STA243 COMPUTATIONAL STATISTICS HOMEWORK3

YUXIANG LIN(ID:917849914) E-MAIL: YUXLIN@UCDAVIS.EDU
ZHIWEI WANG(ID:917869318) E-MAIL: WIZWANG@UCDAVIS.EDU

DATE: MAY 11, 2020

Problem 1. *MNIST handwritten digit dataset classification.*

Solution. *Result in the code.R*

Problem 2.

(a) *Write down the marginal distribution of the Z_i (Hint: What is the probability $p(Z_i = j)$).*

(b) *Calculate $p(Z_i = j|x_i)$. (Hint: Bayes Rule)*

(c) *Prove the following lower bound of the log-likelihood function:*

$$\ell(\theta) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \mathbb{F}_{ij} \frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right] \geq \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{F}_{ij} \log \left[\frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right]$$

Hint: you can use the Jensen's inequality $\log \mathbb{E}X \geq \mathbb{E} \log X$ (You are not required to prove the Jensen's inequality).

(d) *Prove that $\ell(\theta') = Q(\mathbb{F}, \theta')$ when: $\mathbb{F}_{ij} = p_{\theta'}(Z_i = j|x_i)$*

(e) *(M-step for mixture of spherical Gaussians) Under the* **mixture of spherical Gaussians** *model, derive the M-step updating equations for $\mu_j^{(t+1)}, \Sigma_j^{(t+1)}$ and $\pi_j^{(t+1)}$.*

(f) *(M-step for mixture of diagonal Gaussians) Under the* **mixture of diagonal Gaussians** *model, derive the M-step updating equations for $\mu_j^{(t+1)}, \Sigma_j^{(t+1)}$ and $\pi_j^{(t+1)}$.*

Solution.

(a) *The marginal distribution of the Z_i is:*

$$p(Z_i = j) = \pi_j, \ j = 1, 2, \cdots, k$$

(b) *By applying Bayes rule, for $j = 1, 2, \cdots, k$:*

$$\begin{aligned} p(Z_i = j|x_i) &= \frac{p(x_i|Z_i = j)p(Z_i = j)}{p(x_i)} \\ &= \frac{p(x_i|Z_i = j)p(Z_i = j)}{\sum_{j=1}^{k} p(x_i|Z_i = j)p(Z_i = j)} \\ &= \frac{\pi_j p(x_i|Z_i = j)}{\sum_{j=1}^{k} \pi_j p(x_i|Z_i = j)} \end{aligned}$$

Where:

$$p(x_i|Z_i = j) = (2\pi)^{-d/2} |\Sigma_j|^{-1/2} \exp\left(-\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right)$$

(c) *For the third step of the prove, we apply Jensen's inequality $log\mathbb{E}X \geq \mathbb{E}logX$:*

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \mathbb{F}_{ij} \frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right] \\
&= \sum_{i=1}^{n} \log \left[\mathbb{E}_{q_i} \left(\frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right) \right] (Jensen's\ inequality) \\
&\geq \sum_{i=1}^{n} \mathbb{E}_{q_i} \left[\log \left(\frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right) \right] \\
&\geq \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{F}_{ij} \log \left[\frac{p_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right]
\end{aligned}$$

(d) *Given the fact that $f(x) = -logx$ is convex, the equality holds for Jensen's inequality if and only if X is degenerate.*

In this case,

$$X = \frac{p_{\theta'}(x_i, Z_i = j)}{\mathbb{F}_{ij}} = \frac{p_{\theta'}(x_i, Z_i = j)}{p_{\theta'}(Z_i = j | x_i)} = p_{\theta'}(x_i)$$

In addition,

$$\mathbb{E}_{q_i} \left(\frac{p_{\theta'}(x_i, Z_i = j)}{\mathbb{F}_{ij}} \right) = \sum_{j=1}^{k} \mathbb{F}_{ij} \frac{p_{\theta'}(x_i, Z_i = j)}{\mathbb{F}_{ij}} = \sum_{j=1}^{k} p_{\theta'}\left(x_i, Z_i = j\right) = p_{\theta'}(x_i)$$

For random variable X, we have $\mathbb{P}(X = \mathbb{E}X) = 1$, which means X is degenerate. In consequence, the equality holds.

(e) *The lower bound function at $\theta^{(t)}$ is:*

$$Q(\theta^{(t)}, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{F}_{ij}^{(t)} \log \left[\frac{\mathbb{P}_\theta(x_i, Z_i = j)}{\mathbb{F}_{ij}^{(t)}} \right]$$

The problem is equivalent to updating θ such that we maximize the new lower bound at $\theta^{(t)}$:

$$\tilde{Q}(\theta^{(t)}, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{F}_{ij}^{(t)} \log \mathbb{P}_\theta(x_i, Z_i = j)$$

Under Gaussians basis, we have:

$$\log \mathbb{P}_\theta(x_i, Z_i = j) = \log \pi_j - \frac{d}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_j|) - \frac{1}{2} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j)$$

First we derive the updating equations for π_j.

For $j = 1, 2, \cdots k-1$:

$$\frac{\partial \tilde{Q}(\theta^{(t)}, \theta)}{\partial \pi_j} = \frac{\sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)}}{\pi_j} - \frac{\sum_{i=1}^{n} \mathbb{F}_{ik}^{(t)}}{\pi_k}$$

Set the partial derivative to 0, we get:

$$\frac{\pi_j}{\pi_k} = \frac{\sum_{i=1}^n \mathbb{F}_{ij}^{(t)}}{\sum_{i=1}^n \mathbb{F}_{ik}^{(t)}}$$

Sum over index j, we get:

$$\frac{\sum_{j=1}^k \pi_j}{\pi_k} = \frac{\sum_{j=1}^k \sum_{i=1}^n \mathbb{F}_{ij}^{(t)}}{\sum_{i=1}^n \mathbb{F}_{ik}^{(t)}} \Rightarrow \pi_k = \frac{\sum_{i=1}^n \mathbb{F}_{ik}^{(t)}}{n} \Rightarrow \pi_j = \frac{\sum_{i=1}^n \mathbb{F}_{ij}^{(t)}}{n}$$

It is straightforward to verify the second derivative is non-positive, so the updating equations for π_j is as follows:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{F}_{ij}^{(t)}}{n}$$

Second we derive the updating equations for μ_j.

$$\frac{\partial \tilde{Q}(\theta^{(t)}, \theta)}{\partial \mu_j} = \Sigma_j^{-1} \sum_{i=1}^n \mathbb{F}_{ij}^{(t)} (x_i - \mu_j)^T$$

Now we verify the Hessian Matrix is semi-negative defined.

$$\frac{\partial^2 \tilde{Q}(\theta^{(t)}, \theta)}{\partial^2 \mu_j} = \begin{pmatrix} -\sum_{i=1}^n \mathbb{F}_{i1}^{(t)} \Sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & -\sum_{i=1}^n \mathbb{F}_{i2}^{(t)} \Sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\sum_{i=1}^n \mathbb{F}_{ik}^{(t)} \Sigma_k^{-1} \end{pmatrix} \preceq 0$$

Set the partial derivative to 0, we get the updating equatinos for μ_j as follows:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{F}_{ij}^{(t)} x_i}{\sum_{i=1}^n \mathbb{F}_{ij}^{(t)}}$$

Third we derive the updating equations for Σ_j under spherical Gaussians assumption.
Assume that $\Sigma_j = \sigma_j^2 I_d$, then we have:

$$\frac{\partial \tilde{Q}(\theta^{(t)}, \theta)}{\partial \sigma_j^2} = \sum_{i=1}^n \mathbb{F}_{ij}^{(t)} \left(-\frac{d}{2\sigma_j^2} + \frac{1}{2\sigma_j^4} (x_i - \mu_j)^T (x_i - \mu_j) \right)$$

Now we verify the Hessian Matrix is semi-negative defined.

$$\frac{\partial^2 \tilde{Q}(\theta^{(t)}, \theta)}{\partial^2 \sigma_j^2} = \begin{pmatrix} -d\sum_{i=1}^n \mathbb{F}_{i1}^{(t)} \sigma_1^{-4} & 0 & \cdots & 0 \\ 0 & -d\sum_{i=1}^n \mathbb{F}_{i2}^{(t)} \sigma_2^{-4} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -d\sum_{i=1}^n \mathbb{F}_{ik}^{(t)} \sigma_k^{-4} \end{pmatrix} \preceq 0$$

Set the partial derivative to 0, we get the updating equatinos for σ_j^2 as follows:

$$(\sigma_j^2)^{(t+1)} = \frac{\sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)} (x_i - \mu_j^{(t)})^T (x_i - \mu_j^{(t)})}{d \sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)}}$$

Last we derive the updating equations for Σ_j under diagonal Gaussians assumption.
Assume that $\Sigma_j = diag(\sigma_{j1}^2, \sigma_{j2}^2, \cdots, \sigma_{jd})$, for $s = 1, 2, \cdots, d$ we have:

$$\frac{\partial \tilde{Q}(\theta^{(t)}, \theta)}{\partial \sigma_{js}^2} = \sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)} \left(-\frac{1}{2\sigma_{js}^2} + \frac{1}{2\sigma_{js}^4} (x_{is} - \mu_{js})^2 \right)$$

Now we verify the Hessian Matrix is semi-negative defined.

$$\frac{\partial^2 \tilde{Q}(\theta^{(t)}, \theta)}{\partial^2 \sigma_{js}^2} = \begin{pmatrix} -\sum_{i=1}^{n} \mathbb{F}_{i1}^{(t)} \Sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & -\sum_{i=1}^{n} \mathbb{F}_{i2}^{(t)} \Sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\sum_{i=1}^{n} \mathbb{F}_{ik}^{(t)} \Sigma_k^{-2} \end{pmatrix} \preceq 0$$

For $s = 1, 2, \cdots, d$, set the partial derivative to 0, we get the updating equatinos for σ_{js}^2 as follows:

$$(\sigma_{js}^2)^{(t+1)} = \frac{\sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)} (x_{is} - \mu_{js}^{(t)})^2}{\sum_{i=1}^{n} \mathbb{F}_{ij}^{(t)}}$$

Problem 3. *Implement the EM algorithm from last question to cluster the MNIST data set.*

(a) *Program the EM algorithm you derived for mixture of spherical Gaussians. Assume 5 clusters. Terminate the algorithm when the fractional change of the log-likelihood goes under 0.0001. (Try 3 random initializations and present the best one in terms of maximizing the likelihood function).*

(b) *Program the EM algorithm you derived for mixture of diagonal Gaussians. Assume 5 clusters. Terminate the algorithm when the fractional change in the log-likelihood goes under 0.0001. (Try 3 random initializations and present the best one in terms of maximizing the likelihood function).*

Apply following hints to solve three problems:

- *Use the log-sum-exp trick to avoid underflow on the computer. You will run into this problem when computing the log-likelihood. That is, when you calculate $\log \sum_j \exp^{a_j}$ for some sequence of variables a_j, calculate instead $A + \log \sum_j \exp^{a_j - A}$ where $A = \max_j a_j$.*

- *Some pixels in the images do not change throughout the entire dataset. (For example, the top-left pixel of each image is always 0, pure white.) To solve this, after updating the covariance matrix Σ_j for the mixture of diagonal Gaussians, add $0.05I_d$ to Σ_j (ie: add 0.05 to all the diagonal elements).*

- *Be mindful of how you initialize Σ_j. Note that for a diagonal matrix Σ_j, the determinant $|\Sigma_j|$ is the product of all the diagonal elements. Setting each diagonal element to a number too big at initialization will result in overflow on the computer.*

Solution.

A summary of the training and testing process as well as the logic behind:

1. Set up three different random initialized parameters mu,sigma.pi through setting different seeds.
2. For each mixture Gaussian model, pick up the best initialization according to the log-likelihood value.
3. Store the parameters derived from 1000 iterations for each mixture Gaussian model for further prediction.
4. Draw the overlapping matrix between the iterated clusters (1,2,3,4,5) and the true labels 0,1,2,3,4. Find out the best mapping pattern from the clusters to the labels according to the maximum principle.
5. Use the training model's parameters we stored to derive the F matrix for the test image data and then we get the clusters for each test sample.
6. According to the mapping we have derived, we conclude the predicted label for each test sample based on its assigned cluster. Then we use the predicted label and the true label to calculate the classification error rate.

(1) Spherical Models:

ModelID	Randomness	Log-likelihood Value
Spher1	seed(5)	1404204
Spher2	seed(50)	1404446
Spher3	seed(500)	1514350

Table 1: Three Initializations of Spherical Models

According to the Logliklihood values, model Spher3 has the best performance. We utilized its iterated parameters to make predictions:

Labels	True0	True1	True2	True3	True4
Predict0	873	0	35	5	3
Predict1	0	593	0	2	2
Predict2	0	447	4	0	1
Predict3	100	90	906	987	9
Predict4	7	5	87	16	967

Table 2: Confusion Matrix on the Test Data based on Spherical Models

According to the confusion matrix, the Error Rate based on Spherical Models is 0.3337.

(2) Diagonal Models:

ModelID	Randomness	Log-likelihood Value
Diag1	seed(5)	1318299
Diag2	seed(50)	1318309
Diag3	seed(500)	1318312

Table 3: Three Initializations of Diagonal Models

According to the Logliklihood values, model Diag3 has the best performance. We utilized its iterated parameters to make predictions:

Labels	True0	True1	True2	True3	True4
Predict0	892	0	11	3	2
Predict1	26	1129	340	294	128
Predict2	16	3	596	44	8
Predict3	39	3	47	659	0
Predict4	7	0	38	10	844

Table 4: Confusion Matrix on the Test Data based on Diagonal Models

According to the confusion matrix, the Error Rate based on Spherical Models is 0.1983.

Conclusions:

According to the error rate of selected models, the classification performance of both mixture Gaussian spherical and mixture Gaussian diagonal models overwhelms a fair coin (0.5 error rate). However, the prediction accuracy is not very good, especially the spherical one (0.67). So we conclude that the mixture Gaussian model is not a bad choice but might not be an optimal one.

Meanwhile, when we look at the difference between the spherical model and the diagonal model, we can see that the performance of the spherical model is obviously worse than the diagonal model. The intuition behind is quite straightforward: The assumption of spherical model is too strong: the variances of each feature are all the same. This assumption is extremely out of ration especially when the dimension of the feature is large (e.g. 196 in our case).

Pledge

Please sign below (print full name) after checking (✓) the following. If you can not honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

✓We pledge that we are honest students with academic integrity and we have not cheated on this homework.

✓These answers are our own work.

✓We did not give any other students assistance on this homework.

✓We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.

✓We understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

Team Member 1
Yuxiang Lin

Team Member 2
Zhiwei Wang