CSE472 (Machine Learning Sessional)
Assignment 3: Dimensionality Reduction using Principal Component Analysis and Clustering using Expectation-maximization 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 1000 rows and 100 columns. The 1000 rows correspond to 1000 sample points and each sample is represented by a 100-dimensional feature vector.

PCA implementation

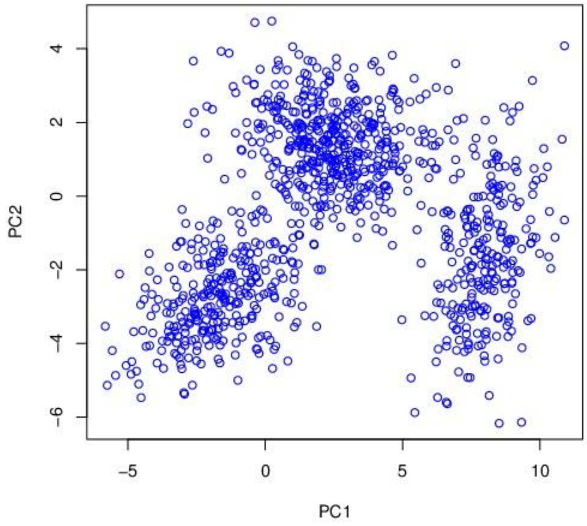
Let X be a $N \times D$ data matrix where D is the number of dimensions and N is the number of instances. Perform principal component analysis (PCA) of X as follows:

1. Construct the co-variance matrix

$$S = \frac{1}{N}\left(\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\right)$$

2. Compute the eigen vectors and eigen values of the co-variance matrix. You can call library functions to perform matrix operations such as eigen decomposition but do not call library functions to perform entire PCA.

3. Now project your data along the two eigen vectors corresponding to the two highest eigen values. You now have 1000 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 k is w_k where $\sum_{k=1}^{K} w_k = 1$ and the probability of generation of **x** from Gaussian distribution k is given as,

$$N_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} e^{\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right)}$$

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 $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients w_i, and evaluate the initial value of the log likelihood.

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

$$p_{ik} = p(z_i = k|\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w}) = \frac{p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w})P(z_i = k|\boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w})}{p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w})}$$

$$= \frac{w_k N_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} w_k N_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

3. **M step**: Re-estimate the parameters using the conditional distribution of latent factors

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} p_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} p_{ik}}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} p_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} p_{ik}}$$

$$w_k = \frac{\sum_{i=1}^{N} p_{ik}}{N}$$

4. **Evaluate** the log likelihood and check for convergence of the log likelihood. If the convergence criterion is not satisfied return to step 2.

$$\ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma},\ \boldsymbol{w}) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} w_k N_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

Submission

1. Upload the codes in Moodle by **9:00 P.M. of 9th November, 2020 (Monday)**. (Strict deadline)
2. You need to submit a report file in pdf format containing the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients w_i, and the PCA plot. No hardcopy is required.
3. Write code in a single *.py file, then rename it with your student id. For example, if your student id is 1505123, then your code file name should be “1505123.py” and the report name should be “1505123.pdf”.
4. 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 may 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. Please ensure an internet connection as you have to instantly download your code from the Moodle and show it.

Warning

1. Don’t 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. Don’t disappoint us!