In the Name of God

Introduction to Machine Learning (25737-2)

Project Phase 1

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

1 Expectation Maximization

Finding the MAP/MLE estimates for the parameters of a mixture model requires solving for two sets of unknowns simultaneously: the latent variables of the data (e.g. which cluster a datapoint belongs to) and the parameters of each mixture. Solving this problem directly is often hard. Suppose Z denotes the cluster each sample belongs to, and $\mathbf{X}$ denotes the observations. Our aim is to maximize the likelihood function:

$$\ln\left(p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\right) = \sum_{i=1}^{N} \ln\left(p_Y(y^{(i)};\boldsymbol{\theta})\right) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} p_{Y,Z}(y^{(i)}, Z^{(i)}=k;\boldsymbol{\theta})\right) \tag{1}$$

Theory Question 1. In your own words, explain how the MM algorithm can deal with non-convex optimization objective functions by considering simpler convex objective functions.

We define hidden variables to model distribution coeffiecients. Then we re-write log likelyhood. As we will say later, the terms would be simpler. We find a tight lower bound for log likelyhood to maximize. We put q as it is said and then update other parameters by derivative.

Theory Question 2. Briefly explain how the formula for mixture models:

$$p(\mathbf{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_Z(z_k;\boldsymbol{\theta}) p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|Z=z_k;\boldsymbol{\theta}),$$

is the same as the sum over all possible values of $Z^{(i)}$ in equation (1). Explain why it's easier to optimize $p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}_n,\mathbf{z}_n;\theta)$ than $p_{\mathbf{Y}}(\mathbf{y}_n;\theta)$ in the context of mixture models.

as we had earlier in eq (3) $p_{\mathbf{Y}|Z}(\mathbf{y}|Z=z_k;\boldsymbol{\theta}) = p_k(\mathbf{y};\boldsymbol{\theta}_k)$ and $p_Z(z_k;\boldsymbol{\theta}) = \pi_k$ so :

$$\sum_{k=1}^{K} p_Z(z_k;\boldsymbol{\theta}) p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|Z=z_k;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y};\boldsymbol{\theta}_k) = p(\mathbf{y};\boldsymbol{\theta})$$

$p_{\mathbf{Y}}(\mathbf{y}_n;\theta)$ has multiple expressions (sum of all distributions) however $p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}_n,\mathbf{z}_n;\theta)$ is a single expressions.

Theory Question 3. Read about variational inference (or variational bayesian methods) and compare it with the procedure we used for the EM algorithm (You might want to check Wikipedia for this!).

2 EM Algorithm for GMM and CMM

In this section, we want you to apply EM algorithm to learn(estimate) parameters for two different mixture model and find closed-form solution of their parameters.

2.1 EM for Gaussian Mixture Model

Theory Question 4. Compute estimate of parameters for Gaussian Mixture Models for N observed data $\{\mathbf{x}_i\}_{i=1}^N$.

1. Determine model parameters and initialize them.

   Model parameters are q_k's and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean vector and covariance matrix respectively.

2. Compute complete dataset likelihood[1].

   by using Bayes rule we have:

$$\begin{aligned} p(\mathcal{D};\boldsymbol{\theta}) = p(\{\mathbf{y}_i\}_{i=1}^N, \{\mathbf{z}_i\}_{i=1}^N;\boldsymbol{\theta}) &= \prod_{n=1}^N p(\mathbf{y}_n, \mathbf{z}_n\ ;\boldsymbol{\theta}) \\ &= \prod_{n=1}^N p(\mathbf{y}_n|\mathbf{z}_n\ ;\boldsymbol{\theta}) p(\mathbf{z}_n;\boldsymbol{\theta}) \\ &= \prod_{n=1}^N p(\mathbf{y}_n|\mathbf{z}_n\ ;\boldsymbol{\theta}_{\mathbf{z}_n}) \pi_{\mathbf{z}_n} \\ &= \prod_{n=1}^N \prod_{k=1}^K (p(\mathbf{y}_n|\mathbf{z}_n;\boldsymbol{\theta}_k)\pi_k)^{\mathbb{1}(k=\mathbf{z}_n)} \\ &= \prod_{n=1}^N \prod_{k=1}^K (\mathcal{N}(y_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\pi_k)^{\mathbb{1}(k=\mathbf{z}_n)} \end{aligned}$$

3. Find closed-form solution for parameters using EM algorithm.

   E step: we put $q_n^* = p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}_n \mid \mathbf{y}_n;\boldsymbol{\theta})$ as said earlier. we use Bayes rule:

$$q_n^{(t+1)}(k) = \frac{p_{\mathbf{Y}|\mathbf{Z}_n}(\mathbf{y}_n \mid \mathbf{z}_n;\boldsymbol{\theta})\, p_{\mathbf{Z}_n(z_n)}}{p_{\mathbf{Y}}(y;\boldsymbol{\theta})} = \frac{\pi_k p_k(\mathbf{z}_n \mid \boldsymbol{\theta}_k)}{\sum_{k'=1}^K \pi_{k'} p_{k'}(\mathbf{y}_n \mid \boldsymbol{\theta}_{k'}^{(t)})} = \frac{\pi_k^{(t)} \mathcal{N}(y_n;\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)})}{\sum_{k'=1}^K \pi_{k'}^{(t)} \mathcal{N}(y_n;\boldsymbol{\mu}_{k'}^{(t)},\boldsymbol{\Sigma}_{k'}^{(t)})}$$

   M step:

   now we will use eq (22) to derive π_k.from part two we can write $l(\theta^{(t)})$:

$$\begin{aligned} l^{(t+1)}(\boldsymbol{\theta}) &= \sum_n \mathbb{E}_{q_n^{(t+1)}}\left[\ln\left(p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y}_n, \mathbf{z}_n;\boldsymbol{\theta})\right)\right] \\ &= \sum_{n=1}^N \sum_{k=1}^K q_n^{(t+1)}(k)\left[\mathbb{1}(k=z_n)(\ln(p_k(y_n;\boldsymbol{\theta}_k)) + \ln(\pi_k))\right] \\ &= \sum_{n=1}^N \sum_{k=1}^K q_n^{(t+1)}(k)(\ln(p_k(y_n;\boldsymbol{\theta}_k)) + \ln(\pi_k)) \end{aligned}$$

   now we optimize above expression. note that $\sum_{k=1}^K \pi_k = 1$.so we solve it with Lagrange multipliers.

$$\frac{\partial}{\partial \pi_l}(l(\theta^{(t+1)}) - \lambda(\sum_{k=1}^K \pi_k - 1))$$

$$\sum_{k=1}^K \pi_k = 1$$

   by solving above equations we have:

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^N q_n^{(t+1)}(k)}{\sum_{n=1}^N \sum_{k=1}^K q_n^{(t+1)}(k)} = \frac{1}{N}\sum_{n=1}^N q_n^{(t+1)}(k)$$

[1] $p(\mathcal{D};\boldsymbol{\theta}) = p(\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{z}_i\}_{i=1}^N;\boldsymbol{\theta})$

now we will maximize it with respect to $\boldsymbol{\theta}$.

$$l^{(t+1)}(\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} q_n^{(t+1)}(k)(\ln(\mathcal{N}(y_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)) + \ln(\pi_k))$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} q_n^{(t+1)}(k)(\ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{y}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k) + \ln(\pi_k))$$

now we differentiate with respect to vector $\boldsymbol{\mu}_k$ and matrix $\boldsymbol{\Sigma}_k$.

$$\frac{\partial}{\partial\boldsymbol{\mu}_k} l^{(t+1)} = \frac{1}{2}\sum_{n=1}^{N} q_n^{(t+1)}(k)(\boldsymbol{\Sigma}_k^{-1} + (\boldsymbol{\Sigma}_k^{-1})^T)(\mathbf{y}_n-\boldsymbol{\mu}_k) = 0$$

$$\longrightarrow \sum_{n=1}^{N} q_n^{(t+1)}\mathbf{y}_n = (\sum_{n=1}^{N} q_n^{(t+1)})\boldsymbol{\mu}_k$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} q_n^{(t+1)}\mathbf{y}_n}{\sum_{n=1}^{N} q_n^{(t+1)}}$$

here we used the vector differentiate identity $\frac{\partial}{\partial x}(x^TAx) = (A+A^T)x$

$$\frac{\partial}{\partial\Delta_k}(\sum_{n=1}^{N}\sum_{k=1}^{K} q_n^{(t+1)}(k)(\ln|\Delta_k| - \frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_k)^T\Delta_k(\mathbf{y}_i-\boldsymbol{\mu}_k)))$$

$$\sum_{n=1}^{N} q_n^{(t+1)}(k)(\Delta_k)^{-1} - \sum_{n=1}^{N} q_n^{(t+1)}(k)(\mathbf{y}_n-\boldsymbol{\mu}_k)(\mathbf{y}_n-\boldsymbol{\mu}_k)^T = 0$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{n=1}^{N} q_n^{(t+1)}(k)(\mathbf{y}_n\mathbf{y}_n^T - \boldsymbol{\mu}_k^{(t+1)}(\boldsymbol{\mu}_k^{(t+1)})^T)}{\sum_{n=1}^{N} q_n^{(t+1)}(k)}$$

2.2 EM for Categorical Mixture Model

Theory Question 5. Compute estimate of parameters for Categorical Mixture Models for N observed data $\{\mathbf{x}_i\}_{i=1}^N$.

1. Determine model parameters and initialize them.

$$p_{\mathbf{Y}}(\mathbf{y};\theta) = \sum_{k=1}^{K}\pi_k Cat(\mathbf{y},\theta_k)$$

the model parameters are π and θ. Initializing these parameters:

$$\pi_k = \frac{1}{K}, \theta_i = \frac{\vec{1}_d}{d}$$

2. Compute complete dataset likelihood.

$$\ln p(\mathcal{D}|\theta) = \ln\prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\theta) = \sum_{n=1}^{N}\ln p(\mathbf{x}_n,\mathbf{z}_n|\theta)$$

$$= \sum_{n=1}^{N}\ln[p(\mathbf{x}_n|\mathbf{z}_n,\theta)p(\mathbf{z}_n|\theta)] = \sum_{n=1}^{N}\ln[\prod_{k=1}^{K}\mathrm{Cat}(\mathbf{x}_n|\theta_k)^{z_{nk}}\prod_{k=1}^{K}\pi_k^{z_{nk}}]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\prod_{c=1}^{C}\theta_{kc}^{x_{nc}} + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\pi_k = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\sum_{c=1}^{C} x_{nc}\ln\theta_{kc} + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\pi_k$$

3. Find closed-form solution for parameters using EM algorithm.

   for Expectation we have:

$$q_{nk}^* = p(\mathbf{z}_n = k|\mathbf{y}_n; \boldsymbol{\theta}) = \frac{\pi_k \text{Cat}(\mathbf{y}_n; \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k \text{Cat}(\mathbf{y}_n; \boldsymbol{\theta}_k)}$$

   For maximization step we have:

$$\pi_k^* = \frac{1}{N} \sum_{n=1}^{N} q_{nk}^*$$

$$\theta_k^* = \frac{\sum_{n=1}^{N} q_{nk}^* y_n}{\sum_{n=1}^{N} q_{nk}^*}$$

---

3 EM Algorithm in Real Applications

Suppose you have a set of MRI images of patients with brain tumors. You want to estimate the volume and location of the tumors from the images. You assume that each image is generated by a mixture of three Gaussian distributions: one for the background, one for the normal brain tissue and one for the tumor tissue. However, some of the images are corrupted by noise or artifacts (missing data). We have given you the three Gaussian distributions and need to estimate their means and variances. Use EM algorithm to estimate the parameters of the mixture model to help us to fill in the missing data.

Answer all simulation questions for both `Image1.csv` and `Image2.csv` datasets and compare the parameters obtained from these two.

The answers to simulation questions are available in notebook file.

Simulation Question 1. Each distribution has 200 data points that are concatenated in a two-dimensional array and given to you. Plot the data with three different colors in a graph.

Simulation Question 2. Write a function that performs the E-step. This means assigning each data to a distribution based on the Euclidean distance. Return as output a 3×600 array specifying which distribution each data belongs to. If the R_{ij} is one, it means that the i–th data is assigned to the j–th distribution. Run this function for one iteration and report the result.

Simulation Question 3. Write a function that performs M-step. This means updating the mean and variance of each distribution. Run this function for one iteration and report the new variances and means of each distribution.

Simulation Question 4. Using the functions you have written, run the EM algorithm until a convergence is reached or the maximum number of steps is passed. Replot the three new distributions and compare with the correct labels.

Simulation Question 5. Compare the parameters obtained from each of the images and explain the reason for their difference.
