Machine Learning
Assignment 3 — E0270
Due: 10th June 2021

Instructions

- Answer all questions
- You need to submit the solutions in two files:
  - A zip file containing the code for all the questions. This file should be named with a .pdf extension, as `[NameOfStudent]_[SR_Number]_Code.zip.pdf`
    It should contain three folders named `q1`, `q2a` and `q2b`, containing code and plots for the corresponding questions.
  - A report in the form of a PDF file, named `[NameOfStudent]_[SR_Number]_Report.pdf`. This should contain all the derivations, details of the experiments etc. Please note that if you do not submit the report or fail to include what has been asked to be included for any particular question in the report, you will not be given marks for that question even if you submit the code.
- Late submissions will be penalized.

1. (20 points) In this problem, you will derive and implement the Expectation-Maximization algorithm for finding the parameters of a mixture of exponential distributions.
   Consider a mixture of K probability distributions, with probability density function

$$f(x;\theta) = \sum_{k=1}^{K} \tau_k f_k(x;\lambda_k),$$

   where $f_k(x;\lambda_k) = \lambda_k e^{-\lambda_k x}$ is the probability density function of an exponential distribution with parameter λ_k, and $\tau = (\tau_1, \ldots, \tau_K)$ is the vector of mixture probabilities. A sample can be obtained from f by sampling $k \in \{1, \ldots, K\}$ with probabilities specified by τ and then sampling from f_k, i.e, the k'th distribution is chosen with probability τ_k. $\theta = (\tau, \lambda)$ is the set of parameters of the mixture distribution, which are unknown and have to be estimated using the EM algorithm.

   Suppose n data points $X = \{x_i\}_{i=1,\ldots,n}$ are given that had been sampled from the mixture distribution. Each sample x_i corresponds to an unknown latent variable z_i that specifies which distribution among $f_1, \ldots, f_k$, x_i has been sampled from. Let $Z = \{z_i\}_{i=1,\ldots,n}$ be the set corresponding unknown latent variables.

   (a) (2 points) Give the expression for the likelihood function $L(\theta; X, Z) = \mathbb{P}(X, Z|\theta)$.

   (b) (3 points) Let $\theta^{(t)} = (\tau^{(t)}, \lambda^{(t)})$ be the current estimate of the unknown parameters θ. For the E-step in the EM algorithm, we need the conditional distribution of Z given the data and the current estimates of the parameters.
   Derive the detailed formula for determining

$$h_{i,j}^{(t)} = \mathbb{P}\left(z_i = j \mid x_i, \theta^{(t)}\right).$$

   The above quantity is the probability that the sample x_i came from the distribution f_j, provided the mixture distribution is parameterized by the current estimates of the parameters.

   (c) (5 points) Now that we have the conditional distribution of Z, we need to find the expectation of the log-likelihood under this distribution. Derive the detailed expression for

$$Q\left(\theta, \theta^{(t)}\right) = \mathbb{E}_{Z|X,\theta^{(t)}}\left[\log L(\theta; X, Z)\right],$$

   in terms of $h_{i,j}^{(t)}, \lambda, \tau$ and x_i's.

   (d) (5 points) The M-step consists of finding the parameters that optimize the above objective. Derive the formula for finding the values of λ and τ that maximize Q, i.e, solve the following optimization problem:

$$(\lambda, \tau) = \arg\max_{(\lambda,\tau)} Q\left((\lambda, \tau), (\lambda^{(t)}, \tau^{(t)})\right).$$

   (e) (5 points) Consider the data points given in the file `mixture_data.txt`, in which each line consists of a random sample from a mixture of two exponential distributions with unknown parameters. Use the EM algorithm derived above for finding

the mixture parameters and the parameters of the individual exponential distributions.

In the report, you have to state the parameters determined by the EM algorithm and include two plots, the first plot depicting the values of λ_1 and λ_2 throughout the progress of the algorithm, the second plot similarly depicting the values of τ_1 and τ_2.

Also submit the python script that you have used for doing the above. It should contain code for reading the data from the given file, running the EM algorithm, displaying the progress and final results of the algorithm, and plotting the evolution of the parameter estimates.

2. (20 points) For this question, you will train Generative Adversarial Networks (GANs) on two datasets:
   (a) (8 points) The two-dimensional data given in the file `gan_data.txt`.
   (b) (12 points) The FashionMNIST dataset.

   The report should contain the following:

   - Description of the network architectures used for each of these two experiments, the training methodology, the loss functions and hyperparameters used, description of any preprocessing done, as well any other relevant details and observations.
   - Plots depicting the generator and discriminator losses throughout the training process.
   - Ten visualizations generated at regular intervals during the training process, depicting samples generated by the partially trained generator network. For example, if you are updating the neural network parameters 2000 times, then generate a plot visualizing the samples generated by the generator network after every 200 update steps.
     - For question 2(a), each visualization should be a 2D scatter plot of 500 samples generated by the partially trained generator network.
     - For question 2(b), **each** visualization should contain 100 sample images generated by the partially trained generator network. An example visualization is shown in Figure 1.

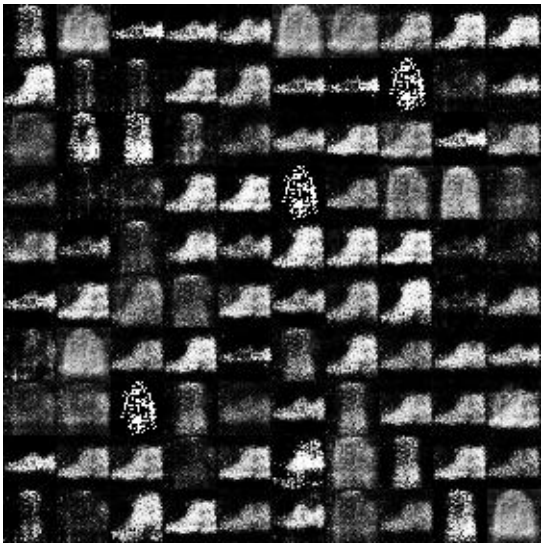

Figure 1: Example visualization of a set of 100 FashionMNIST samples generated by a partially trained generator network. Ten such visualizations should be included in the report, created at regular intervals during the training process.