CSE472 (Machine Learning Sessional)
Assignment 2: PCA and EM algorithm

Introduction

Principal component analysis (PCA) and the expectation-maximization (EM) algorithm are two of the most widely used unsupervised methods in machine learning. In this assignment, you will use PCA for dimensionality reduction and apply the EM algorithm for Gaussian mixture model to cluster the data with dimensionality reduced.

Dataset

You are given a tab separated file titled "data.txt" to be used as the dataset for this assignment. The file contains 500 rows and 100 columns. The 500 rows correspond to 500 sample points and each sample is represented by a 100 dimensional feature vector.

PCA implementation

Let X be a $N \times m$ data matrix where N is the number of dimensions and m is the number of instances (notice that your dataset is not in this format). Perform principal component analysis (PCA) of X as follows:

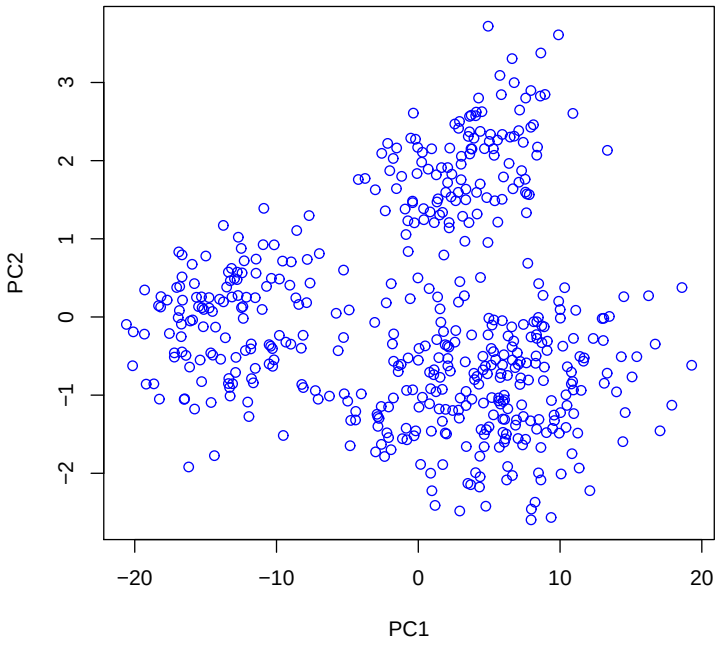
1. Scale and center each dimension of X such that they have zero mean and unit variance.
2. Construct the co-variance matrix $\Sigma = XX^T$.
3. Perform eigen decomposition of Σ

$$\Sigma = Q\Lambda Q^T$$

   The columns of Q will give the eigen vectors and diagonal entries of Λ will give the eigen values.

Note: You can call library functions to perform matrix operations such as eigen decomposition but do not call library functions to perform entire PCA.

Now project your data along the two eigen vectors corresponding to the two highest eigen values. You now have 500 samples each having two dimensions. Plot of the data should look like below (or some rotation of that).



EM implementation

Now we will cluster the two dimensional data assuming a Gaussian mixture model using the EM algorithm.

Let a vector x with dimension D can be generated from any one of the k Gaussian distribution where the probability of selection of Gaussian distribution i is w_i where $\sum_{i=1}^{k} w_i = 1$ and the probability of generation of x from Gaussian distribution i is given as,

$$N_i(x_j|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{(-\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1}(x_j-\mu_i))}$$

To learn a Gaussian mixture model using EM algorithm, we need to maximize the likelihood function with respect to the parameters (comprising

the means and covariances of the components and the mixing coefficients). The steps are given below.

1. Initialize the means μ_i, covariances Σ_i and mixing coefficients w_i, and evaluate the initial value of the log likelihood. Use the plot to estimate number of clusters and initialize the parameters.

2. **E-step:** Evaluate the conditional distribution of latent factors using the current parameter values

$$p_{ij} = p(z_j = i|x_j, \mu, \Sigma, w) = \frac{p(x_j|z_j = i, \mu, \Sigma, w)P(z_j = i|\mu, \Sigma, w)}{p(x_j|\mu, \Sigma, w)} = \frac{w_i N_i(x_j|\mu_i, \Sigma_i)}{\sum_{i=1}^{k} w_i N_i(x_j|\mu_i, \Sigma_i)}$$

3. **M-step:** Evaluate the conditional distribution of latent factors using the current parameter values

$$\mu_i = \frac{\sum_{j=1}^{N} p_{ij} x_j}{\sum_{j=1}^{N} p_{ij}}$$

$$\Sigma_i = \frac{\sum_{j=1}^{N} p_{ij}(x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{N} p_{ij}}$$

$$w_i = \frac{\sum_{j=1}^{N} p_{ij}}{N}$$

4. Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.
   $\ln p(X|\mu, \Sigma, w) = \sum_{j=1}^{N} \ln p(x_j|\mu, \Sigma, w) = \sum_{j=1}^{N} \ln(\sum_{i=1}^{k} w_i N_i(x_j|\mu_i, \Sigma_i))$

Submission

1. Upload the codes in Moodle within 9:00 P.M. of 7th December, 2018 (Sunday). (Strict deadline)

2. Write code in a single *.py file, then rename it with your student id. For example, if your student id is 1405123, then your code file name should be 1405123.py and the report name should be 1405123.pdf.

3. Finally make a main folder, put the code and report in it, and rename the main folder as your student id. Then zip it and upload it.

Evaluation

1. You have to reproduce your experiments during in-lab evaluation. Keep everything ready to minimize delay.
2. You are likely to give online tasks during evaluation which will require you to modify your code.
3. You will be tested on your understanding through viva-voce.
4. If evaluators like performance, efficiency or modularity of a particular code, they can give bonus marks. This will be completely at the discretion of evaluators.
5. You are encouraged to bring your computer in the sessional to avoid any hassle. But in that case, ensure an internet connection as you have to instantly download your code from the Moodle and show it.

Warning

1. Dont copy! We regularly use copy checkers.
2. First time copier and copyee will receive negative marking because of dishonesty. Their default is bigger than those who will not submit.
3. Repeated occurrence will lead severe departmental action and jeopardize your academic career. We expect Fairness and honesty from you. Dont disappoint us!