Chapter 11

stevenjin8

January 16, 2020

Comments

I found that the formulae for the EM algorithms could have been a bit more explicit. More specifically, I did not really understand what Q was until I realized that

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) &= \mathbb{E}[\ell_c(\boldsymbol{\theta}) | \mathcal{D}, \boldsymbol{\theta}^{t-1}] \\ &= \sum \mathbb{E}[\log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) | \mathbf{x}_i, \boldsymbol{\theta}^{t-1}]. \end{aligned}$$

In the case of mixture models with unknown latent variables, we can further expand to

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = \sum_{i=1}^{N} \sum_{k=1}^{L} \log(p(\mathbf{x}_i, z_i = k | \boldsymbol{\theta})) p(z_i = k | \mathbf{x}_i, \boldsymbol{\theta}^{t-1})$$

In the case of GMMs, I think a more straightforward derivation of Q is

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) &= \mathbb{E}\left[\sum_i \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) \middle| \mathcal{D}, \boldsymbol{\theta}^{t-1}\right] \\ &= \sum_i \mathbb{E}\left[\log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) \middle| \mathbf{x}_i, \boldsymbol{\theta}^{t-1}\right] \\ &= \sum_i \sum_k \log[p(\mathbf{x}_i, z_i = k | \boldsymbol{\theta})] p(z_i = k | \mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \\ &= \sum_i \sum_k r_{ik} \log[p(\mathbf{x}_i | z_i = k, \boldsymbol{\theta}) p(z_i = k | \boldsymbol{\theta})] \\ &= \sum_i \sum_k r_{ik} \log[\pi_k p(\mathbf{x}_i | z_i = k, \boldsymbol{\theta}))] \\ &= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k \log p(\mathbf{x}_i | z_i = k, \boldsymbol{\theta})). \end{aligned}$$

Note that r_{ik} is with respect to $\boldsymbol{\theta}^{t-1}$ and π_k is with respect to $\boldsymbol{\theta}$.

Exercises

Exercise 1

Recall that with $D = 1$, equation 11.61 is

$$\mathcal{T}\left(x_i|\mu,\sigma^2,\upsilon\right) = \int_0^\infty \mathcal{N}\left(x_i \mid \mu, \sigma^2/z\right) \mathrm{Ga}\left(z|\frac{\upsilon}{2},\frac{\upsilon}{2}\right) dz \tag{11.61'}$$

We have to show that this is equivalent to

$$\mathcal{T}\left(x_i|\mu,\sigma^2,\upsilon\right) = \frac{\Gamma((\upsilon+1)/2)}{\Gamma(\upsilon/2)\sqrt{\upsilon\pi}\sigma}\left[1+\frac{1}{\upsilon}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]^{-\frac{\upsilon+1}{2}}. \tag{2.71'}$$

Recall that the pdf of the gamma distribution is

$$\mathrm{Ga}(T|a,b) = \frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb},$$

and the gamma function is

$$\Gamma(u) = \int_0^\infty x^{u-1}e^{-x}dx.$$

With that out of the way, we first expand equation 11.61':

$$\frac{1}{\sigma\sqrt{2\pi}\Gamma(\upsilon/2)}\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}\int\sqrt{z}\exp\left[\frac{-z}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\exp\left[\frac{\upsilon-1}{2}\right]z^{\frac{\upsilon-2}{2}}dz$$
$$=\frac{1}{\sigma\sqrt{2\ pi}\Gamma(\upsilon/2)}\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}\int\exp\left[\frac{-z}{2}\left(\left(\frac{x-\mu}{\sigma}\right)^2+\upsilon\right)\right]z^{\frac{\upsilon-1}{2}}dz.$$

Performing a u-substitution with $u = z\gamma$ where

$$\gamma = \frac{1}{2}\left(\left(\frac{x-\mu}{\sigma}\right)^2+\upsilon\right)$$

gives

$$\frac{1}{\sigma\sqrt{2\pi}\Gamma(\upsilon/2)}\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}\int e^{-u}u^{\frac{\upsilon-1}{2}}\gamma^{-\left(\frac{\upsilon+1}{2}\right)}du.$$

Using the pdf of the gamma distribution, we have

$$\frac{\Gamma(\frac{\upsilon-1}{2})}{\sigma\sqrt{2\ pi}\Gamma(\upsilon/2)}\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}\gamma^{-\frac{\upsilon+1}{2}}$$

$$= \frac{\Gamma(\frac{\upsilon-1}{2})}{\sigma\sqrt{\upsilon\pi}\Gamma(\upsilon/2)}\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon+1}{2}}\left(\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2+\frac{\upsilon}{2}\right)^{-(\frac{\upsilon+1}{2})}$$

$$= \frac{\Gamma(\frac{\upsilon-1}{2})}{\sigma\sqrt{\upsilon\pi}\Gamma(\upsilon/2)}\left(1+\frac{1}{\upsilon}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\upsilon+1}{2}}.$$

Exercise 5

a. We have

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\sum_{i=1}^{N}\log\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma_j}) \\ &= \sum_i \frac{\pi_k\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma_k})\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)}{p(\mathbf{x}_i|\boldsymbol{\theta})} \\ &= \sum_i r_{ik}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k).\end{aligned}$$

b. We have

$$\begin{aligned}\frac{\partial \ell}{\partial \pi_k} &= \frac{\partial}{\partial \pi_k}\sum_{i=1}^{N}\log\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma_j}) \\ &= \sum_i \ frac\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)p(\mathbf{x}_i|\theta).\end{aligned}$$

c. Using the results from part (b), we have

$$\begin{aligned}\frac{\partial \ell}{\partial w_k} &= \sum_{j=1}^{K}\frac{\partial \ell}{\partial \pi_j}\frac{\partial \pi_j}{\partial w_k} \\ &= \sum_i \ frac\pi_k p(\mathbf{x}_i,\boldsymbol{\theta})\left(-\sum_{j=1}^{K}[\pi_j\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)]+\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \\ &= \sum_i \frac{\pi_k}{p(\mathbf{x}_i,\boldsymbol{\theta})}\left(-p(\mathbf{x}_i|\boldsymbol{\theta})+\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \\ &= \sum_i r_{ik}-\pi_k\end{aligned}$$

d.

Recall that $\left.\frac{\partial f}{\partial \mathbf{A}}\right|_{\mathbf{A}}$ is a matrix such that

$$f(\mathbf{A} + \partial \mathbf{A}) \approx f(\mathbf{A}) + \operatorname{Tr}\left(\frac{\partial f}{\partial \mathbf{A}}^T \partial \mathbf{A}\right).$$

Here, the trace can be thought of as a matrix "dot product."

We can rewrite the question as

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\Sigma}_k} &= \sum_{i=1}^N \frac{\pi_k}{p(\mathbf{x}_i)} \frac{\partial}{\partial \Sigma_k} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{i=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{p(\mathbf{x}_i | \theta)} \frac{1}{\mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{i=1}^N r_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{i=1}^N r_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left[-\frac{D}{2} \log(2\ pi) - \frac{1}{2} \log \det \boldsymbol{\Sigma}_k - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right].
\end{aligned}$$

Using the fact that

$$\begin{aligned}
\partial \log \det \mathbf{A} &= \operatorname{Tr}(\mathbf{A}^{-T} \partial \mathbf{A}), \\
\partial (\mathbf{A}^{-1}) &= -\mathbf{A}^{-1} \partial \mathbf{A} \mathbf{A}^{-1},
\end{aligned}$$

and

$$\partial (\mathbf{x}^T \mathbf{A} \mathbf{x}) = \operatorname{Tr}(\mathbf{x}\mathbf{x}^T \partial \mathbf{A}),$$

we have

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log \det \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^{-1},$$

and

$$\begin{aligned}
\partial \left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right] &= \operatorname{Tr} \left[(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \partial (\boldsymbol{\Sigma}_k^{-1}) \right] \\
&= -\operatorname{Tr} \left[(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \partial \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1} \right] \\
&= -\operatorname{Tr} \left[\boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \partial \boldsymbol{\Sigma}_k \right].
\end{aligned}$$

Giving us our result:

$$\frac{\partial \ell}{\partial \boldsymbol{\Sigma}_k} = \sum_{i=1}^N r_{ik} \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} - \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right).$$

Recall that $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_k^{-1}$ are symmetric.

e. To stop notation from become clunky, let $\mathbf{a}_{ik} = \mathbf{x}_i - \boldsymbol{\mu}_k$.

Using the results from part e, and the fact that

$$\partial(\mathbf{A}^T\mathbf{A}) = \partial\mathbf{A}^T\mathbf{A} + \mathbf{A}^T\partial\mathbf{A},$$

we have

$$\begin{aligned}
\partial\left[(\mathbf{a}_{ik})^T\mathbf{\Sigma}_k^{-1}(\mathbf{a}_{ik})\right] &= \mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\partial\mathbf{\Sigma}_k) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\partial(\mathbf{R}_k^T\mathbf{R}_k)) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}(\partial\mathbf{R}_k^T\mathbf{R}_k + \mathbf{R}_k^T\partial\mathbf{R}_k)) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\partial\mathbf{R}_k^T\mathbf{R}_k + \mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\mathbf{R}_k^T\partial\mathbf{R}_k) \\
&= \mathrm{Tr}(\mathbf{R}_k^T\partial\mathbf{R}_k\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1} + \mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\mathbf{R}_k^T\partial\mathbf{R}_k) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\mathbf{R}_k^T\partial\mathbf{R}_k + \mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{\Sigma}_k^{-1}\mathbf{R}_k^T\partial\mathbf{R}_k) \\
&= 2\,\mathrm{Tr}(\mathbf{\Sigma}_k^{-1}\mathbf{a}_{ik}\mathbf{a}_{ik}^T\mathbf{R}_k^{-1}\partial\mathbf{R}_k).
\end{aligned}$$

Also,

$$\begin{aligned}
\partial\log\det\mathbf{\Sigma}_k^{-1} &= \mathrm{Tr}(\mathbf{\Sigma}_k^{-T}\partial\mathbf{\Sigma}_k^{-1}) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-T}\partial(\mathbf{R}_k^T\mathbf{R}_k)) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_k^{-T}(\partial\mathbf{R}_k^T\mathbf{R}_k + \mathbf{R}_k^T\partial\mathbf{R}_k)) \\
&= 2\,\mathrm{Tr}(\mathbf{R}_k^{-T}\partial\mathbf{R}_k).
\end{aligned}$$

Finally, the answer is

$$\frac{\partial\ell}{\partial\mathbf{\Sigma}_k} = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ik}\left(-\mathbf{R}_k^{-1} - \mathbf{R}_k^{-T}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}\right).$$

But when performing gradient descent, we should change all the values of the gradient that are below the diagonal to zero, forcing $\mathbf{R}_k$ to be upper-triangular.

Exercise 13

Recall from Chapter 4 that

$$\mathcal{N}(x_j|\theta,\sigma_j^2)\mathcal{N}(\theta|\mu,\tau^2) = \mathcal{N}\left(\theta|\frac{\sigma_j^2\theta + \tau^2\mu}{\sigma_j^2+\tau^2}, \frac{\sigma_j^2\tau^2}{\sigma_j^2+\tau^2}\right).$$

It follows that

$$
\begin{aligned}
Q(\eta^t, \eta^{t-1}) &= \sum_j \mathbb{E}\left[\log \mathcal{N}(\theta | m_{j,t}, s_{j,t}^2) | x_j, m_{j,t-1}, s_{j,t-1}^2\right] \\
&= \sum_j \mathbb{E}\left[-\frac{1}{2}\log(2\pi s_{j,t}^2) - \frac{1}{2}\left(\frac{\theta - m_{j,t}}{s_{j,t}}\right)^2 \middle| x_j, m_{j,t-1}, s_{j,t-1}^2\right] \\
&= -\frac{1}{2}\sum_j \log(2\pi s_{j,t}^2) + \frac{1}{s_{j,t}^2}\mathbb{E}\left[\theta^2 - 2\theta m_{j,t} + m_{j,t}^2 \middle| x_j, m_{j,t-1}, s_{j,t-1}^2\right] \\
&= -\frac{1}{2}\sum_j \log(2\pi s_{j,t}^2) + \frac{1}{s_{j,t}^2}\left(s_{j,t-1}^2 + m_{j,t-1}^2 - 2m_{j,t-1}m_{j,t} + m_{j,t}^2\right),
\end{aligned}
$$

where $m_{j,t} = \frac{\sigma_j^2 \mu_t + \tau_t^2 x_j}{\sigma_j^2 + \tau_t^2}$ and $s_{j,t}^2 = \frac{\sigma_j^2 \tau_t^2}{\sigma_j^2 + \tau_t^2}$.

Next, we optimize wrt to μ_t:

$$
\frac{\partial m_{j,t}}{\partial \mu_t} = \frac{\sigma_j^2}{\sigma_j^2 + \tau_t^2} = 1 - \frac{\tau_t^2}{\sigma_j^2 + \tau_t^2}
$$

and

$$
\begin{aligned}
\frac{\partial Q}{\partial \mu_t} &= -\frac{1}{2}\sum_j \frac{1}{s_{j,t}^2}\left(s_{j,t-1}^2 + m_{j,t-1}^2 - 2m_{j,t-1}\frac{\partial}{\partial \mu_t}(m_{j,t}) + \frac{\partial}{\partial \mu_t}\left(m_{j,t}^2\right)\right) \\
&= -\frac{1}{2}\sum_j \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2 \tau_t^2}\left(s_{j,t-1}^2 + m_{j,t-1}^2 - 2m_{j,t-1}\frac{\sigma_j^2}{\sigma_j^2 + \tau_t^2} + 2m_{j,t}\frac{\sigma_j^2}{\sigma_j^2 + \tau_t^2}\right) \\
&= -\frac{1}{2\tau_t^2}\sum_j \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}s_{j,t-1}^2 + \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}m_{j,t-1}^2 - 2m_{j,t-1} + 2m_{j,t}.
\end{aligned}
$$

Now we set equal to 0 and solve:

$$
\begin{aligned}
\frac{\partial Q}{\partial \mu_t} &= 0 \\
-\frac{1}{2\tau_t^2}\sum_j \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}s_{j,t-1}^2 + \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}m_{j,t-1}^2 - 2m_{j,t-1} + 2m_{j,t} &= 0 \\
\sum_j \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}s_{j,t-1}^2 + \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}m_{j,t-1}^2 - 2m_{j,t-1} + 2m_{j,t} &= 0 \\
\sum_j \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}s_{j,t-1}^2 + \frac{\sigma_j^2 + \tau_t^2}{\sigma_j^2}m_{j,t-1}^2 - 2m_{j,t-1} = -\sum_j 2m_{j,t}&.
\end{aligned}
$$

You get the idea...