CSE 847 (Spring 2022): Machine Learning— Homework 5

Instructor: Jiayu Zhou

1 Clustering: K-means

1. Elaborate the relationship between k-means and spectral relaxation of k-means. Is it possible that we obtain exact k-means solution using spectral relaxed k-means?

2. Implementation of k-means. Submit all the source code to D2L along with a short report on your observation.

   - Implement the k-means in MATLAB using the alternating procedure introduced in the class (you will not get the credit if you use the build-in `kmeans` function in MATLAB).
   - Implement the spectral relaxation of k-means. Create a random dataset and compare the k-means and spectral relaxed k-means.

2 Principle Component Analysis

1. Suppose we have the following data points in 2d space $(0,0), (-1,2), (-3,6), (1,-2), (3,-6)$.

   - Draw them on a 2-d plot, each data point being a dot.
   - What is the first principle component? Given 1-2 sentences justification. You do not need to run MATLAB to get the answer.
   - What is the second principle component? Given 1-2 sentences justification. You do not need to run MATLAB to get the answer.

2. **Experiment**: We apply data pre-processing techniques to a collection of handwritten digit images from the USPS dataset (data in MATLAB format: USPS.mat)[1]. You can load the whole dataset into MATLAB by `load USPS.mat`. The matrix A contains all the images of size 16 by 16. Each of the 3000 rows in A corresponds to the image of one handwritten digit (between 0 and 9). To visualize a particular image, such as the second one, first you need to convert the vector representation of the image to the matrix representation by `A2 = reshape(A(2,:), 16, 16)`, and then use `imshow(A2')` for visualization.

   Implement Principal Component Analysis (PCA) using SVD and apply to the data using $p = 10, 50, 100, 200$ principal components. Reconstruct images using the selected principal components from part 1.

   - Show the source code links for parts 1 and 2 to your github account.
   - The total reconstruction error for $p = 10, 50, 100, 200$.
   - A subset (the first two) of the reconstructed images for $p = 10, 50, 100, 200$.

   Note: The USPS dataset is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps`. The image size is 16 by 16, thus the data dimensionality of the original dataset is 256. We used a subset of 3000 images in this homework.

[1] `https://github.com/jiayuzhou/CSE847/blob/master/data/USPS.mat?raw=true`