CSCI 5521: Introduction to Machine Learning (Spring 2020)[1]

Homework 3 Solution

Questions

1. (**30 points**) Derive the EM algorithm for estimating a mixture of Laplacian distributions (*double exponential distributions*). The probability density function of a Laplacian distribution for class C_i is

$$p(x|C_i) = f(x|\mu_i, \sigma_i) = \frac{1}{2b_i}\exp\Big(-\frac{|x-\mu_i|}{\sigma_i}\Big),\ \sigma_i > 0,$$

where μ_i and b_i are referred to as the location and diversity parameters of the distribution. The mixture density of K Laplacian distributions is

$$P(x) = \sum_{i=1}^{K} P(x|C_i)P(C_i) = \sum_{i=1}^{K} \pi_i \frac{1}{2\sigma_i}\exp\Big(-\frac{|x-\mu_i|}{\sigma_i}\Big),$$

where $\sum_{i=1}^{K} \pi_i = 1$. Given the data $\mathcal{X} = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$, define the log-likelihood function and the complete log-likelihood function, and derive the EM equations including the expectation for the responsibility $\gamma(z_i^t)$ (z_i^t is the binary indicator of sample t in cluster i) and maximum likelihood learning for $\{\pi_i, \mu_i, \sigma_i\}_{i=1,\ldots,K}$.

Important hint:

   (a) There is no easy way to solve the maximum likelihood learning of μ_i. An approximation is to binarize $\gamma(z_i^t)$ as $b_i^t = 1$ if $i = argmax_j \gamma(z_j^t)$ and otherwise $b_i^t = 0$ (the same as we do in k-means) before estimating μ_i. You should still use $\gamma(z_i^t)$ to estimate π_i and σ_i.

   (b) The absolute error $\sum_{t=1}^{N} |\theta - x^t|$ is minimized if θ is the median of the N numbers).

[1]Instructor: Rui Kuang (kuang@cs.umn.edu). TAs: Tianci Song (song0309@umn.edu) and Ruyuan (wanxx199@umn.edu).

Answer:

Based on the description, it is easy to have the log-likelihood fucntion:

$$L(\mu_i, \sigma_i, \pi_i) = \sum_{t=1}^{N} \log \sum_{i=1}^{K} \frac{\pi_i}{2\sigma_i} e^{-\frac{|x-\mu_i|}{\sigma_i}}$$

Then by introducing the variable z_i^t we have the following complete log-likelihood function (to make it consistent with the textbook, here we use ϕ to indicate the parameter vector which includes π. μ, σ):

$$\log L(\phi|X, Z) = \sum_{t=1}^{N} \sum_{i=1}^{K} z_i^t \left[\log(\pi) - log(2\sigma) - \frac{|x-\mu_i|}{\sigma_i}\right]$$

In the E-Step:

$$Q(\phi|\phi^l) = \sum_{t=1}^{N} \sum_{i=1}^{K} E\left[z_i^t|X, \phi^l\right] \left[\log(\pi) - log(2\sigma) - \frac{|x-\mu_i|}{\sigma_i}\right]$$

where $\gamma(z_i^t) = E\left[z_i^t|X, \phi^l\right] = \frac{\pi_i/2\sigma_i exp(-\frac{|x-\mu_i|}{2\sigma_i})}{\sum_{i=1}^{K} \pi_i/2\sigma_i exp(-\frac{|x-\mu_i|}{2\sigma_i})}$

In the M-Step:

Once $\gamma(z_i^t)$ we get in the E-step, we introduce Lagrange multiplier λ to enforce the constriant $\sum_{i=1}^{K} \pi_i = 1$, and the derivative of loss function w.r.t. π_i becomes:

$$\frac{\partial \sum_{t=1}^{N} \sum_{i=1}^{K} \gamma(z_i^t)\left[\log(\pi) - \log(2\sigma) - \frac{|x-\mu_i|}{\sigma_i}\right] + \lambda(\sum_{i=1}^{K} \pi_i - 1)}{\partial \pi} = \frac{\sum_{t=1}^{N} \gamma(z_i^t)}{\pi_i} - \lambda$$

Then we set it to be zero, and obtain $\pi_i = \frac{1}{\lambda} \sum_{t=1}^{N} \gamma(z_i^t)$, then combine $\sum_{i=1}^{K} \pi_i = 1$, then easily know that $\lambda = \sum_{t=1}^{N} \sum_{i=1}^{K} \gamma(z_i^t) = N_i$, thus $\pi_i = \frac{\sum_{t=1}^{N} \gamma(z_i^t)}{N_i}$

Similarly, we can have the derivative of loss function w.r.t. μ_i and σ_i and set them to be zeros separately:

$$\frac{\partial \sum_{t=1}^{N}\sum_{i=1}^{K}\gamma(z_i^t)\left[\log(\pi)-\log(2\sigma)-\frac{|x-\mu_i|}{\sigma_i}\right]}{\partial \sigma_i} = \sum_{t=1}^{N}\gamma(z_i^t)\left[-\frac{1}{\sigma_i}+\frac{|x-\mu_i|}{\sigma_i^2}\right] = 0$$

$$\sigma_i = \frac{\sum_{t=1}^{N}\gamma(z_i^t)|x^t-\mu_i|}{\sum_{t=1}^{N}\gamma(z_i^t)}$$

$$\frac{\partial \sum_{t=1}^{N}\sum_{i=1}^{K}\gamma(z_i^t)\left[\log(\pi)-log(2\sigma)-\frac{|x-\mu_i|}{\sigma_i}\right]}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}\sum_{t=1}^{N}-\gamma(z_i^t)\frac{|x^t-\mu_i|}{\sigma_i} = 0$$

$\mu_i = median(x^t)$ where $\gamma(z_i^t) = 1$

---

2. In this question, we will implement the EM algorithm to estimate a mixture of Gaussian distributions for image compression.

   (a) (**20 points**) Implement the EM algorithm to estimate a mixture of k Gaussian distributions and run it on the image file "stadium.bmp". Cluster the pixels into $k = \{4, 8, 12\}$ clusters and plot the compressed images for each value of k as described below. (Note: your program might fail if Σ is singular; in this case, restart your EM again. We will fix the problem in part (d)).

   (b) (**10 points**) Run your EM implementation on the "stadium.bmp" image for $k = \{4, 8, 12\}$ and plot the expected complete log-likelihood function $\mathcal{Q}(\Phi|\Phi^l)$ after each E-step and M-step of the EM algorithm in one curve for each value of k. Use different colors to plot the log-likelihood after the E-step and M-step in the curve. Briefly explain the results.

   (c) (**10 points**) Try to run your EM implementation on the image "goldy.bmp" with $k = 7$ and report your observation (**Hint:** If your algorithm falls here, don't panic. Continue to the rest part.). Next, use the k-means function in cluster module of sklearn package to cluster the pixels with $k = 7$. Plot the compressed image given by kmeans. Explain why kmeans and EM behaved differently on the image.

   (d) (**20 points**) Next, implement an improved version of EM to handle singular covariance matrix. In the likelihood function, we can add the following regularization term, $-\frac{\lambda}{2}\Sigma_{i=1}^{k}\Sigma_{j=1}^{d}(\Sigma_i^{-1})_{jj}$, where $(\Sigma_i^{-1})_{jj}$ is the (j, j)-th entry of matrix Σ_i^{-1} and $\lambda > 0$. This regularization term encourages the diagonal of Σ_i^{-1} to be small such that Σ_i is not singular. After adding this regularization term, the expectation step is unchanged and in the

maximization step, the maximum likelihood learning of μ_is and π_is are also unchanged. Derive the maximum likelihood learning of Σ_is using the following result,

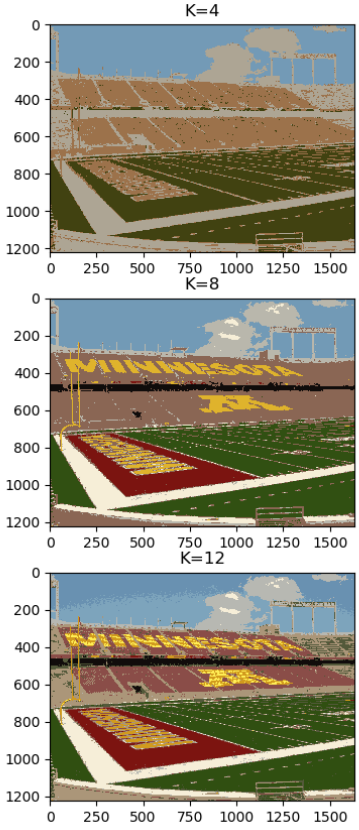
$$\frac{\partial(-\frac{\lambda}{2}\Sigma_{i=1}^{k}\Sigma_{j=1}^{d}(\Sigma_i^{-1})_{jj})}{\partial\Sigma_i^{-1}} = -\frac{\lambda I}{2}$$

(hint: modify the derivation on slide 32 in parametric.pdf to solve the problem).

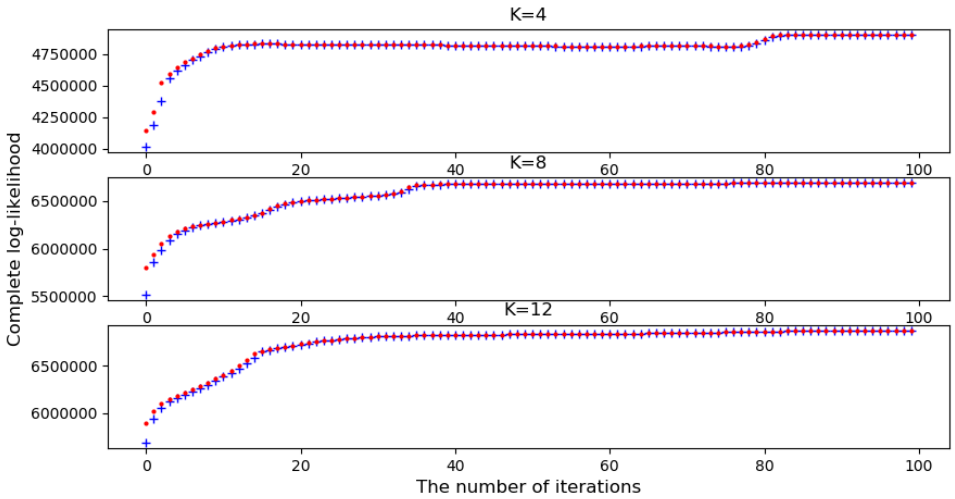
(e) (**10 points**). Implement the new model and test the new model on "goldy.bmp". Explain your observations.

Answer:

(a) The compressed images are the following:



(b) The following image shows the complete log-likelihood function after each E-step and M-step:



(c) The EM implementation should fail with error "SIGMA must be a square, symmetric, positive definite matrix". This happens because the picture has very few colors and data points will have value equals (or close) to the mean. The covariance matrix will have rows containing only zeros, and thus will not be positive definite. K-means uses a hard-assignment approach and does not calculate the covariance matrix.
