STA 243: Homework 3

- Homework due in Canvas: 05/20/2020 at 11:59PM. Please follow the instructions provided in Canvas about homeworks, carefully.

1. Please download the MNIST handwritten digit dataset (training set images and training set labels) from `http://yann.lecun.com/exdb/mnist/`. It contains 28×28-pixel images for the hand-written digits {0,1,...,9} by storing each pixel value ranging between 0 and 255, and their corresponding true labels.

   Load the data into Python. Preprocess the data by compressing each image to 1/4 of the original size in the following way: Divide each 28×28 image into 2×2 non-overlapping blocks. Calculate the mean pixel value of each 2×2 block, and create a new 14×14 image. This preprocessing step will drastically help your computation. We will be clustering the digits $\{0, 1, 2, 3, 4\}$ in this homework. For visualization purpose, we view each data sample as 14×14 matrix. For using in an algorithm, treat each sample as a vector - you just simply stack each column of the 14×14 matrix into a 196 dimensional vector.

2. (**10 Points**) Closely following the derivation in class, we now derive the EM algorithm for Gaussian mixture models. More specifically, we will use 2 models, the "mixture of spherical Gaussians" and the "mixture of diagonal Gaussians". By the end of this question, you should have derived two EM algorithms, one for each model. Below, the following denotes the meaning of each symbol:

   - each $\boldsymbol{\mu}_j$ is a d-dimensional vector representing the cluster center for cluster j
   - each $\boldsymbol{\Sigma}_j$ is a $d \times d$ matrix representing the covariance matrix for cluster j
   - $\sum_{j=1}^{k} \pi_j = 1$ and $\forall j, \pi_j \geq 0$ where all the π_j are the mixing coefficients.
   - Z_i represents the missing variable associated with $\boldsymbol{x}_i$ for $i \in \{1, \ldots, n\}$. It takes integer values $1, \ldots, k$.

   The parameters of the model to be estimated are $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j\}_{j=1}^{k})$. To be more explicit, we will the notation $p(\boldsymbol{x}_i; \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j\}_{j=1}^{k})$ to denote the probability density function $p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. That is, $p(\boldsymbol{x}_i; \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j\}_{j=1}^{k}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. The likelihood of the data for k clusters is:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(\boldsymbol{x}_i; \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j\}_{j=1}^{k}) \qquad (1) \\
&= \prod_{i=1}^{n} \sum_{j=1}^{k} p(\boldsymbol{x}_i | Z_i = j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) p(Z_i = j) \\
&= \prod_{i=1}^{n} \sum_{j=1}^{k} \frac{\pi_j}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \right)
\end{aligned}
$$

   In a **mixture of diagonal Gaussians** model, $\boldsymbol{\Sigma}_j = \text{diag}(\sigma_{j_1}^2, \ldots, \sigma_{j_d}^2)$ is a diagonal matrix for all $j = 1, \ldots, k$. So, we only have d parameters to estimate for the covariance matrix of each cluster.

In a **mixture of spherical Gaussians** model, $\boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}_d$ for all $j = 1, \ldots, k$. Here $\sigma_j > 0$ is a scalar and $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. So we only have 1 parameter to estimate for the covariance matrix of each cluster.

(i) First, we do some preliminary work that will be useful later on.

- **Part a:** Write down the marginal distribution of the Z_i (Hint: What is the probability $p(Z_i = j)$).
- **Part b:** Calculate $p(Z_i = j|\boldsymbol{x}_i)$. (Hint: Bayes Rule)

(ii). We now derive the E- and M- steps. We denote $\boldsymbol{\theta} := \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \pi_j\}_{j=1}^k$. From Equation 1, we can write down the log-likelihood and derive a lower bound:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\
&= \sum_{i=1}^n \log \left[\sum_{j=1}^k p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j) \right] \\
&= \sum_{i=1}^n \log \left[\sum_{j=1}^k \mathbf{F}_{ij} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j)}{\mathbf{F}_{ij}} \right],
\end{aligned}$$

where $F_{ij} > 0$ for all i, j and satisfies

$$\sum_{j=1}^k F_{ij} = 1 \quad for\ i = 1, \ldots, n.$$

Note that if $i = 1$, $\mathbf{F}_{1j} \in \mathbb{R}^k$ corresponds to the the distributions that we pick in the notes, denoted as q_1. Similarly for all i from 1 to n.

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

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log \left[\sum_{j=1}^k \mathbf{F}_{ij} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j)}{\mathbf{F}_{ij}} \right] \geq \sum_{i=1}^n \sum_{j=1}^k \mathbf{F}_{ij} \log \left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j)}{\mathbf{F}_{ij}} \right]. \tag{2}$$

  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).

- **Part d:** (E-Step) We define

$$Q(\mathbf{F}, \boldsymbol{\theta}) := \sum_{i=1}^n \sum_{j=1}^k \mathbf{F}_{ij} \log \left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j)}{\mathbf{F}_{ij}} \right] \tag{3}$$

  to be the lower bound function of $\ell(\boldsymbol{\theta})$. Let $\boldsymbol{\theta}'$ be a fixed value of $\boldsymbol{\theta}$ (e.g., $\boldsymbol{\theta}'$ could be the parameter value of the current iteration). Recall from Part c that we designed $Q(\mathbf{F}, \boldsymbol{\theta})$ to be a lower bound for $\ell(\boldsymbol{\theta})$. We want to make this lower bound as tight as possible. Prove that $\ell(\boldsymbol{\theta}') = Q(\mathbf{F}, \boldsymbol{\theta}')$ when

$$\mathbf{F}_{ij} = p_{\boldsymbol{\theta}'}(Z_i = j|\boldsymbol{x}_i). \tag{4}$$

(iii). Once the E-step is derived, we now derive the M-step. First, we plug

$$\mathbf{F}_{ij} = p_{\boldsymbol{\theta}^{(t)}}(Z_i = j|\boldsymbol{x}_i)$$

into Equation 3 and define the lower bound function at $\boldsymbol{\theta}^{(t)}$ to be

$$Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}) := \sum_{i=1}^{n}\sum_{j=1}^{k} p_{\boldsymbol{\theta}^{(t)}}(Z_i = j|\boldsymbol{x}_i) \log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Z_i = j)}{p_{\boldsymbol{\theta}^{(t)}}(Z_i = j|\boldsymbol{x}_i)}\right]. \tag{5}$$

- **Part e** (M-step for mixture of spherical Gaussians) In the M-step, we aim to find $\boldsymbol{\theta}^{(t+1)}$ that maximize the lower bound function $Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta})$. Under the **mixture of spherical Gaussians** model, derive the M-step updating equations for $\boldsymbol{\mu}_j^{(t+1)}, \boldsymbol{\Sigma}_j^{(t+1)}$ and $\pi_j^{(t+1)}$.
- **Part f** (M-step for mixture of diagonal Gaussians) Under the **mixture of diagonal Gaussians** model, derive the M-step updating equations for $\boldsymbol{\mu}_j^{(t+1)}, \boldsymbol{\Sigma}_j^{(t+1)}$ and $\pi_j^{(t+1)}$.

Hint: Almost all parts above are provided in the class notes. You are required to understand it, replicate it and fill in the missing parts, if any, from the class notes.

3. (**20 Points**) We now implement the EM algorithm from last question to cluster the MNIST data set.
(i) 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).

(ii) 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).

Note that to assign a sample $\boldsymbol{x}_i$ to a cluster j, you first calculate $\mathbf{F}_{ij}$ using the parameters from the last iteration of EM algorithm you implemented. Next, assign sample $\boldsymbol{x}_i$ to the cluster j for which $\mathbf{F}_{ij}$ is maximum, i.e., the probability of sample i belonging to cluster j is maximum. Recall that the dataset has the true labels for each classes. Calculate the error of the algorithm (for the two different model). Here, error is just the number of mis-clustered samples divided by the total number of samples. In your opinion, were mixture models of Gaussian distributions suitable for modeling the MNIST data?

Hint: For these implementations, you will run into three different problems. Apply these following hints to solve each problem.

- 1) 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 $A = \max_j a_j$.
- 2) 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 $\boldsymbol{\Sigma}_j$ for the mixture of diagonal Gaussians, add $0.05\boldsymbol{I}_d$ to $\boldsymbol{\Sigma}_j$ (ie: add 0.05 to all the diagonal elements).

- 3) Be mindful of how you initialize $\mathbf{\Sigma}_j$. Note that for a diagonal matrix $\mathbf{\Sigma}_j$, the determinant $|\mathbf{\Sigma}_j|$ 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.

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 Team Member 2