ISYE6740-HW4

pkubsad

March 2021

(a) (5 points) Select one image of "2" and one image of "6", and visualize the two images.

Answer: Based on the distribution plots we can conclude that:

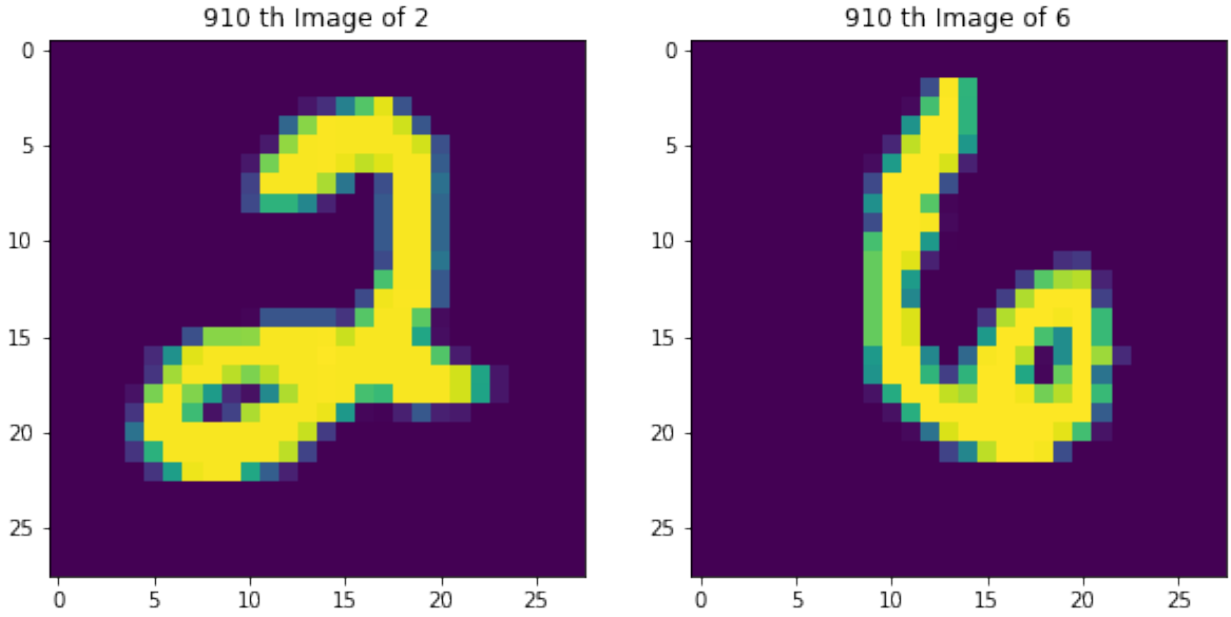

Figure 1: Image of 2 and 6

(b) (10 points) Write down detailed expression of the E-step and M-step in the EM algorithm

Answer:
Expectation step:

$$\tau_k^i = \frac{\pi_k N(x^i|\mu_k, \Sigma_k)}{\Sigma_{k'}^i \pi_{k'} N(x^i|\mu_{k'}, \Sigma_{k'})}$$

As per the homework walk through video,

$$N(x^i|\mu, \Sigma) = \frac{1}{\sqrt{2\pi^n}.\sqrt{|\Sigma|}} * exp(-\frac{1}{2}(X-\mu)^T.\Sigma^{-1}.(X-\mu))$$

$$\tau_k^i = \frac{\pi_k \frac{1}{\sqrt{2\pi^n}.\sqrt{|\Sigma_k|}} * exp(-\frac{1}{2}(X-\mu_k)^T.\Sigma_k^{-1}.(X-\mu_k))}{\Sigma_{k'}^i \pi_{k'} \frac{1}{\sqrt{2\pi^n}.\sqrt{|\Sigma_{k'}|}} * exp(-\frac{1}{2}(X-\mu_{k'})^T.\Sigma_{k'}^{-1}.(X-\mu_k'))}$$

$$\tau_k^i = \frac{\pi_k \frac{1}{\sqrt{|\Sigma_k|}} * exp(-\frac{1}{2}(X-\mu_k)^T.\Sigma_k^{-1}.(X-\mu_k))}{\Sigma_{k'}^i \pi_{k'} \frac{1}{\sqrt{|\Sigma_{k'}|}} * exp(-\frac{1}{2}(X-\mu_{k'})^T.\Sigma_{k'}^{-1}.(X-\mu_k'))}$$

Maximization step: update (μ_k, Σ_k, π_k)

$$\pi_k = \frac{\sum_i \tau_k^i}{m}$$

$$\mu_k = \frac{\sum_i \tau_k^i.x^i}{m}$$

$$\Sigma_k = \frac{\sum_i \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^T}{\sum_i \tau_k^i}$$

where

$$(k = 1 \ldots K, i = 1 \ldots m)$$

(c) (15 points) Implement EM algorithm yourself. Plot the log-likelihood function versus the number of iterations to show your algorithm is converging.

Answer: I have used the demo code provided prof. X. Based on the output of the code, we can see there the log-likelihood varies until iteration 6 and then it settles down varying marginally.The algorithm finally converges after 22 iterations. Please find below the plot of log-likelihood vs iteration:

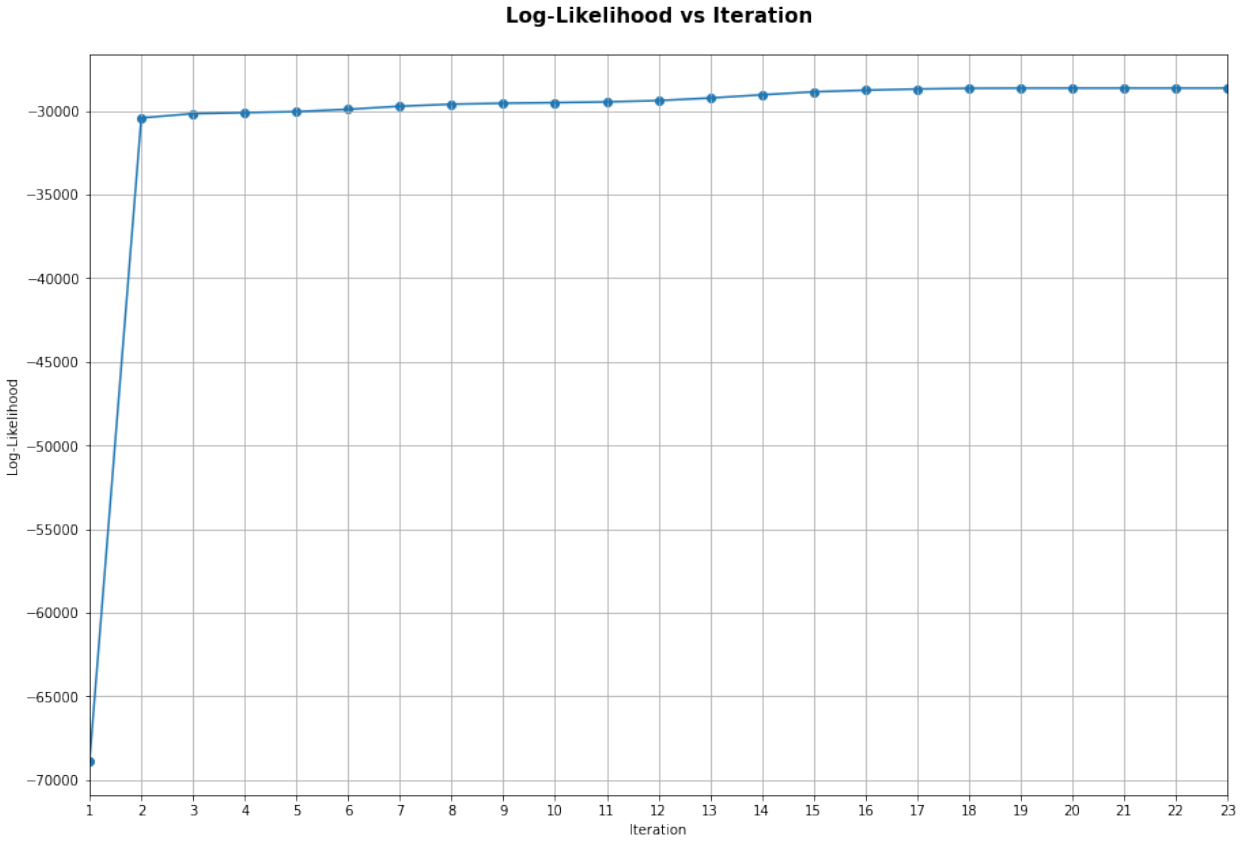

Figure 2: log-likelihood vs iterations

(d) (15 points) Report, the fitted GMM model when EM has terminated in your algorithms as follows. Make sure to report the weights for each component, and the mean of each component (you can reformat the vector to make them into 28-by-28 matrices and show images).

Answer: Mean values representation:

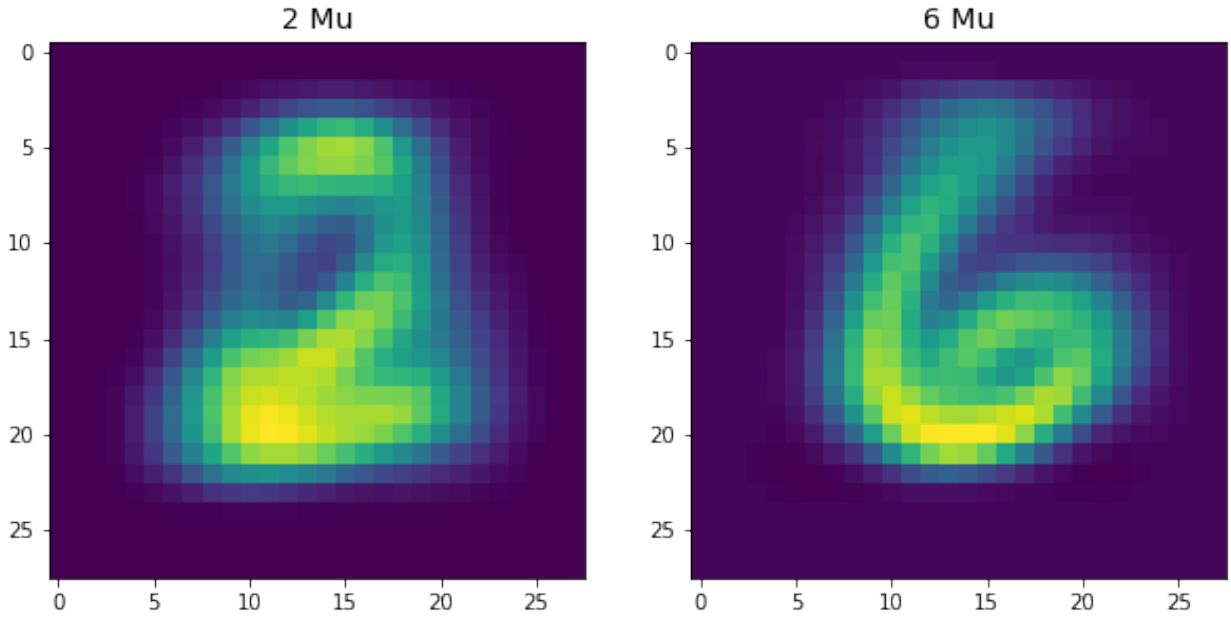

Figure 3: mean value representation of 2 components

Covariance values representation:

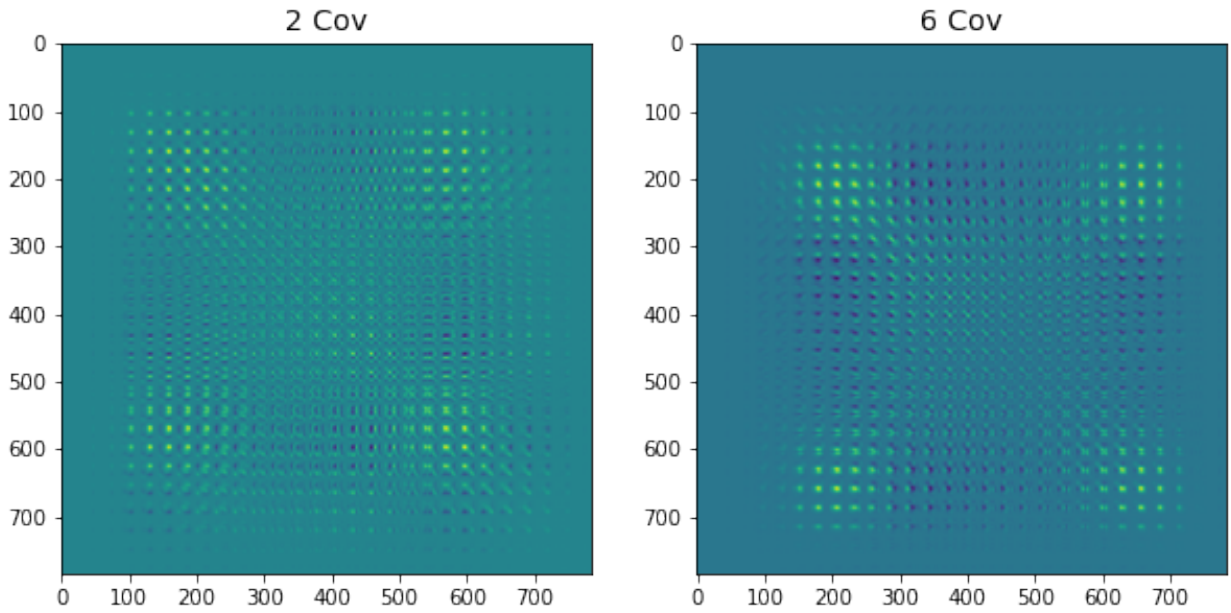

Figure 4: Covariance value representation of 2 components

(e) (15 points) Use the τ_k^i to infer the labels of the images, and compare with the true labels. Report the mis-classification rate for digits "2" and "6" respectively. Perform K-means clustering with $K = 2$ (you may call a package or use the code from your previous homework). Find out the mis-classification rate for digits

"2" and "6" respectively, and compare with GMM. Which one achieves the better performance?

Answer: Below is observations for number of miscalculations for 2 and 6:

C	EM	K-Means
2	51/1032= 4.94%	55/1032 = 5.33%
6	9/958 = 0.93%	68/958 = 7.10 %

Based on this observation, we can conclude EM algorithm performs better than K-Means clustering for this use case.

(2.1) (10 points) Show step-by-step mathematical derivation for the gradient of the cost function

Answer: As mentioned by prof in lecture, to calculate the gradient of the cost function, we will calculate partial derivative wrt theta.

$$l(\theta) = \sum_{i=1}^{n} \{-log(1 + exp{-}\theta x_i) + (y_i - 1)\theta x_i\}$$

$$\frac{\partial(l(\theta))}{\partial\theta} = \frac{\partial(\sum_{i=1}^{n}\{-log(1 + exp{-}\theta x_i) + (y_i - 1)\theta x_i\})}{\partial\theta}$$

$$\frac{\partial(l(\theta))}{\partial\theta} = \frac{\partial(\sum_{i=1}^{n}\{-log(1 + exp{-}\theta x_i)\}}{\partial\theta} + \frac{\partial\{(y_i - 1)\theta x_i\})}{\partial\theta}$$

Applying the chain rule to the derivative:

$$= \sum_{i=1}^{n} -\frac{(1 + exp\{-\theta x_i\})'}{(1 + exp\{-\theta x_i\})} + \sum_{i=1}^{n}(y_i - 1)x_i$$

$$= \sum_{i=1}^{n} \frac{x_i.exp\{-\theta x_i\}}{(1 + exp\{-\theta x_i\})} + \sum_{i=1}^{n}(y_i - 1)x_i$$

$$= \sum_{i=1}^{n} \frac{x_i.exp\{-\theta x_i\}}{(1 + exp\{-\theta x_i\})} + (y_i - 1)x_i$$

(2.2) (10 points) Write a pseudo-code for performing **gradient descent** to find the optimize

Answer: We cannot set the above function to 0 because that would not give closed formed solution. Instead we can perform gradient descent by following the below steps:

(a) Begin with randomly initialized point θ^0, in general, θ^i

(b) Select a step size γ and perform the below calculation:

$$\theta^{i+1} = \gamma.gradient\_function, f(\theta)$$

$$\theta^{i+1} = \gamma.\frac{\partial(l(\theta))}{\partial\theta}$$

$$\theta^{i+1} = \gamma.\sum_{i=1}^{n} \frac{x_i.exp\{-\theta x_i\}}{(1 + exp\{-\theta x_i\})} + (y_i - 1)x_i$$

(c) Define optimality condition, like a threshold value t. If θ^t - θ^{t+1} <= t, stop the process.

(d) Select decreasing step size at every iteration to avoid oscillation when the algorithm reaches optimal value.

(2.3) (10 points) Write the pseudo-code for performing the **stochastic gradient descent** algorithm to solve the training of logistic regression problem (**??**). Please explain the difference between gradient descent and stochastic gradient descent for training logistic regression.

Answer: Stochastic Gradient Descent (SGD) process:

(a) Split the entire data into K buckets.

(b) Perform the Gradient Descent algorithm mentioned above for each bucket.

(c) Update global parameters after ever bucket is executed.

(d) Iterate through all the buckets calculating the cost function for each bucket.

The main difference between Stochastic gradient descent (SGD) and gradient descent (GD) is, in GD we run the algorithm on the entire data set. If the data set is large, this is not an efficient way, wrt , compute and memory to iterate and calculate gradient descent. In SGD, we break the entire data into smaller buckets and perform GD on each bucket. This is lot faster compared to running GD on whole data set.

(2.4) (10 points) We will **show that the training problem in basic logistic regression problem is concave.** Derive the Hessian matrix of $\ell(\theta)$ and based on this, show the training problem (**??**) is concave (note that in this case, since we only have one feature, the Hessian matrix is just a scalar). Explain why the problem can be solved efficiently and gradient descent will achieve a unique global optimizer, as we discussed in class.

Answer: *Source* : *https* : *//math.wikia.org/wiki/Hessian_matrix* From the source mentioned above, definition of Hessian matrix , I am taking derivative of the gradient of the log likelihood function from Q2.1.

$$= \frac{\partial(\sum_{i=1}^{n} \frac{x_i.exp\{-\theta x_i\}}{(1+exp\{-\theta x_i\})} + (y_i - 1)x_i)}{\partial\theta}$$

$$= \frac{\partial(\sum_{i=1}^{n} \frac{x_i.exp\{-\theta x_i\}+(y_i-1)x_i(1+exp\{-\theta x_i\}))()}{(1+exp\{-\theta x_i\})}}{\partial\theta}$$

$$= \frac{\partial}{\partial}(\sum_{i=1}^{n} \frac{x_i[y_i exp\{-x_i\} + y_i - 1]}{1 + exp\{-\theta x_i\}})$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \frac{(x_i.y_i.exp(-\theta x_i) + x_i y_i - x_i)}{1 + exp\{-\theta x_i\}}$$

Applying the chain rule and quotient rule to the derivative. (source: https://web.iit.edu/sites/web/files/departments/ac affairs/academic-resource-center/pdfs/product_quotient_rule.pdf)

$$= \sum_{i=1}^{n} \frac{(-x_i^2 * y_i.exp\{-\theta x_i\}).(1 + exp\{-\theta x_i\}) - (-x_i.exp\{-\theta x_i\}).(x_i.y_i.exp\{-\theta x_i\} + x_i.y_i - x_i)}{(1 + exp\{-\theta x_i\})^2}$$

$$= \sum_{i=1}^{n} \frac{-x_i^2.exp(-\theta x_i)}{1 + exp(-\theta x_i)^2}$$

This expression is a quadratic function which proves second derivative of the likelihood function is a concave optimization problem.