CSE 847 (Spring 2016): Machine Learning— Homework 5

Instructor: Jiayu Zhou

Due on Tuesday, April 19th, before class.

1 Clustering: K-means and Gaussian Mixture Model

1. Textbook Page 456, Question 9.6
2. Textbook Page 456, Question 9.9
3. Given N data points $x_i, (i = 1, ..., N)$, Kmeans will group them into K clusters by minimizing the distortion function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \|x_n - \mu_k\|^2$, where μ_k is the center of the k^{th} cluster; and $r_{n,k} = 1$ if x_n belongs to the k^{th} cluster and $r_{n,k} = 0$ otherwise. In this exercise, we will use the following iterative procedure

   - Initialize the cluster center $\mu_k, (k = 1, ..., K)$;
   - Iterate until convergence
     - Update the cluster assignments for every data point x_n: $r_{n,k} = 1$ if $k = \text{argmin}_j \|x_n - \mu_j\|^2$; $r_{n,k} = 0$ otherwise.
     - Update the center for each cluster k: $\mu_k = \frac{\sum_{n=1}^{N} r_{n,k} x_n}{\sum_{n=1}^{N} r_{n,k}}$

   Remember in Gaussian Mixture Model, $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $\pi_k = p(z_k = 1)$ is the prior for the k^{th} component; and μ_k,Σ_k are the mean and covariance matrix for k^{th} component respectively. In the E-step, we will update $p(z_k = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}$ Now suppose that

   (1) $\Sigma_k = \epsilon \mathbf{I}$ where ϵ is some *given* number;

   (2) $\pi_k \neq 0$ $(k = 1, ..., K)$;

   (3) $\|x_n - \mu_i\| \neq \|x_n - \mu_j\|$ for any $i \neq j$.

   Under the above assumptions, prove that when $\epsilon \to 0$, $p(z_k = 1|x_n) = r_{n,k}$, where $r_{n,k}$ is the cluster assignment used in 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

[1] `https://github.com/jiayuzhou/CSE847-2016Spring/blob/master/homework/USPS.mat`

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.

Apply Principal Component Analysis (PCA) 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.