2. Assign each data point to the closes centroid.
3. Recompute centroids from the assignments.
4. Iterate the past two steps.

This is called the K-means algorithm. Let's see it on colab.

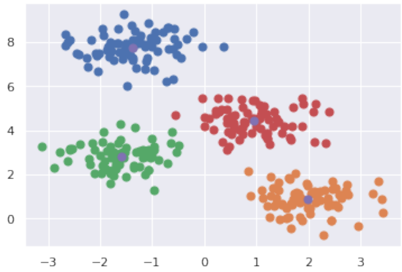
Important points of K-means

Automatic inference of latent variables

The point-to-cluster assignment variable is **unknown/latent/hidden**, and must be infered together with the parameters.

Limited to spherical and equally populated clusters

The assignment criterion is the Euclidean distances $\Rightarrow$ groups are spherical and equally populated.



Generalising K-means: GMM

Remark 2.1: Gaussian mixture model (GMM):

- For each data point $\boldsymbol{x}_n$ there is a hidden variable z_n taking discrete values from 1 to K: $z_n \in \{1, \ldots, K\}$.

Generalising K-means: GMM

Remark 2.1: Gaussian mixture model (GMM):

- For each data point $\boldsymbol{x}_n$ there is a hidden variable z_n taking discrete values from 1 to K: $z_n \in \{1, \ldots, K\}$.
- Its prior probability is defined as: $p(z_n = k) = \pi_k$, $\sum_{k=1}^{K} \pi_k = 1$.

Generalising K-means: GMM

Remark 2.1: Gaussian mixture model (GMM):

- For each data point $\boldsymbol{x}_n$ there is a hidden variable z_n taking discrete values from 1 to K: $z_n \in \{1, \ldots, K\}$.
- Its prior probability is defined as: $p(z_n = k) = \pi_k$, $\sum_{k=1}^{K} \pi_k = 1$.
- Given z_n, the data point is modeled as a multivariate Gaussian:

$$p(\boldsymbol{x}_n | z_n = k) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Advantages:

1. Having $\pi_1, \ldots, \pi_K$ means that groups can be differently populated.
2. The shape of the groups is modeled by $\boldsymbol{\Sigma}_k$.

Maximum likelihood for GMM

Let's compute $p(\boldsymbol{x}_n)$

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n, z_n = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The log-likelihood:

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

with $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

Compute either $\frac{\partial \mathcal{L}}{\partial \pi_k}$, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k}$ or $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k}$.

Maximum likelihood for GMM

Let's compute $p(\boldsymbol{x}_n)$

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n, z_n = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The log-likelihood:

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{X}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

with $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

Compute either $\frac{\partial \mathcal{L}}{\partial \pi_k}$, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k}$ or $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k}$. Very difficult to optimise.

Maximum Likelihood for GMM: the EM algorithm

Log-likelihood?

We have see that $\log p(\boldsymbol{x})$ does not work well with derivatives.

However, $\log p(\boldsymbol{x}, z)$ does!

Problem: z is not observed, we must take the expectation w.r.t. z.

Log-likelihood?

We have see that $\log p(\boldsymbol{x})$ does not work well with derivatives.

However, $\log p(\boldsymbol{x}, z)$ does!

Problem: z is not observed, we must take the expectation w.r.t. z.

We propose to do it using the posterior distribution $p(z|\boldsymbol{x})$:
(we will justify this choice later on)

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(z|\boldsymbol{x};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{x}, z; \boldsymbol{\Theta})$$

This function is called: expected complete-data log-likelihood.

Towards the EM algorithm for GMM

Notation: observations $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, latent variables $\boldsymbol{Z} = \{z_n\}_{n=1}^N$ and parameters $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$.

Towards the EM algorithm for GMM

Notation: observations $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$, latent variables $\boldsymbol{Z} = \{z_n\}_{n=1}^{N}$ and parameters $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

Remark 2.3: Given $\boldsymbol{\Theta}^0$, we use the *expected complete-data log-likelihood* $\mathcal{Q}$:

1. Expectation: $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$
2. Maximisation: $\boldsymbol{\Theta}^1 = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0)$

Towards the EM algorithm for GMM

Notation: observations $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, latent variables $\boldsymbol{Z} = \{z_n\}_{n=1}^N$ and parameters $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$.

Remark 2.3: Given $\boldsymbol{\Theta}^0$, we use the *expected complete-data log-likelihood* $\mathcal{Q}$:

1. Expectation: $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$
2. Maximisation: $\boldsymbol{\Theta}^1 = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0)$

We can look back to K-means:

1. Infer latent variables (assignment) given the parameters (centroids).
2. Estimate the parameters (centroids) given the assignments.

EM for GMM

Let's recall: $p(z_n = k) = \pi_k$ & $p(\boldsymbol{x}_n | z_n = k) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Expectation: Compute $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$.

- Start with $p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$.

EM for GMM

Let's recall: $p(z_n = k) = \pi_k$ & $p(\boldsymbol{x}_n | z_n = k) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Expectation: Compute $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$.

- Start with $p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$. And name it $\eta_{nk} = p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$.
  η_{nk} is the posterior probability that $\boldsymbol{x}_n$ belongs to group k.
- Then $p(\boldsymbol{x}_n, z_n | \boldsymbol{\Theta})$

EM for GMM

Let's recall: $p(z_n = k) = \pi_k$ & $p(\boldsymbol{x}_n | z_n = k) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Expectation: Compute $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$.

- Start with $p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$. And name it $\eta_{nk} = p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$.
  η_{nk} is the posterior probability that $\boldsymbol{x}_n$ belongs to group k.
- Then $p(\boldsymbol{x}_n, z_n | \boldsymbol{\Theta}) = \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Also $\mathbb{E}_{p(z_n|\boldsymbol{Z}_n;\boldsymbol{\Theta}^0)} \log p(z_n, \boldsymbol{x}_n; \boldsymbol{\Theta})$

EM for GMM

Let's recall: $p(z_n = k) = \pi_k$ & $p(\boldsymbol{x}_n | z_n = k) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Expectation: Compute $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}^0)} \log p(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\Theta})$.

- Start with $p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$. And name it $\eta_{nk} = p(z_n = k | \boldsymbol{x}_n; \boldsymbol{\Theta}^0)$. η_{nk} is the posterior probability that $\boldsymbol{x}_n$ belongs to group k.
- Then $p(\boldsymbol{x}_n, z_n | \boldsymbol{\Theta}) = \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.
- Also $\mathbb{E}_{p(z_n|\boldsymbol{z}_n;\boldsymbol{\Theta}^0)} \log p(z_n, \boldsymbol{x}_n; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \eta_{nk} \log \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Remark 2.4:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Let's recall:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

This splits:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \pi_k + \eta_{nk} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

We consider now the Lagrangian for $\pi_1, \ldots, \pi_K$:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^0) = \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_{nk} \log \pi_k + \beta \Big(1 - \sum_{k=1}^{K} \pi_k \Big).$$

EM for GMM (III)

Exercise 2.1: Prove that the optimal parameters write:

$$\pi_k^* = \frac{1}{N} S_k, \qquad S_k = \sum_{n=1}^{N} \eta_{nk},$$

and:

$$\boldsymbol{\mu}_k^* = \frac{1}{S_k} \sum_{n=1}^{N} \eta_{nk} \boldsymbol{x}_n \qquad \boldsymbol{\Sigma}_k^* = \frac{1}{S_k} \sum_{n=1}^{N} \eta_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^*)(\boldsymbol{x}_n - \boldsymbol{\mu}_k^*)^\top.$$

The EM algorithm

The EM in general

Let us assume a probabilistic graphical model, with observed variables $\boldsymbol{X}$, hidden variables $\boldsymbol{z}$ and parameters $\boldsymbol{\Theta}$.

Remark 2.5: Initialise the parameters $\boldsymbol{\Theta}^0$. For iteration $r = 1, \ldots, R$:

E-step Compute $p(\boldsymbol{z}|\boldsymbol{X}; \boldsymbol{\Theta}^{r-1})$ and $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

M-step Compute $\boldsymbol{\Theta}^r = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

Comments

- EM is sensible to initialisation.
- It may converge to a local maxima or saddle point.
- We still need to compute and optimise $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{r-1})$.

But why does it work?

The main mathematical object in EM is $\mathcal{Q}$.
(The expected complete-data log-likelihood).

What is the relationship with the log-likelihood?
Let's take any distribution of $\boldsymbol{z}$: $q(\boldsymbol{z})$ and ignore $\boldsymbol{\Theta}$ for the time being.

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \mathbb{E}_{q(\boldsymbol{z})}\Big[\log p(\boldsymbol{x})\Big] \\
&= \mathbb{E}_{q(\boldsymbol{z})}\Big[\log p(\boldsymbol{x})\frac{p(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})q(\boldsymbol{z})}\Big] \\
&= \mathbb{E}_{q(\boldsymbol{z})}\Big[\log \frac{p(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\Big] + D_{\mathrm{KL}}\Big(q(\boldsymbol{z})\Big\|p(\boldsymbol{z}|\boldsymbol{x})\Big)
\end{aligned}$$

But why does it work? (II)

$$\log p(\boldsymbol{x};\boldsymbol{\Theta}) = \underbrace{\mathbb{E}_{q(\boldsymbol{z})}\Big[\log\frac{p(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\Big]}_{\text{M-step}} + \underbrace{D_{\text{KL}}\Big(q(\boldsymbol{z})\Big\|p(\boldsymbol{z}|\boldsymbol{x})\Big)}_{\text{E-step}}$$

Another interpretation. Given $\boldsymbol{\Theta}^0$:

1. Set $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\Theta}^0)$.
2. Optimise w.r.t. $\boldsymbol{\Theta}$:

$$\mathbf{E}_{q(\boldsymbol{z})}\Big[\log\frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\Theta})}{q(\boldsymbol{z})}\Big]$$

Why?

E-step: reduce the distance between log-likelihood and $\mathcal{Q}$.
M-step: push $\mathcal{Q}$ and therefore push the log-likelihood.

But why does it work? (III)

$$\log p(\boldsymbol{x}; \boldsymbol{\Theta}) = \underbrace{\mathbb{E}_{q(\boldsymbol{z})}\Big[\log \frac{p(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\Big]}_{\mathcal{L}(q,\boldsymbol{\Theta})} + \underbrace{D_{\text{KL}}\Big(q(\boldsymbol{z})\Big\|p(\boldsymbol{z}|\boldsymbol{x})\Big)}_{\text{KL}(q\|p)}$$