ML estimation and Bayesian estimation

Luu Huu Phuc
Pattern Recognition Spring class 2020
Kyoto University

I. Question

ML estimation

- Derive the update formulas of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$ in page 22 by letting the partial derivative of the lower bound in page 20 w.r.t each parameter equal to zero.

Bayesian estimation

- Derive the variational posteriors of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$ in page 47 by using the formulas in page 46.

Test different values for K and discuss appropriate value of K.

II. Answer

A. Update formulas of the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$

For one data point $\mathbf{x_n}$, the inequality in page 20 can be rewritten as

$$\begin{aligned}
\log p(\mathbf{x_n};\theta) &\geq \sum_{k=1}^{K} q(z_{n_k}=1)\log p(\mathbf{x_n}, z_{n_k}=1;\theta) \\
&= \sum_{k=1}^{K} q(z_{n_k}=1)\{\log N(\mathbf{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Lambda}_k^{-1}) + \log \pi_k]\} \\
&= \sum_{k=1}^{K} \gamma_{n_k}\{\frac{1}{2}\log(\det \boldsymbol{\Lambda}_k) - \frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x_n}-\boldsymbol{\mu}_k) + \log \pi_k\} + C
\end{aligned}$$

,where C denotes a constant value that is independent to the parameters $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$. For all data points $\{\mathbf{x_1}, ..., \mathbf{x_N}\}$, we have:

$$\sum_{n=1}^{N}\log p(\mathbf{x_n};\theta) \geq \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{n_k}\{\frac{1}{2}\log(\det \boldsymbol{\Lambda}_k) - \frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x_n}-\boldsymbol{\mu}_k) + \log \pi_k\} + C$$

Therefore, the lower bound can be written as follows.

$$LB = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{n_k}\{\frac{1}{2}\log(\det \boldsymbol{\Lambda}_k) - \frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\mathbf{x_n}-\boldsymbol{\mu}_k) + \log \pi_k\} + \text{constant}$$

We update the parameters by finding θ to maximize the lower bound.

$$\arg\max_{\theta} \quad LB \qquad \text{s.t} \quad \sum_{k=1}^{K} \pi_k = 1$$

or

$$\arg\max_{\theta,\lambda} \quad L$$

, where $L = LB + \lambda(1 - \sum_{k=1}^{K} \pi_k)$ and λ is the Lagrange multiplier's parameter. Taking the derivatives of L w.r.t $\boldsymbol{\pi}$ and λ we have:

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 \tag{1}$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma_{n_k}}{\pi_k} - \lambda \tag{2}$$

Setting (1) and (2) to 0, we have

$$\lambda = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{n_k} = S.[1]$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{n_k}}{\lambda} = \frac{S_k[1]}{S.[1]}$$

Taking the derivatives of L w.r.t $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, we have:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\mu}_k} &= -\sum_{n=1}^{N} \gamma_{n_k} \boldsymbol{\Lambda}_k^T (\mathbf{x_n} - \boldsymbol{\mu}_k) \boldsymbol{\Lambda}_k \\ &= -\boldsymbol{\Lambda}_k^T \sum_{n=1}^{N} \gamma_{n_k} (\mathbf{x_n} - \boldsymbol{\mu}_k) \boldsymbol{\Lambda}_k \end{aligned} \tag{3}$$

$$\frac{\partial L}{\partial \boldsymbol{\Lambda}_k} = \frac{1}{2} \sum_{n=1}^{N} \gamma_{n_k} \{ \boldsymbol{\Lambda}_k^{-1} - (\mathbf{x_n} - \boldsymbol{\mu}_k)(\mathbf{x_n} - \boldsymbol{\mu}_k)^T \} \tag{4}$$

Setting (3) to 0, we have

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{n_k} \mathbf{x_n}}{\sum_{n=1}^{N} \gamma_{n_k}} = \frac{S_k[\mathbf{x}]}{S_k[1]}$$

Setting (4) to 0 and using $S_k[\mathbf{x}] = S_k[1]\boldsymbol{\mu}_k$, we have

$$
\begin{aligned}
\mathbf{\Lambda}_k^{-1} &= \frac{1}{S_k[1]}\{\sum_{n=1}^{N}\gamma_{n_k}\mathbf{x_n}\mathbf{x_n}^T - \boldsymbol{\mu}_k S_k[\mathbf{x}]^T - S_k[\mathbf{x}]\boldsymbol{\mu}_k^T + S_k[1]\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\} \\
&= \frac{S_k[\mathbf{xx}]}{S_k[1]} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T
\end{aligned}
$$

We have finished deriving the update formulas of the parameters in page 22.

B. Variational posteriors of the parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$, and $\mathbf{\Lambda}$

B.1 Variational posterior for $\boldsymbol{\pi}$

From page 46, we have

$$
\log q^*(\boldsymbol{\pi}) = \log p(\boldsymbol{\pi}) + \mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Z}|\boldsymbol{\pi})] + \text{const} \tag{5}
$$

The first term of (5) can be expanded as

$$
\log p(\boldsymbol{\pi}) = \log(\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}) + \log\frac{\Gamma(\sum_{k=1}^{K}\alpha_{0k})}{\prod_{k=1}^{K}\Gamma(\alpha_{0k})}
$$

The second term of (5) can be expanded as

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{Z}|\boldsymbol{\pi})] &= \int q(\mathbf{Z})\log p(\mathbf{Z}|\boldsymbol{\pi})d\mathbf{Z} \\
&= \sum_{n=1}^{N}\int q(\mathbf{z}_n)\log p(\mathbf{z}_n|\boldsymbol{\pi})d\mathbf{z}_n \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N}\int q(\mathbf{z}_n)z_{n_k}\log\pi_k d\mathbf{z}_n \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N}\log\pi_k\int q(\mathbf{z}_n)z_{n_k}d\mathbf{z}_n \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_{n_k}\log\pi_k \quad \text{derived from (6)} \\
&= \sum_{k=1}^{K}\log\pi_k^{\sum_{n=1}^{N}\gamma_{n_k}} \\
&= \log\prod_{k=1}^{K}\pi_k^{S_k[1]} \quad \text{where} \quad S_k[1] = \sum_{n=1}^{N}\gamma_{n_k}
\end{aligned}
$$

Note that, z_{n_k} only takes binary values and $q(\mathbf{z}_n) = \prod_{k=1}^{K}\gamma_{n_k}^{z_{n_k}}$, therefore

$$\int q(\mathbf{z}_n) z_{n_k} d\mathbf{z}_n = \gamma_{n_k} \tag{6}$$

From the above equations, we have

$$\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \log \prod_{k=1}^{K} \pi_k^{\alpha_{0k}+S_k[1]-1} + \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_{0k})}{\prod_{k=1}^{K} \Gamma(\alpha_{0k})} + \text{const} \\
&= \log \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} + \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_{0k})}{\prod_{k=1}^{K} \Gamma(\alpha_{0k})} + \text{const} \quad \text{where} \quad \alpha_k = \alpha_{0k} + S_k[1]
\end{aligned}$$

By normalizing the above equation, we derive the variational posterior on parameter $\boldsymbol{\pi}$ as in page 47.

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

B.2 Variational posterior for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$

From page 46, we have

$$\log q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \mathbb{E}_{q(\mathbf{Z})}[p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const} \tag{7}$$

The first term of (7) can be expanded as follows.

$$\begin{aligned}
\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \sum_{k=1}^{K} \{\log N(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\boldsymbol{\beta}_0 \boldsymbol{\Lambda}_k)^{-1}\} + \log W(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, v_0)) \\
&= \sum_{k=1}^{K} \{\frac{1}{2} \log \det(\boldsymbol{\Lambda}_k) - \frac{\beta_0}{2} (\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0)\} \\
&+ \sum_{k=1}^{K} \{\frac{v_0 - d - 1}{2} \log \det(\boldsymbol{\Lambda}_k) - \frac{1}{2} \text{Trace}(\boldsymbol{W}_0^{-1} \boldsymbol{\Lambda}_k)\} + \text{const}
\end{aligned}$$

Here, d is in the size $d \times d$ of matrix $\boldsymbol{W}_0$. The second term of (7) can be expanded as follows.

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{Z})}[\log p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) d\mathbf{Z} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \int q(\mathbf{z}_n) z_{n_k} \log N(\mathbf{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Lambda}_k^{-1}) d\mathbf{z}_n \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \log N(\mathbf{x_n}|\boldsymbol{\mu_k},\boldsymbol{\Lambda}_k^{-1}) \int q(\mathbf{z}_n) z_{n_k} d\mathbf{z}_n \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{n_k} \log N(\mathbf{x_n}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}) \quad \text{derived from (6)} \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} \gamma_{n_k} \{ \frac{1}{2} \log\det(\boldsymbol{\Lambda}_k) - \frac{1}{2}(\boldsymbol{\mu}_k - \mathbf{x_n})^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{x_n}) \}
\end{aligned}
$$

Since $\log q^*(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \log q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda}) + \log q^*(\boldsymbol{\Lambda})$, we first find $\log q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda})$ by only considering the terms on the right hand side of (7) which depends on $\boldsymbol{\mu}$.
From (7), we have

$$
\begin{aligned}
\log q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda}) = -\frac{1}{2} \sum_{k=1}^{K} \{ &\beta_0 (\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0) \\
&+ \sum_{n=1}^{N} \gamma_{n_k} (\boldsymbol{\mu}_k - \mathbf{x_n})^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{x_n}) \} + \text{const}
\end{aligned}
$$

Expand the above equation properly and using $\beta_k = \beta_0 + S_k[1]$, $S_k[\mathbf{x}] = \sum_{n=1}^{N} \gamma_{n_k} \mathbf{x}_n$, and $\mathbf{m}_k = \frac{\beta_0 m_k + S_k[\mathbf{x}]}{\beta_k}$ we have

$$
\log q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda}) = -\frac{1}{2} \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T \beta_k \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_k) + \text{const}
$$

By normalizing the above equation, we have

$$
q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k | \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \tag{8}
$$

Next, we find $\log q^*(\boldsymbol{\Lambda})$ from $\log q^*(\boldsymbol{\Lambda}) = \log q^*(\boldsymbol{\mu},\boldsymbol{\Lambda}) - \log q^*(\boldsymbol{\mu}|\boldsymbol{\Lambda})$.
We have

$$
\begin{aligned}
\log q^*(\mathbf{\Lambda}) &= \log q^*(\boldsymbol{\mu}, \mathbf{\Lambda}) - \log q^*(\boldsymbol{\mu}|\mathbf{\Lambda}) \\
&= \sum_{k=1}^{K} \{ \frac{1}{2} \log \det(\mathbf{\Lambda}_k) - \frac{\beta_0}{2} (\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \mathbf{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0) \\
&+ \frac{v_0 - d - 1}{2} \log \det(\mathbf{\Lambda}_k) - \frac{1}{2} \mathrm{Trace}(\boldsymbol{W}_0^{-1} \mathbf{\Lambda}_k) \\
&+ \sum_{n=1}^{N} \gamma_{n_k} [\frac{1}{2} \log \det(\mathbf{\Lambda}_k) - \frac{1}{2} (\boldsymbol{\mu}_k - \mathbf{x_n})^T \mathbf{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{x_n})] \\
&+ \frac{1}{2} (\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T \beta_k \mathbf{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_k) \} \\
&+ \mathrm{const}
\end{aligned}
$$

Expand the above equation properly and using

$$
\begin{aligned}
v_k &= v_0 + S_k[1] \\
S_k[\mathbf{x}\mathbf{x}^T] &= \sum_{n=1}^{N} \gamma_{n_k} \mathbf{x}_n \mathbf{x}_n^T \\
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + \beta_0 \mathbf{m}_0 \mathbf{m}_0^T + S_k[\mathbf{x}\mathbf{x}^T] - \beta_k \mathbf{m}_k \mathbf{m}_k^T
\end{aligned}
$$

,we have

$$
\log q^*(\mathbf{\Lambda}) = \sum_{k=1}^{K} \{ \frac{v_k - d - 1}{2} \log \det(\mathbf{\Lambda}_k) - \frac{1}{2} \mathrm{Trace}(\mathbf{\Lambda}_k \boldsymbol{W}_k^{-1}) \} + \mathrm{const}
$$

Normalizing the above equation gives us the formula for $q^*(\mathbf{\Lambda})$ as follows.

$$
q^*(\mathbf{\Lambda}) = \prod_{k=1}^{K} W(\mathbf{\Lambda}_k | \boldsymbol{W}_k, v_k) \tag{9}
$$

From equations (8) and (9), we derive the variational posterior of parameters $\boldsymbol{\mu}$ and $\mathbf{\Lambda}$ as in page 47.

$$
q^*(\boldsymbol{\mu}, \mathbf{\Lambda}) = \prod_{k=1}^{K} N(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \mathbf{\Lambda}_k)^{-1}) W(\mathbf{\Lambda}_k | \mathbf{W}_k, v_k)
$$

III. Appropriate value for K

To find an appropriate value for the numbers of clusters K, we tried with different values of $K \in \{2, 3, 4, 5, 6\}$. As shown in fig 1, the log-likelihood decreases when K increases from 2 to 4 and remains the same for $K \geq 4$. This is a good suggestion that $K = 4$ would be a appropriate value for the number of clusters.

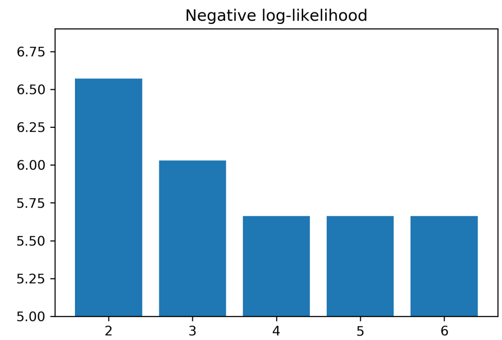

Figure 1: The negative log-likelihood of different values for K

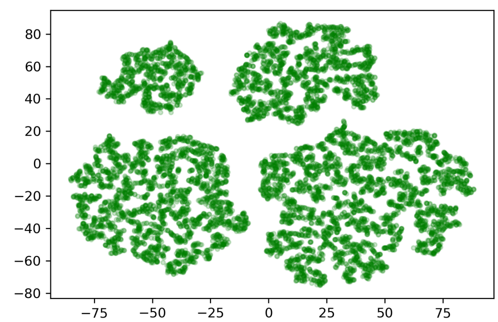

Figure 2: Plotting the data in 2-D using TSNE

Furthermore, when we look at the weights $\boldsymbol{\pi}$ of different K, we can see that when $K \geq 4$, there only exists 4 clusters out of K clusters with major weights and all the other $K-4$ clusters have very small weights. This observation once again suggest that $K = 4$ would be a proper value[1].

Additionally, as shown in fig 2, when using TSNE to embed the data into 2-D space for visualization, it is clear that the data are laid in 4 different clusters.

Figures 3 and 4 shows the classification result of the data into 4 classes when we fit a Gaussian Mixture Model with $K = 4$ (a data point $\mathbf{x}_n$ is classified according to its posterior probability $\mathbf{z}_n$).

[1] The discussion here is based on the implementation of EM algorithm. The same arguments can be applied for the case of VB algorithm.

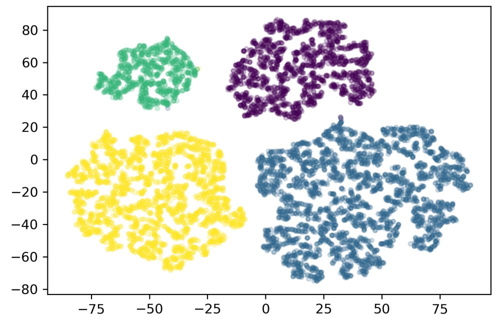

Figure 3: Classification of the data when fitting a GMM with K = 4 using the EM algorithm

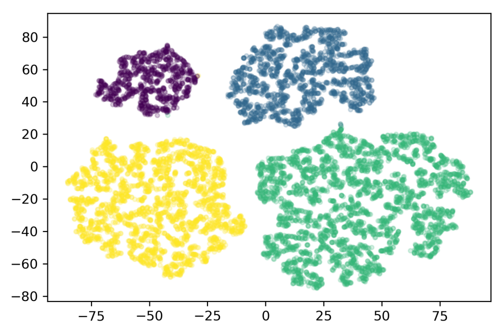

Figure 4: Classification of the data when fitting a GMM with K = 4 using the VB algorithm