CMPE 544: Pattern Recognition (Fall 2020)

Homework #2

Due December 25, 2020 by 11:59pm on Moodle

In this homework, you will practice Expectation-Maximization (EM) algorithm for Gaussian mixture model. Please read the explanations below and implement the EM algorithm for mixture of Gaussians.

EM for Gaussian Mixtures

Suppose we have a dataset of observations $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ and we wish to model the data using mixture of Gaussians. The Gaussian mixture distribution can be written as a linear superposition of Gaussians as given below:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

We can introduce a latent (unobserved/unknown) variable $\mathbf{z} \in \mathbb{R}^K$ binary vector, where $z_k = 1$ and rest is zero if the data point is from kth normal distribution. The marginal distribution over $\mathbf{z}$ is defined in term of the mixing coefficient π_k:

$$p(z_k = 1) = \pi_k$$

As a result, π_k must satisfy the conditions below:

$$0 \leq \pi_k \geq 1 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1$$

The conditional distribution of $\mathbf{x}$ for a given value of $\mathbf{z}$ follows the Normal distribution:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

The conditional probability of $\mathbf{z}$ given $\mathbf{x}$ can be obtain using the Bayes' theorem:

$$\begin{aligned}\gamma(z_k) = p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)\, p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)\, p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \mathbf{\Sigma}_j)}\end{aligned}$$

The log likelihood for the Gaussian mixture model:

$$\ln p(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k) \right]$$

If we try to maximize the log likelihood with respect to μ_k and $\mathbf{\Sigma}_k$, we obtain

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, \mathbf{x}_n$$

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, (\mathbf{x}_n - \mu_k)\, (\mathbf{x}_n - \mu_k)^T$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Since we need to make sure $\sum_{k=1}^{K} \pi_k = 1$, Lagrange multiplier method is used to optimize the log likelihood with respect to π_k:

$$\ln p\left(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}\right) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right)$$

which gives $\pi_k = \frac{N_k}{N}$.

Steps of EM Algorithm for Gaussian Mixtures

1. Initialize the means μ_k, covariances $\mathbf{\Sigma}_k$ and the mixing coefficients π_k. Evaluate the initial value of the log likelihood.

2. **E-step:** Compute

$$\gamma\left(z_n k\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n|\mu_j, \mathbf{\Sigma}_j\right)}$$

3. **M-step:**

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n$$

$$\mathbf{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(\mathbf{x}_n - \mu_k^{new}\right) \left(\mathbf{x}_n - \mu_k^{new}\right)^T$$

$$\pi_k^{new} = \frac{N_k}{N}$$

   where

$$N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right)$$

4. Compute the value of the log likelihood to check the convergence. Convergence criteria can be stopping when the difference between the current and previous values of the log likelihood is negligible.

$$\ln p\left(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}\right) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k\right)\right]$$

Task

Download the synthetic data, `dataset.npy`, provided on Moodle. The synthetic dataset is a mixture of 3 Gaussians. Implement and run the EM algorithm for the synthetic data. After the convergence, plot the cluster assignments denoted by different colors. Also report the estimated values for the mean and the covariance matrix. Please submit your code as a .py file. Notebooks will not be accepted. Your code should be able to run from terminal when located in the same folder as the dataset. If your code has bugs, 30 points will be deducted. Please submit a report including your

plot and results. Please describe how you implemented the EM algorithm in your report. If you use any script from any source online, please cite the references. Please do not copy paste any text from internet, or from another student. This is not a group activity. If cheating is detected, you will get -100.