UCSB Spring 2019

ECE 283: Homework 3

Topics: Unsupervised Learning, part 1 (Gaussian mixtures and EM algorithm; K-means and soft K-means
Assigned: Monday April 29
Due: Friday May 10
Reading: Posted notes on EM algorithm, Gaussian mixtures and K-means and variants; Bishop, Chapter 9

Experiments with 2D data
For this, let us use the same data statistics as in HW1, except that we will not use the labels. Thus, we have three Gaussian components, with $\pi_1 = 1/2$, $\pi_2 = 1/6$, $\pi_3 = 1/3$ (component 1 corresponds to class 0 in HW1, components 2 and 3 are components of class 1 in HW1).
Component 1: Gaussian with mean vector $\mathbf{m} = (0, 0)^T$ and covariance matrix $\mathbf{C}$ with eigenvalue, eigenvector pairs:
$\lambda_1 = 2$, $\mathbf{u}_1 = (\cos\theta, \sin\theta)^T$, $\lambda_2 = 1$, $\mathbf{u}_2 = (-\sin\theta, \cos\theta)^T$, with $\theta = 0$.
Component 2: $\mathbf{m} = (-2, 1)^T$, $\mathbf{C}$ with eigenvalue, eigenvector pairs: $\lambda_1 = 2$, $\mathbf{u}_1 = (\cos\theta, \sin\theta)^T$, $\lambda_2 = 1/4$, $\mathbf{u}_2 = (-\sin\theta, \cos\theta)^T$, with $\theta = -\frac{3\pi}{4}$.
Component 3: $\mathbf{m} = (3, 2)^T$, $\mathbf{C}$ with eigenvalue, eigenvector pairs: $\lambda_1 = 3$, $\mathbf{u}_1 = (\cos\theta, \sin\theta)^T$, $\lambda_2 = 1$, $\mathbf{u}_2 = (-\sin\theta, \cos\theta)^T$, with $\theta = \frac{\pi}{4}$.

1) Generate $N = 200$ data samples from the preceding model, saving both the data point $\mathbf{x}_i$ and $\mathbf{z}_i \in \{0, 1\}^3$, the one-hot encoding of which component the data point belongs to. Implement the K-means algorithm with different values of $K = 2, 3, 4, 5$. For each K, start with several different random initializations, and choose the run that leads to the smallest mean squared error. Let $\{\mathbf{m}_k, k = 1, ..., K\}$ denote the cluster centers, and for each data point, compute $\mathbf{a}_i \in \{0, 1\}^K$, the one-hot encoding of which cluster the data point is assigned to. Plot the empirical probabilities $P[a_i[k] = 1|z_i[l] = 1]$, $l = 1, 2, 3$, $k = 1, ..., K$ in a $3 \times K$ table, indicating how the "ground truth" components map to the clusters you learn.

2) Using the results of 1) as a starting point (there are some details to be worked out on exactly how), implement the EM algorithm to estimate the mean and covariance for a Gaussian mixture model with $K = 2, 3, 4, 5$ components. Plot the average values of $p(k|i), k = 1, ..., K$ for data points drawn from each ground truth component, $l = 1, 2, 3$, in a $3 \times K$ table.

3) Comment (using 2D plots) on how the means and covariances of the mixture components you find relate geometrically to the "ground truth" mixture components for different values of K.

Experiments with data in higher dimensions
Let us now see what happens when we increase the number of dimensions to d (d to be played with, but the nominal value is $d = 30$), while keeping the "effective dimension" (I'm not defining this formally, but you should soon see what I mean once I tell you how to generate the data) smaller than d.

4) Write the following program to generate a random vector $\mathbf{u}$ in d dimensions as follows:
The components of $\mathbf{u}$ are i.i.d., with

$$P[u[i] = 0] = 2/3, \quad P[u[i] = +1] = 1/6, \quad P[u[i] = -1] = 1/6$$

Let $\{\mathbf{u}_j, j = 1, ..., 7\}$ be i.i.d. draws from your program in 4). Draw them once, and then fix them. Check that the vectors are quasi-orthogonal. If two of them are "too correlated" (what does that mean?), then purge one of them and draw another vector. We will use these vectors to generate data samples coming from a Gaussian mixture distribution, as follows.

5) Write a program to use the fixed vectors from 4) to generate d-dimensional data samples for a Gaussian mixture distribution with 3 equiprobable components, as follows. In order to generate any given data point $\mathbf{X}$, we will use i.i.d. draws from a standard Gaussian ($N(0, 1)$)

distribution that we will denote $\{Z_m\}$, and we will also draw a "noise vector" $\mathbf{N} \sim N(0, \sigma^2 \mathbf{I}_d)$ (default value $\sigma^2 = 0.01$). A sample from each of the three components can now be described as follows (remember, a given data sample belongs to exactly one of these components):

Component 1: Generate $\mathbf{X} = \mathbf{u}_1 + Z_1 \mathbf{u}_2 + Z_2 \mathbf{u}_3 + \mathbf{N}$.

Component 2: Generate $\mathbf{X} = 2\mathbf{u}_4 + \sqrt{2} Z_1 \mathbf{u}_5 + Z_2 \mathbf{u}_6 + \mathbf{N}$.

Component 3: Generate $\mathbf{X} = \sqrt{2}\mathbf{u}_6 + Z_1(\mathbf{u}_1 + \mathbf{u}_2) + \frac{1}{\sqrt{2}} Z_2 \mathbf{u}_5 + \mathbf{N}$.

Note that the vectors $\{\mathbf{u}_j\}$ stay the same across data samples, but the random numbers Z_1 and Z_2, and the noise vector $\mathbf{N}$ are drawn afresh for each sample.

6) Generate $N =??$ (to be determined) data samples from the preceding model, saving both the data point $\mathbf{x}_i$ and $\mathbf{z}_i \in \{0,1\}^3$, the one-hot encoding of which component the data point belongs to. Implement the K-means algorithm with different values of $K = 2, 3, 4, 5$. For each K, start with several different random initializations, and choose the run that leads to the smallest mean squared error. Let $\{\mathbf{m}_k, k = 1, ..., K\}$ denote the cluster centers, and for each data point, compute $\mathbf{a}_i \in \{0,1\}^K$, the one-hot encoding of which cluster the data point is assigned to. Plot the empirical probabilities $P[a_i[k] = 1 | z_i[l] = 1]$, $l = 1, 2, 3$, $k = 1, ..., K$ in a $3 \times K$ table, indicating how the "ground truth" components map to the clusters you learn.

7) Try to provide geometric insight into how the cluster centers found by K-means relate to the vectors $\{\mathbf{u}_j\}$ in the model.

8) Run the EM algorithm with several different values of K. Comment on how the eigenvectors of the covariance matrices you find relate to the parameters of the Gaussian mixture model.

9) **Optional bonus problem:** Play with deterministic annealing for this dataset, and report on your findings.