CSCI 5090/7090 Machine Learning, Spring 2018: Homework 4

Due: Monday, April 30th, 9:00 PM

1 K-means Clustering [100 points]

In this assignment you will implement k-means clustering algorithm on the MNIST dataset of handwritten digits, which consists of 60,000 handwritten digits (0-9) that have been scanned in and scaled to 28 × 28 pixels. The data is available at `http://yann.lecun.com/exdb/mnist/`.

1. Implement K-means algorithm. For initial cluster centers, use random points. Repeat the random start 10 times for each clustering run. After getting the K-means result with 10 different initializations, how can you determine the best starting point? For the following questions, use the best initialization for your final result.

2. We define the objective function of K-means as the sum of the squared distances of each point to its cluster centers, $\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ki} - \mu_k)^2$. Run your program with $K = 10$ and plot the values of objective function against iterations. Is it monotonically decreasing?

3. Try running it with $K = 16$ and plot the objective function again. How is the behavior of the objective function different from when $K = 10$?

4. Clustering performance is hard to evaluate. However, since we have the true labels, we can use the following heursitics. For each cluster C, we find the most frequent (true) label Y_C and label the instances in that cluster with the majority label Y_C . Report your precision (number of correctly labeled instances / number of all instances) and final value of the objective function for $K = 1, 5, 10, 16, 20$.

5. Among the five values you tried above, what would you choose to be the optimal number of clusters and why?