Task:

The task is to cluster images in the MNIST dataset. We wish to find one cluster for each digit in an unsupervised manner. For the purpose of simplicity, we only consider instances with label 2, 3, 4. So our task is to find 3 clusters, which correspond to 2, 3, 4 respectively, without looking at training labels. Here we use the Bernoulli mixtures as the clustering algorithm. Bernoulli mixtures method assumes feature values are binary. The original data are represented as pixels ranging from 0 to 255. So we first normalize features to numbers between 0 and 1 and then further make them into binary values by thresholding at 0.5.

Questions

These are five interesting things I did:

1. Using EM, we could get μ_k, corresponds to each component. So we can find change of digits by drawing μ_k for first a few iterations.
2. To monitor the progress of EM, we evaluate the objective function of EM after each iteration. So drawing objective against iteration numbers is my second section.
3. Bernoulli mixture provides density estimation for the given data, so we could sample digits from this density so that the sampled ones look like the given ones.
4. We could also use the trained Bernoulli mixture model for classification. We can use, say a subset images (select 2,3,4 only) to train BM, and use the remaining (also select 2,3,4 only) to test classification accuracy. The training is unsupervised. After training, we could get 3 centroids and weights corresponding to three digits. And we should look at the 3 centroids and know which cluster corresponds to which digit. For a test data point x, we could compute $p(z = k|x) = \frac{p(z=k,x)}{p(x)}$, and pick whichever cluster k gives the max probability. Then for each x, we can predict the digit for cluster k.
5. It is also interesting to check what happens if we set k is bigger than 3.