More discussion on EM algorithm

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

March 9, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Finished discussing Gaussian mixture modeling
  - Indirectly develop EM algorithm to derive GMM parameters
- Today's agenda
  - Discuss an alternative view of EM algorithm
  - Prove that EM maximizes log likelihood while maximizing the lower bound

Simplification of log likelihood using Jensen inequality (recap)

- We wanted to maximize the log likelihood

$$L = \log p(\mathbf{X} \mid \Delta, \boldsymbol{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right))$$

- Recall that above log likelihood was obtained by marginalizing over latent variable z_{nk}
- Summation inside logarithm makes finding optimal parameter $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}$ difficult
- EM algorithm overcomes this problem by deriving a lower bound on this likelihood
- To calculate lower bound, multiply and divide inside summation over k by latent variable q_{nk}
  - q_{nk} is some probability distribution over the K components for the nth object

$$L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} q_{nk} \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} = \sum_{n=1}^{N} \log \mathbf{E}_{q_{nk}} \left\{ \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right\}$$

- Applying Jensen's inequality, we can lower bound the log likelihood:

$$L = \sum_{n=1}^{N} \log \mathbf{E}_{q_{nk}} \left\{ \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right\} \geq \sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right\}$$

Simplification of log likelihood using Jensen inequality (2)

- Expanding the expression gives us something which we could maximize

$$\mathcal{B} = \sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right\} = \sum_{n=1}^{N} \sum_{k=1}^{K} q_{nk} \log \left(\frac{\pi_k p\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{q_{nk}} \right)$$

- We maximized lower bound $\mathcal{B}$ to calculate $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}$
- q_{nk} could be interpreted as the posterior probability that object n was generated by component k

$$p(z_{nk} = 1 \mid \mathbf{x}_n) = \frac{p\left(z_{nk} = 1\right) p\left(\mathbf{x}_n \mid z_{nk} = 1\right)}{\sum_{j=1}^{K} p\left(z_{nj} = 1\right) p\left(\mathbf{x}_n \mid z_{nj} = 1\right)} = q_{nk}$$

- Re-state the EM algorithm (remember these steps for next two slides)
  - Calculate posterior distribution of latent variable z_{nk} i.e., $p(z_{nk} = 1 \mid \mathbf{x}_n)$, which is denoted as q_{nk}
  - Calculate $\mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$ and maximize $\mathcal{B} = \sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{X}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$ to calculate $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}$
  - Iterate the above two steps
- Quantity is called complete data likelihood (CDLL)

Discussion on variable z_{nk}

- In many applications there will be characteristics of objects of interest, not provided in given data
- GMM: we used indicator variables z_{nk}, where $z_{nk} = 1$ if nth object was generated by kth component
- These variables (also known as latent variables) do not really exist but enable us to build models
    - z_{nk} is a latent variable – it does not exist in reality

An alternative view of EM algorithm (1)[1]

- Present a view of the EM algorithm that recognizes key role played by latent variables
- Goal of the EM algorithm is to find maximum likelihood solutions for models with latent variables
  - Denote the set of all observed data by $\mathbf{X}$, in which the nth row represents $\mathbf{x}_n^T$
  - Denote the set of all latent variables by $\mathbf{Z}$, in which the nth row represents $\mathbf{Z}_n^T$
  - Set of all model parameters is denoted by $\boldsymbol{\theta}$, which in GMM are $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$
- Log likelihood function can be expressed as

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log\left\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right\} \tag{1}$$

- Recall the log likelihood for GMM model

$$L = \log p(\mathbf{X} \mid \Delta, \boldsymbol{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p\left(\mathbf{X}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)) \tag{2}$$

- GMM likelihood in (2) has same form as (1) with $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}\}$
- Observe that the summation over latent variables appears inside the logarithm
  - $\log p(\mathbf{X}|\boldsymbol{\theta})$ is not easy to maximize due to this summation

[1] PRML, Chap 9.3

An alternative view of EM algorithm (2)

- Suppose for each observation in $\mathbf{X}$, we were told the corresponding value of the latent variable $\mathbf{Z}$
- We call $\{\mathbf{X}, \mathbf{Z}\}$ complete data set, and we refer to the actual observed data $\mathbf{X}$ as incomplete
  - Complete-data log likelihood (CDLL) is $\log p(\mathbf{X}, \mathbf{Z}|\theta)$
  - EM algorithm ssumes that maximization of CDLL is straightforward
- In practice, however, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$ but only incomplete data $\mathbf{X}$
  - For example in GMM, we did not know the assignments z_{nk}
- Our knowledge of latent variables $\mathbf{Z}$ is given only by posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$
  - For example in GMM, we knew only $p(z_{nk} = 1 \mid \mathbf{X}_n, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}\}$
- This implies that we cannot consider complete-data log likelihood (CDLL)
  - For example in GMM, we did not use consider CDLL $\log \frac{\pi_k p(\mathbf{X}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}}$
- Instead consider expected value of CDLL under posterior of latent variable i.e., $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$
  - For example in GMM, calculate $\mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{X}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$
  - This corresponds to E step of EM algorithm
- In subsequent M step, we maximize this expectation
  - For example, in GMM we maximized $\mathcal{B} = \sum_{n=1}^{N} \mathbf{E}_{q_{nk}} \left\{ \log \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}} \right\}$

Summary of the alternative view of EM algorithm

- If the current estimate for the parameters is denoted $\boldsymbol{\theta}^{old}$, then EM algorithm is
  1. E step: use current parameter values $\boldsymbol{\theta}^{old}$ to find posterior of latent variables $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$
  2. Use $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ to find expectation of CDLL evaluated for some general $\boldsymbol{\theta}$

$$\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

  3. M step: determine the revised parameter estimate $\boldsymbol{\theta}^{new}$ by maximizing this function

$$\boldsymbol{\theta}^{new} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}).$$

- Note that in definition of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$, logarithm acts directly on CDLL $\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$
  - So the corresponding M-step maximization will, by supposition, be tractable
- EM calculates
  - posterior distribution over hidden variable in Step 1. For example in GMM, we calculated $p(z_{nk} = 1 \mid \mathbf{X}_n, \boldsymbol{\theta}^{old})$ with $\boldsymbol{\theta}^{old} = \{\boldsymbol{\mu}_k^{old}, \boldsymbol{\Sigma}_k^{old}, \boldsymbol{\pi}^{old}\}$
  - point estimate of parameter $\boldsymbol{\theta}$ by maximizing expected value (using above posterior) of CDLL
  - For example, we maximized $\mathbf{E}_{q_{nk}}\left\{\log \frac{\pi_k p(\mathbf{X}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{nk}}\right\}$ to calculate $\boldsymbol{\theta}^{new} = \{\boldsymbol{\mu}_k^{new}, \boldsymbol{\Sigma}_k^{new}, \boldsymbol{\pi}^{new}\}$

Formal proof that EM algorithm maximize the log likelihood (1)

- Recall we want to maximize the log likelihood $\log p(\mathbf{X}|\boldsymbol{\theta})$ which can equivalently be written as

$$
\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &\overset{(a)}{=} \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) \overset{(b)}{=} \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} \right) \\
&\overset{(c)}{=} \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} \right) \\
&= \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)}_{\mathcal{L}(q,\boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right)}_{KL(q||p)} \\
&= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p)
\end{aligned}
$$

- Equality (a) is obtained because $\sum_{\mathbf{Z}} q(\mathbf{Z}) = 1$. Equality (b) uses Bayes rule $p(\mathbf{X}|\boldsymbol{\theta}) = \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}$.
- Equality (c) is obtained multiply and divide by $q(\mathbf{Z})$ inside log
- Kullback-Leibler divergence $KL(q \parallel p) \geq 0$, with equality if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$

Formal proof that EM algorithm maximize the log likelihood (2)

- EM algorithm is a two-stage iterative technique for finding maximum likelihood solutions

$$
\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \parallel p), \text{ where} \qquad (3)\\
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)\\
KL(q \parallel p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right)
\end{aligned}
$$

- Suppose current value of parameter vector is $\boldsymbol{\theta}^{\text{old}}$
- E step maximizes the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to $q(\mathbf{Z})$ by fixing $\boldsymbol{\theta}^{old}$
    - Note that $\log p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ in (3) does not depend on $q(\mathbf{Z})$, and will remain constant in this maximization
- Largest value of $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ will occur when $KL(q \parallel p) = 0$ i.e., $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
    - lower bound will now be equal to the log likelihood

Formal proof that EM algorithm maximize the log likelihood (3)

- EM algorithm is a two-stage iterative technique for finding maximum likelihood solutions

$$\begin{aligned} \log p(\mathbf{X}|\boldsymbol{\theta}) &= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p), \text{ where} \\ \mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right) \\ KL(q \parallel p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right) \end{aligned} \tag{4}$$

- M step fixes $q(\mathbf{Z})$, and maximizes $L(q,\boldsymbol{\theta})$ wrt $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$
  - This will increase $\mathcal{L}$ (unless it is already maximum), which will increase log likelihood
- Because $q(\mathbf{Z})$ is determined using $\boldsymbol{\theta}^{\text{old}}$ rather than $\boldsymbol{\theta}^{\text{new}}$, and is held fixed during M step
  - It will not equal new posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{new}})$, and there will be a nonzero KL divergence
- Increase in log likelihood function is therefore greater than increase in lower bound, and it increases

Formal proof that EM algorithm maximize the log likelihood (4)

- EM algorithm is a two-stage iterative technique for finding maximum likelihood solutions

$$
\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p), \text{ where} \\
\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right) \\
KL(q \parallel p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right)
\end{aligned}
$$

- To see, what is M step maximizing, we substitute $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$ in $\mathcal{L}(q,\boldsymbol{\theta})$:

$$
\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \log\left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})}\right) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \log\left(p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \log\left(p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})\right) \\
&= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + \text{constant (entropy of } q \text{ distribution)}
\end{aligned}
$$

- M step maximizes expectation of the complete-data log likelihood (CDLL) $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$