Machine Learning for Data Science (CS 4786)

Lecture 15: EM Algorithm for Gaussian Mixture Models and Why EM works!

1 Gaussian Mixture Models

Each $\theta \in \Theta$ consist of mixture distribution π which is a distribution over the choices of the K clusters, $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ the choices of the K means for the corresponding gaussians and $\Sigma_1, \ldots, \Sigma_K$ the choices of the K covariance matrices. The latent variables are $c_1, \ldots, c_n$ the cluster assignments for the n points and $x_1, \ldots, x_n$ are the n observations.

1.1 E-step

On iteration i, for each data point $t \in [n]$, set

$$Q_t^{(i)}(c_t) = P(c_t | x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned}
Q_t^{(i)}(c_t) &= P(c_t | x_t, \theta^{(i-1)}) \\
&\propto p(x_t | c_t \theta^{(i-1)}) \times P(c_t | \theta^{(i-1)}) \\
&\propto \frac{1}{\sqrt{(2\pi)^d |\Sigma_{c_t}|}} \exp\left(-(x_t - \mu_{c_t})^\top \Sigma_{c_t} (x_t - \mu_{c_t})/2\right) \pi_{c_t}
\end{aligned}$$

1.2 M-step for GMM

For the M-step (for MLE) we would like to find

$$\theta = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta)$$

To this end note that

$$\begin{aligned}
\sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) &= \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(\log \phi(x_t | \mu_k, \Sigma_k) + \log \pi_k\right) \\
&= \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(\frac{1}{2} \log\left(\frac{1}{(2 * 3.14)^d |\Sigma_k|}\right) - \frac{1}{2}(x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) + \log \pi_k\right) \\
&= \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(-\frac{1}{2} \log \det(\Sigma_k) - \frac{1}{2}(x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) + \log \pi_k\right) + \text{constant terms}
\end{aligned}$$

For notational convenience define:

$$L(\mu_{1:K}, \Sigma_{1:K}, \pi) = \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(-\frac{1}{2} \log \det(\Sigma_k) - \frac{1}{2}(x_t - \mu_k)^\top \Sigma_k^{-1}(x_t - \mu_k) + \log \pi_k \right)$$

Our goal is to find parameters that maximize $L(\mu_{1:K}, \Sigma_{1:K}, \pi)$.

M-step for mean: To optimize with respect to mean we take derivative and equate to 0,

$$\begin{aligned} \frac{\partial}{\partial \mu_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) &= -\frac{1}{2} \frac{\partial}{\partial \mu_k} \left(\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t - \mu_k)^\top \Sigma_k^{-1}(x_t - \mu_k) \right) \\ &= -\sum_{t=1}^{n} Q_t^{(i)}(k) \Sigma_k^{-1}(x_t - \mu_k) = -\Sigma_k^{-1} \left(\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t - \mu_k) \right) \end{aligned}$$

To maximize over μ_k we set derivative equal to 0. Hence

$$\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t - \mu_k) = \sum_{t=1}^{n} Q_t^{(i)}(k) x_t - \mu_k \left(\sum_{t=1}^{n} Q_t^{(i)}(k) \right) = 0$$

Or equivallently:

$$\mu_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

M-step for mixture distribution: Since we want to optimize over π subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we want to optimize the following term w.r.t. π_k and λ

$$L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$$

Hence taking derivative of above w.r.t. π we get,

$$\frac{\partial}{\partial \pi_k} \left(L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda(1 - \sum_{k=1}^{K} \pi_k) \right) = \frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) - \lambda$$

But,

$$\frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) = \frac{\partial}{\partial \pi_k} \sum_{t=1}^{n} Q_t^{(i)}(k) \log(\pi_k) = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{\pi_k}$$

Hence,

$$\frac{\partial}{\partial \pi_k} \left(L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda(1 - \sum_{k=1}^{K} \pi_k) + \sum_{i=1}^{K} \lambda_i \pi_i \right) = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{\pi_k} - \lambda$$

Setting derivative to 0 we discover that

$$\pi_k \propto \sum_{t=1}^{n} Q_t^{(i)}(k)$$

Since π needs to be a valid distribution, this yields that

$$\pi_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{\sum_{k=1}^{K}\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

However notice that since $Q_t^{(i)}$ is a distribution over K clusters, $\sum_{k=1}^{K}\sum_{t=1}^{n} Q_t^{(i)}(k) = \sum_{t=1}^{n} 1 = n$. Hence,

$$\pi_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}$$

M-step for Covariance: This one needs being able to take derivative w.r.t. matrices and so I will only sketch the proof here. Let us consider optimizing w.r.t. some Σ_k. It makes the problem easier if we instead think of the problem as optimizing over Σ_k^{-1} and then invert the solution.

Here are two facts that come in handy:

$$\frac{\partial}{\partial \mathbf{X}} \log\det(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}$$

and for any vector v,

$$\frac{\partial}{\partial \mathbf{X}} v^\top \mathbf{X} v = v v^\top$$

Now note that

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) &= \frac{\partial}{\partial \Sigma_k}\left(\sum_{t=1}^{n} Q_t^{(i)}(k)\left(-\frac{1}{2}\log\det(\Sigma_k) - \frac{1}{2}(x_t-\mu_k)^\top \Sigma_k^{-1}(x_t-\mu_k)\right)\right)\\
&= \left(\sum_{t=1}^{n} Q_t^{(i)}(k)\left(\frac{1}{2}(\Sigma_k^{-1})^{-1} - \frac{1}{2}(x_t-\mu_k)(x_t-\mu_k)^\top\right)\right)
\end{aligned}$$

Hence equating to 0 we get that

$$\Sigma_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t-\mu_k)(x_t-\mu_k)^\top}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

that is the weighted sample variance. (**there is a bit of a fudge here since μ_k is also an optimiation variable. But we skip the details of this for now.**)

2 EM Algorithm: Why it works?

Log likelihood only decreases after one iteration of EM algorithm. Why?

We will show below that EM algorithm can never lead to a worsening of the objective in any step and can only imrpvie likelihood.

$$
\begin{aligned}
\log P(x_1, \ldots, x_n|\theta^{(i+1)}) &= \sum_{t=1}^{n} \log P(x_t|\theta^{(i+1)}) && \text{(x's drawn independently)} \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P(x_t, c_t|\theta^{(i+1)}) \right) && \text{(marginalizing over } c_t\text{'s)} \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} \frac{Q^{(i+1)}(c_t)}{\color{red}{Q^{(i+1)}(c_t)}} \color{red}{P(x_t, c_t|\theta^{(i+1)})} \right)
\end{aligned}
$$

Logarithm is a concave function and by Jensen's inequality $\log(E[X]) \geq E[\log(X)]$ for any R.V. X. Treat the term in red as the random variable and the probability distribution is specified by $Q^{(i+1)}$, now using Jensen,

$$
\begin{aligned}
&\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(\frac{P(x_t, c_t|\theta^{(i+1)})}{Q^{(i+1)}(c_t)} \right) \\
&= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i+1)}) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right)
\end{aligned}
$$

Since in M-step $\theta^{(i+1)}$ is exactly the maximizer of $\sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i+1)}) \right)$, we conclude that this term is larger than $\sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i)}) \right)$ and so

$$
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i)}) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right)
$$

Now note that $P(x_t, c_t|\theta^{(i)}) = P(c_t|x_t, \theta^{(i)}) P(x_t|\theta^{(i)}) = Q^{(i+1)}(c_t) P(x_t|\theta^{(i)})$ and so,

$$
\begin{aligned}
&= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t|\theta^{(i)}) \times Q^{(i+1)}(c_t) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right) \\
&= \sum_{t=1}^{n} \log P(x_t|\theta^{(i)}) \\
&= \log P(x_1, \ldots, x_n|\theta^{(i)})
\end{aligned}
$$

Hence we have shown that running the EM algorithm yields, $\log P(x_1, \ldots, x_n|\theta^{(i)}) \leq \log P(x_1, \ldots, x_n|\theta^{(i+1)})$, that is the Likelihood value never decreases and could only improve.