Machine Learning 2 - Homework 5

Pascal M. Esser

May 7, 2018

Collaborators: Sindy Löwe

1

Consider a Gaussian mixture model

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

1.1

Given the expected value of the complete-data log-likelihood (9.40 in Bishop's book)

$$\mathbb{E}_{posterior}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n}^{N} \sum_{k}^{K} \gamma(z_{nk})\{\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{2}$$

Derive update rules for $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Write the Lagrangian as:

$$\mathcal{L} = \sum_{n}^{N} \sum_{k}^{K} \gamma(z_{nk})\{\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} + \lambda \left(\sum_{k} \pi_k - 1 \right) \tag{3}$$

Derive $\boldsymbol{\pi}_k$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n} \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0 \tag{4}$$

$$\Rightarrow \pi_k = \frac{\sum_n \gamma(z_{nk})}{-\lambda} = \frac{N_k}{-\lambda} \tag{5}$$

$$\text{with } -\lambda \cdot 1 = -\lambda \sum_{k} \pi_k = \sum_{k} \sum_{n} \gamma(z_{nk}) = \sum_{k} N_k \Rightarrow \lambda = -N \tag{6}$$

$$\Rightarrow \pi_k = \frac{N_k}{N} \tag{7}$$

Derive $\boldsymbol{\mu}_k$

$$\ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \sum_n \gamma(z_{nk}) \frac{\partial \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \tag{9}$$

$$= \sum_n \frac{1}{2}\gamma(z_{nk}) \frac{\partial (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\partial (\boldsymbol{x}_n - \boldsymbol{\mu}_k)} \frac{\partial (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k} \tag{10}$$

$$= \sum_n -\frac{1}{2}\gamma(z_{nk}) 2\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(-\mathbb{I}) \tag{11}$$

$$= \sum_n \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{12}$$

$$\boldsymbol{\Sigma}_k^{-1} \sum_n \gamma(z_{nk}) \boldsymbol{x}_n = \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \sum_n \gamma(z_{nk}) \tag{13}$$

$$\Rightarrow \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) \boldsymbol{x}_n \tag{14}$$

Derive $\boldsymbol{\Sigma}_k$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \sum_n \gamma(z_{nk}) \frac{\partial \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k} \tag{15}$$

$$= \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\frac{\partial \ln|\boldsymbol{\Sigma}_k|}{\partial \boldsymbol{\Sigma}_k} + \frac{\partial (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\partial \boldsymbol{\Sigma}_k} \right] \tag{16}$$

$$\text{using Matrix Cookbook 57 and 61} \tag{17}$$

$$= \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\boldsymbol{\Sigma}_k^{-T} - \boldsymbol{\Sigma}_k^{-\top}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-\top} \right] \tag{18}$$

$$= -\frac{1}{2}\boldsymbol{\Sigma}_k^{-\top} \sum_n \gamma(z_{nk}) - \boldsymbol{\Sigma}_k^{-\top} \sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-\top} = 0 \tag{19}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \tag{20}$$

1.2

Consider a special case of the model above, in which the covariance matrices $\boldsymbol{\Sigma}_k$ of the components are all constrained to have a common value $\boldsymbol{\Sigma}$. Derive EM equations for maximizing the likelihood function under such a model.

We don't have any changes for $\boldsymbol{\pi}_k$ and $\boldsymbol{\mu}_k$ as they do not include terms for $\boldsymbol{\Sigma}_k$.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}} = \sum_k \sum_n \gamma(z_{nk}) \frac{\partial \ln \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} \tag{21}$$

$$= \sum_k \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\frac{\partial \ln|\boldsymbol{\Sigma}_k|}{\partial \boldsymbol{\Sigma}} + \frac{\partial (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\partial \boldsymbol{\Sigma}} \right] \tag{22}$$

$$\text{using Matrix Cookbook 57 and 61} \tag{23}$$

$$= \sum_k \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\boldsymbol{\Sigma}^{-T} - \boldsymbol{\Sigma}^{-\top}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-\top} \right] \tag{24}$$

$$= \sum_k -\frac{1}{2}\boldsymbol{\Sigma}^{-\top} \sum_n \gamma(z_{nk}) - \sum_k \boldsymbol{\Sigma}^{-\top} \sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-\top} = 0 \tag{25}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_k \sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \tag{26}$$

2

Suppose we wish to use the EM algorithm to maximize the posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{X})$ for a model (Figure 1) containing latent variables z and observed variables $\boldsymbol{x}$. Show that the E step remains the same as in the maximum likelihood case, where as in the M step, the quantity to be maximized is

$$\sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}^{old}) \ln p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) \tag{27}$$

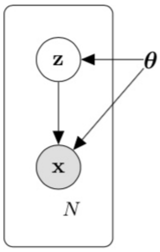

Figure 1: A simple generative model.

For adjusting the EM algorithm to maximize the posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{X})$ we first rewrite the posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{X})$ using Bishop [1] 9.76 and 9.77 as

$$\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{X}) =& \frac{p(\boldsymbol{\Theta}, \boldsymbol{X})}{p(\boldsymbol{X})} && (28)\\
\ln p(\boldsymbol{\Theta}|\boldsymbol{X}) =& \ln p(\boldsymbol{\Theta}, \boldsymbol{X}) - \ln p(\boldsymbol{X}) && (29)\\
=& \ln p(\boldsymbol{X}|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) - \ln p(\boldsymbol{X}) && (30)\\
=& \mathcal{L}(q, \boldsymbol{\Theta}) + \mathcal{D}_{KL}(q||p) + \ln p(\boldsymbol{\Theta}) - \ln p(\boldsymbol{X}) && (31)\\
\geq& \mathcal{L}(q, \boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) - \ln p(\boldsymbol{X}) && (32)\\
=& \mathcal{L}(q, \boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) - \text{const} && (33)\\
& && (34)
\end{aligned}$$

This means that for an optimization with respect to q we can reuse the same E-Step as before as q only appears in $\mathcal{L}(q, \boldsymbol{\Theta})$.

Starting from Bishop [1] 9.32 and 9.33 we can furthermore rewrite the M-Step with the prior term. The original M-Step can be written as:

$$\boldsymbol{\Theta}^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \tag{35}$$

with

$$\mathcal{Q}(\Theta, \Theta^{old}) = \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}^{old}) \ln p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\Theta}) \tag{36}$$

we can now rewrite this for the maximum likelihood case by adding $\ln p(\boldsymbol{\Theta})$ to the optimization as follows:

$$\mathcal{L}(q, \boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) - \text{const} = \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}^{old}) \ln p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) - \text{const} \tag{37}$$

3

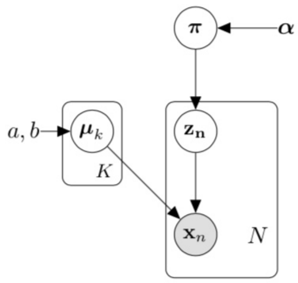

Figure 2: Mixtures of Bernoulli distribution

$$\boldsymbol{\pi}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{38}$$

$$\boldsymbol{z}_n|\boldsymbol{\pi} \sim Mult(\boldsymbol{z}_n|\boldsymbol{\pi}) \tag{39}$$

$$\boldsymbol{\mu}_k|a_k b_k \sim Beta(\boldsymbol{\mu}_k|a_k b_k) \tag{40}$$

$$\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\} \sim \prod_k^K (Bern(\boldsymbol{x}_n|\boldsymbol{\mu}_k))^{z_{nk}} \tag{41}$$

Derive the EM algorithm for maximizing the posterior probability $p(\boldsymbol{\mu}, \boldsymbol{\pi}|\{\boldsymbol{x}_n\}_{n=1}^N)$. (The E step is given in Bishop's Book, you only need to do the M step)

optimize $_{\boldsymbol{z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})] + \ln p(\boldsymbol{\Theta})$:

$$M = \sum_n^N \sum_k^K \gamma(z_{nk}) \left[\ln \pi_k + \sum_i^D x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right] + \ln p(\boldsymbol{\Theta}) \tag{42}$$

$$\text{with} \tag{43}$$

$$\ln p(\boldsymbol{\Theta}) = \sum_k \ln Beta(\boldsymbol{\mu}_k|a_k, b_k) + \ln Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{44}$$

$$= \sum_k \sum_i (a_k - 1) \ln \mu_{ki} + (b_k - 1) \ln(1 - \mu_{ki}) + f(a_k, b_k) + \sum_k (a_k - 1) \ln \pi_k + g(\alpha_k) \tag{45}$$

$$\frac{\partial M}{\partial \mu_{ki}} = \sum_n \gamma(z_{nk}) \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) + \frac{a_k - 1}{\mu_{ki}} - \frac{b_k - 1}{1 - \mu_{ki}} = 0 \tag{46}$$

$$\frac{1}{\mu_{ki}} \left(\sum_n \gamma(z_{nk}) x_{ni} + a_k - 1 \right) = \frac{1}{1 - \mu_{ki}} \left(\sum_n \gamma(z_{nk})(1 - x_{ni}) + b_k - 1 \right) \tag{47}$$

$$(1 - \mu_{ki}) \left(\sum_n \gamma(z_{nk}) x_{ni} + a_k - 1 \right) = \mu_{ki} \left(\sum_n \gamma(z_{nk})(1 - x_{ni}) + b_k - 1 \right) \tag{48}$$

$$\sum_n \gamma(z_{nk}) x_{ni} + a_k - 1 - \mu_{ki} \sum_n \gamma(z_{nk}) x_{ni} - \mu_{ki}(a_k - 1) = -\mu_{ki} \sum_n \gamma(z_{nk}) x_{ni} + -\mu_{ki} \sum_n \gamma(z_{nk}) + \mu_{ki}(b_k - 1) \tag{49}$$

$$\sum_n \gamma(z_{nk}) x_{ni} + a_k - 1 = \mu_{ki}(a_k - 1 + N_k + b_k - 1) \tag{50}$$

$$\Rightarrow \mu_{ki} = \frac{\sum_n \gamma(z_{nk}) x_{ni} + a_k - 1}{a_k + N_k + b_k - 2} \tag{51}$$

alternatively we can write μ in vector form by keeping the sum over i to get $\boldsymbol{\mu}_k$ as follows (the derivation says the same with the exact of the use of the vector instead of the double index that is derived above):

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(z_{nk})\boldsymbol{x}_n + a_k - 1}{a_k + N_k + b_k - 2} \tag{52}$$

$$\frac{\partial M + \lambda\left(\sum_k \pi_k - 1\right)}{\partial \pi_k} = \sum_n \frac{\gamma(z_{nk})}{\pi_i} + \frac{\alpha_k - 1}{\pi_k} + \lambda = 0 \tag{53}$$

$$\Rightarrow \pi_k = \frac{\sum_n \gamma(z_{nk}) + \alpha_k - 1}{-\lambda} \tag{54}$$

$$= \frac{N_k + \alpha_k - 1}{-\lambda} \tag{55}$$

$$\text{with } -\lambda \sum_k \pi_k = \sum_k \sum_n \gamma(z_{nk}) + \alpha_k - K \Rightarrow -\lambda = N + \sum_k \alpha_k - K \tag{56}$$

$$\Rightarrow \pi_k = \frac{N_k + \alpha_k - 1}{N + \sum_k \alpha_k - K} \tag{57}$$

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.