DDA3020 Machine Learning: Lecture 16 Expectation Maximization

JIA, Kui
School of Data Science, CUHK-SZ

November 21, 2023

Outline

EM for GMM Clustering

- Last time: We have introduced EM algorithm as a way of fitting a Gaussian Mixture Model for clustering
  - E-step: Compute probability each datapoint came from certain cluster, given model parameters
  - M-step: Adjust parameters of each cluster to maximize probability it would generate data it is currently responsible for
- This lecture: derive EM from principled approach and see how EM can be applied to general latent variable models

Latent Variable Models

- Recall: variables which are always unobserved are called **latent variables** or sometimes hidden variables
- In a mixture model, the identity of the component that generated a given datapoint is a latent variable
- Why use latent variables if introducing them complicates learning?
    - We can build a complex model out of simple parts - this can simplify the description of the model
    - We can sometimes use the latent variables as a representation of the original data (e.g. cluster assignments in a GMM model)

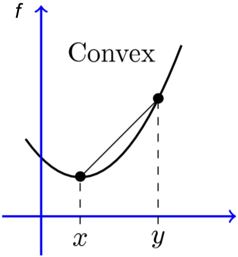
Preliminaries: Convex and Concave Functions

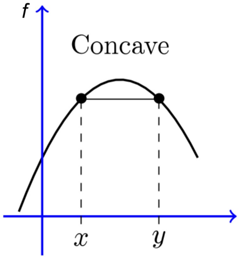
- **Theorem 1**: Suppose f is a convex function, for any two input points x and y, as well as any scalar value $\alpha \in [0, 1]$, we have

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f(x) + (1-\alpha)f(y).$$

- **Theorem 2**: Suppose f is a concave function, for any two input points x and y, as well as any scalar value $\alpha \in [0, 1]$, we have

$$f\big(\alpha x + (1-\alpha)y\big) \geq \alpha f(x) + (1-\alpha)f(y).$$





Preliminaries: Jensen's Inequality

The above theorems can be extended to Jensen's Inequality.

Theorem (Jensen's Inequality)

Suppose f is a convex function, and X is a random variable, then we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

If f is a concave function, then we have

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)].$$

When the equality holds?

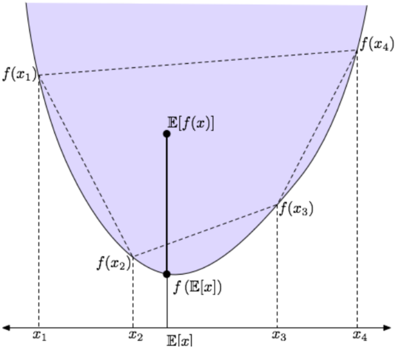
- X has a unique state
- f is not strongly convex/concave

Try to prove the above theorem and claims by yourself. (Hint: using mathematical induction to prove)

Preliminaries: Jensen's Inequality

For example, as shown in the right figure, f is a convex fucntion, and there are four candidate states of X, *i.e.*, x_1, x_2, x_3, x_4. Given any setting of the probability distribution (*i.e.*, $P(X = x_i) = \alpha_i$), it always has

$$f(\sum_{i=1}^{4} \alpha_i x_i) \leq \sum_{i=1}^{4} \alpha_i f(x_i).$$



Notations of Latent Variable Models

- In this lecture, we'll be using $\mathbf{x}$ to denote **observed data** and z to denote the **latent variables**
- We assume we have an observed dataset $\mathcal{D} = \left\{\mathbf{x}^{(n)}\right\}_{n=1}^{N}$ and would like to fit $\boldsymbol{\theta}$ using maximum log likelihood:

$$\log p(\mathcal{D}; \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p\left(\mathbf{x}^{(n)}; \boldsymbol{\theta}\right)$$

- To compute $p(\mathbf{x}; \boldsymbol{\theta})$, we have to **marginalize** over z:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{z} p(z, \mathbf{x}; \boldsymbol{\theta}),$$

  where $p(z, \mathbf{x}; \boldsymbol{\theta})$ denotes the probabilistic model we should define. Note that
  - Anything following a semicolon denotes a parameter of the distribution
  - We're not treating the parameters as random variables

Difficulty of Fitting Latent Variable Models

- Typically, there is no closed form solution to the maximum likelihood problem

$$\log p(\mathcal{D}; \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p\left(\mathbf{x}^{(n)}; \boldsymbol{\theta}\right) = \sum_{n=1}^{N} \log \left(\sum_{z^{(n)}} p\left(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}\right) \right).$$

- Key difficulty: once z is marginalized out, $p(\mathbf{x}; \theta)$ could be complex (*e.g.*, a mixture distribution).
- As shown in GMM (see last slides), if our objective is in terms of $\log p(z, \mathbf{x}; \boldsymbol{\theta})$, which can be fully decomposed, then the optimization is very simple.
- To accomplish this, we need to move the summation outside the log.

Auxiliary Distribution of Latent Variables

- We firstly introduce a new distribution *w.r.t.* each latent variable $z^{(n)}$, denoted as $q_n(z^{(n)})$.
- We assume that the distributions *w.r.t.* different latent variables could be different, and they are independent, *i.e.*,

$$q(\mathbf{z}) = \prod_{n=1}^{N} q_n(z^{(n)}).$$

- Note that here we don't specify the parameter value of $q_n(z^{(n)})$, which will be learned later. And, be careful that

$$q_n(z^{(n)}) \neq p(z; \boldsymbol{\pi}).$$

Decomposition of Log Likelihood

- We start from one pair of observed and latent variables, *i.e.*, $\{\mathbf{x}, z\}$. Utilizing $q(z)$, we have

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = \mathbb{E}_{q(z)}\left[\ln\left(\frac{p(\mathbf{x};\boldsymbol{\theta})\cdot q(z)}{q(z)}\right)\right] = \mathbb{E}_{q(z)}\left[\ln\left(\frac{p(\mathbf{x},z;\boldsymbol{\theta})}{q(z)}\cdot\frac{q(z)}{p(z|\mathbf{x};\boldsymbol{\theta})}\right)\right]$$
$$= \mathbb{E}_{q(z)}\left[\ln\left(\frac{p(\mathbf{x},z;\boldsymbol{\theta})}{q(z)}\right)\right] + \mathbb{E}_{q(z)}\left[\ln\left(\frac{q(z)}{p(z|\mathbf{x};\boldsymbol{\theta})}\right)\right]$$

- It is natural to extend the above decomposition to the log likelihood of the whole data set $\mathcal{D}$, *i.e.*,

$$\begin{aligned}\ln p(\mathcal{D};\boldsymbol{\theta}) =& \sum_{n=1}^{N}\mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{p(\mathbf{x}^{(n)},z^{(n)};\boldsymbol{\theta})}{q_n(z^{(n)})}\right)\right] \\ &+ \sum_{n=1}^{N}\mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{q_n(z^{(n)})}{p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta})}\right)\right] \\ =& \mathcal{L}(\mathbf{q};\boldsymbol{\theta}) + \mathrm{KL}\big(\mathbf{q}(\mathbf{z})||p(\mathbf{z}|\mathcal{D};\boldsymbol{\theta})\big)\end{aligned} \tag{1}$$

Note that the summation over the latent variable ($i.e.$, $\mathbb{E}_{q(z)}$) is out of log in both $\mathcal{L}(\mathbf{q};\boldsymbol{\theta})$ and $\mathrm{KL}(\mathbf{q}(\mathbf{z})||p(\mathbf{z}|\mathcal{D};\boldsymbol{\theta}))$.

Decomposition of Log Likelihood

Theorem

$$\ln p(\mathcal{D};\boldsymbol{\theta}) \geq \mathcal{L}(\mathbf{q};\boldsymbol{\theta}), \ \forall \mathbf{q}, \boldsymbol{\theta}.$$

Proof 1: Since $\ln(\cdot)$ is concave, utilizing the Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_{q(z)}\left[\ln\left(\frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(z)}\right)\right] &\leq \ln \mathbb{E}_{q(z)}\left(\frac{p(\mathbf{x},z;\boldsymbol{\theta})}{q(z)}\right) \\ &= \ln \sum_{k}^{K} q(z=k)\cdot\frac{p(\mathbf{x},z=k;\boldsymbol{\theta})}{q(z=k)} = \ln p(\mathbf{x};\boldsymbol{\theta}).\end{aligned}$$

Then, it is easy to prove the above theorem.

Decomposition of Log Likelihood

Theorem

$$\ln p(\mathcal{D};\boldsymbol{\theta}) \geq \mathcal{L}(\mathbf{q};\boldsymbol{\theta}),\ \forall \mathbf{q},\boldsymbol{\theta}.$$

Proof 2: According to the non-negative property of KL divergence, we have

$$\mathrm{KL}\big(\mathbf{q}(\mathbf{z})||p(\mathbf{z}|\mathcal{D};\boldsymbol{\theta})\big) \geq 0,$$

where the equality holds only when $\mathbf{q}(\mathbf{z}) = p(\mathbf{z}|\mathcal{D};\boldsymbol{\theta})$. Utilizing the decomposition of the log likelihood (*i.e.*, Eq. (1)), we can prove the above theorem.

Maximizing the Lower Bound of Log Likelihood

- Since learning $\boldsymbol{\theta}$ by maximizing $\ln p(\mathcal{D}; \boldsymbol{\theta})$ is difficult, we resort to maximize its lower bound $\mathcal{L}(\mathbf{q}; \boldsymbol{\theta})$ with some auxiliary distribution $\mathbf{q}(\mathbf{z})$, *i.e.*,

$$\max_{\mathbf{q}(\mathbf{z}), \boldsymbol{\theta}} \mathcal{L}(\mathbf{q}; \boldsymbol{\theta}) \equiv \max_{\mathbf{q}(\mathbf{z}), \boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[\ln \left(\frac{p(\mathbf{x}^{(n)}, z^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})} \right) \right],$$

  with the constraint $\sum_{z^{(n)}=1}^{K} q_n(z^{(n)}) = 1, \forall n$.
- We adopt the coordinate descent algorithm to solve the above optimization problem, with the following alternative steps:
  - Given $\boldsymbol{\theta}$, update $\mathbf{q}(\mathbf{z})$;
  - Given $\mathbf{q}(\mathbf{z})$, update $\boldsymbol{\theta}$.
- The whole algorithm for fitting the latent variable model is called Expectation Maximization (EM) algorithm.

Expectation Maximization: E step

Given $\boldsymbol{\theta}$, update $\mathbf{q}(\mathbf{z})$ by solving the following sub-problem:

$$
\begin{aligned}
\max_{\mathbf{q}(\mathbf{z})} \mathcal{L}(\mathbf{q}; \boldsymbol{\theta}) &\equiv \max_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{p(\mathbf{x}^{(n)}, z^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right)\right] \\
&\equiv \max_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}) \cdot p(\mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right)\right] \\
&\equiv \max_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right) + \ln p(\mathbf{x}^{(n)}; \boldsymbol{\theta})\right] \\
&\equiv \max_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right)\right] + \text{constant} \\
&\equiv \min_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln\left(\frac{q_n(z^{(n)})}{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta})}\right)\right] \\
&\equiv \min_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathrm{KL}\big(q_n(z^{(n)}) || p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta})\big),
\end{aligned}
$$

with the constraint $\sum_{k=1}^{K} q_n(z^{(n)} = k) = 1, \forall n$.

Expectation Maximization: E step

- Given $\boldsymbol{\theta}$, update $\mathbf{q}(\mathbf{z})$ by solving the following sub-problem:

$$\min_{\mathbf{q}(\mathbf{z})} \sum_{n=1}^{N} \mathrm{KL}\big(q_n(z^{(n)})||p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta})\big),$$

  with the constraint $\sum_{k=1}^{K} q_n(z^{(n)}=k)=1, \forall n$.

- According to the property of KL divergence, it is easy to find the optimal solution, as follows:

$$q_n^*(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}).$$

  And this solution also satisfies the equality constraint.

- It is interesting to see that
  - The optimal auxiliary distribution $q_n^*(z^{(n)})$ is exactly the posterior distribution $p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta})$
  - Since $\mathrm{KL}\big(\mathbf{q}^*(\mathbf{z})||p(\mathbf{z}|\mathcal{D};\boldsymbol{\theta})\big)=0$, then

$$\ln p(\mathcal{D};\boldsymbol{\theta}) = \mathcal{L}(\mathbf{q}^*;\boldsymbol{\theta}).$$

    It means that the gap between $\ln p(\mathcal{D};\boldsymbol{\theta})$ and its lower bound $\mathcal{L}(\mathbf{q}^*;\boldsymbol{\theta})$ becomes 0, given the current $\boldsymbol{\theta}$.

Expectation Maximization: M step

- Given $\mathbf{q}(\mathbf{z})$, update $\boldsymbol{\theta}$ by solving the following sub-problem:

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{q}; \boldsymbol{\theta}) \equiv \max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\ln \left(\frac{p(\mathbf{x}^{(n)}, z^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})} \right) \right]$$

$$\equiv \max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n\left(z^{(n)}\right)} \left[\log p \left(\mathbf{x}^{(n)}, z^{(n)}; \boldsymbol{\theta} \right) \right] - \underbrace{\mathbb{E}_{q_n\left(z^{(n)}\right)} \left[\log q_n \left(z^{(n)} \right) \right]}_{\text{constant w.r.t. } \boldsymbol{\theta}}$$

- Substitute in $q_n\left(z^{(n)}\right) = p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right)$:

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \mathbb{E}_{p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right)} \left[\log p \left(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta} \right) \right]$$

- This is the expected complete data log-likelihood, which is easy to optimize.

Rcap from last lecture - GMM E-Step: expected log likelihood

Once we computed $\gamma_k^{(n)} = p\left(z^{(n)} = k \mid \mathbf{x}^{(n)}\right)$, we can compute the expected log likelihood, as follows:

$$\begin{aligned}
&\sum_n \mathbb{E}_{P(z^{(n)}|\mathbf{x}^{(n)})}\left[\ln\left(P(\mathbf{x}^{(n)}, z^{(n)} \mid \Theta)\right)\right] \\
=&\sum_n \sum_k \gamma_k^{(n)} \left(\ln\left(P(z^{(n)} = k \mid \Theta)\right) + \ln\left(P(\mathbf{x}^{(n)} \mid z^{(n)} = k, \Theta)\right)\right) \\
=&\sum_n \sum_k \gamma_k^{(n)} \left(\ln\left(\pi_k\right) + \ln\left(\mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)\right) \\
=&\sum_n \sum_k \gamma_k^{(n)} \ln\left(\pi_k\right) + \sum_n \sum_k \gamma_k^{(n)} \ln\left(\mathcal{N}(\mathbf{x}^{(n)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right),
\end{aligned}$$

where $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$. Note that the above expectation is fully decomposed to each data n and each cluster k, which will facilitate the parameter learning in the following maximization step.

Rcap from last lecture - GMM M-Step

Maximization step:

- Given the posterior probability $\gamma_k^{(n)} = p\left(z^{(n)} = k \mid \mathbf{x}^{(n)}\right)$, we want to update the model parameters $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ by maximizing the expected log likelihood, *i.e.*,

$$\max_{\Theta} \sum_{n}^{N} \sum_{k}^{K} \gamma_k^{(n)} \ln(\pi_k) + \sum_{n}^{N} \sum_{k}^{K} \gamma_k^{(n)} \ln\left(\mathcal{N}(\mathbf{x}^{(n)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right), \text{ s.t. } \sum_{k}^{K} \pi_k = 1.$$

- Following the derivations introduced in previous slides (see page 12-17), it is easy to obtain the following solutions:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^{(n)} \mathbf{x}^{(n)}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^{(n)} \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_k\right)^{\top}$$

$$\pi_k = \frac{N_k}{N}, \text{ with } N_k = \sum_{n=1}^{N} \gamma_k^{(n)}$$

Expectation Maximization: Summary

- The EM algorithm alternates between making the bound tight at the current parameter values and then optimizing the lower bound
- If the current parameter value is $\boldsymbol{\theta}^{\text{old}}$:
  - **E-step**: Given $\boldsymbol{\theta}^{\text{old}}$, we update the auxiliary distribution $\mathbf{q}(\mathbf{z})$ to make the bound tight:

$$\mathbf{q}(\mathbf{z}) = \underset{\mathbf{q}(\mathbf{z})}{\text{argmax}}\ \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}). \tag{2}$$

    It leads to $q_n\left(z^{(n)}\right) = p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right), \forall n$, and makes

$$\log p\left(\mathcal{D}; \boldsymbol{\theta}^{\text{old}}\right) = \mathcal{L}\left(q; \boldsymbol{\theta}^{\text{old}}\right)$$

  - **M-step**: Given $\mathbf{q}(\mathbf{z})$ updated above, we update $\boldsymbol{\theta}$ by optimizing the lower bound:

$$\begin{aligned}\boldsymbol{\theta}^{\text{new}} &= \underset{\boldsymbol{\theta}}{\text{argmax}}\ \mathcal{L}(q, \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\text{argmax}}\ \sum_{n=1}^{N} \mathbb{E}_{q_n\left(z^{(n)}\right)}\left[\log \frac{p\left(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}\right)}{q_n\left(z^{(n)}\right)}\right]\end{aligned}$$

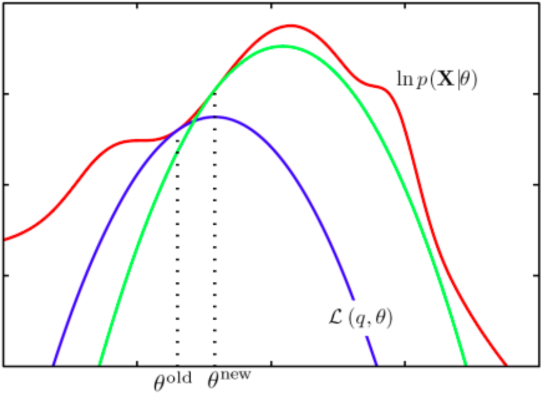
EM Convergence

- We can deduce that an iteration of EM will improve the log-likelihood by using the fact that the bound is tight at $\boldsymbol{\theta}^{\text{old}}$ after the E-step
- Let q denote the q_n after the E-step, *i.e.*, $q_n\left(z^{(n)}\right) = p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right)$

$$\begin{aligned}
\log p\left(\mathcal{D}; \boldsymbol{\theta}^{\text{new}}\right) &\geq \mathcal{L}\left(q, \boldsymbol{\theta}^{\text{new}}\right) && \text{since } \log p(\mathcal{D}; \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta}) \text{ always holds} \\
&\geq \mathcal{L}\left(q, \boldsymbol{\theta}^{\text{old}}\right) && \text{since } \boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \mathcal{L}(q, \boldsymbol{\theta}) \\
&= \log p\left(\mathcal{D}; \boldsymbol{\theta}^{\text{old}}\right) && \text{since } \log p\left(\mathcal{D}; \boldsymbol{\theta}^{\text{old}}\right) = \mathcal{L}\left(q; \boldsymbol{\theta}^{\text{old}}\right)
\end{aligned}$$

- It tells that the log likelihood objective keeps increasing after each iteration of EM, until convergence.

EM Visualization



- The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values.

Revisiting Gaussian Mixture Models

- Let's revisit the Gaussian mixture models from last lecture and derive the updates using our general EM algorithm
- Recall our model was:

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_z p(\mathbf{x}, z; \boldsymbol{\theta}) = \sum_z p(\mathbf{x}|z;\boldsymbol{\theta})p(z|\boldsymbol{\theta}) \tag{3}$$

$$p(z=k;\boldsymbol{\theta}) = \pi_k, \ \sum_{k=1}^{K} \pi_k = 1. \tag{4}$$

$$p(\mathbf{x} \mid z=k;\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{5}$$

- In this scenario, we have $\boldsymbol{\theta} = \{\boldsymbol{\pi_k}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\}_{k=1}^{K}$.

E-Step for Gaussian Mixture Models

- Let the current parameters be $\boldsymbol{\theta}^{\text{old}} = \left\{\boldsymbol{\pi}_{\boldsymbol{k}}^{\text{old}}, \boldsymbol{\mu}_{\boldsymbol{k}}^{\text{old}}, \boldsymbol{\Sigma}_{\boldsymbol{k}}^{\text{old}}\right\}_{k=1}^{K}$
- **E-step**: For all n, set $q_n\left(z^{(n)}\right) = p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right)$, *i.e.*,

$$\begin{aligned}\gamma_k^{(n)} :=& q_n\left(z^{(n)} = k\right) = p\left(z^{(n)} = k \mid \mathbf{x}^{(n)}; \theta^{\text{old}}\right) \\ =& \frac{\pi_k^{\text{old}} \mathcal{N}\left(\mathbf{x}^{(n)} \mid \boldsymbol{\mu}_{\boldsymbol{k}}^{\text{old}}, \boldsymbol{\Sigma}_{\boldsymbol{k}}^{\text{old}}\right)}{\sum_{j=1}^{K} \pi_j^{\text{old}} \mathcal{N}\left(\mathbf{x}^{(n)} \mid \mu_j^{\text{old}}, \Sigma_j^{\text{old}}\right)}\end{aligned}$$

M-Step for Gaussian Mixture Models

M-step:

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \mathbb{E}_{q_n\left(z^{(n)}\right)} \left[\log p\left(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}\right)\right], \text{ s.t. } \sum_{k=1}^{K} \pi_k = 1.$$

- Substitute in:
  - $\log p\left(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}\right) = \sum_{k=1}^{K} 1_{\{z^{(n)}=k\}} \left(\log \pi_k + \log \mathcal{N}\left(\mathbf{x}^{(n)}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right)\right)$
  - $q_n\left(z^{(n)}\right) = p\left(z^{(n)} \mid \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}\right)$:
- We have:

$$\begin{aligned}\boldsymbol{\theta}^{\text{new}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \mathbb{E}_{q_n\left(z^{(n)}\right)} \left[\sum_{k=1}^{K} 1_{\{z^{(n)}=k\}} \left(\log \pi_k + \log \mathcal{N}\left(\mathbf{x}^{(n)}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right)\right)\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^{(n)} \left(\log \pi_k + \log \mathcal{N}\left(\mathbf{x}^{(n)}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right)\right)\end{aligned}$$

M-Step for Gaussian Mixture Models

M-step:

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^{(n)} \left(\log \pi_k + \log \mathcal{N}\left(\mathbf{x}^{(n)}; \boldsymbol{\mu}_{\boldsymbol{k}}, \boldsymbol{\Sigma}_{\boldsymbol{k}} \right) \right)$$

- Taking derivatives and setting to zero, and utilizing the constraint $\sum_{k=1}^{K} \pi_k = 1$, we get the exactly same updates from last lecture:

$$\boldsymbol{\mu}_{\boldsymbol{k}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^{(n)} \mathbf{x}^{(n)}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{k}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^{(n)} \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_{\boldsymbol{k}} \right) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}_{\boldsymbol{k}} \right)^T$$

$$\pi_k = \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n=1}^{N} \gamma_k^{(n)}.$$

EM Recap

- A general algorithm for optimizing many latent variable models, such as GMMs, mixture of Bernoulli distribution
- Iteratively computes a lower bound then optimizes it.
- Converges but maybe to a local minima.
- Can use multiple restarts.
- Can initialize from k-means for mixture models.
- Limitation - need to be able to compute $p(z|\mathbf{x};\boldsymbol{\theta})$, not possible for more complicated models.

References

- Further reading 1: Chapter 9 in the book "Pattern Recognition and Machine Learning". Link
- Further reading 2: Wikipedia https://en.wikipedia.org/wiki/Expectatio
E2%80%93maximization_algorithm
- Demo with code: https://www.kaggle.com/code/charel/learn-by-examp
notebook