In the Name of God

Introduction to Machine Learning (25737-2)

Project Report (Phase 1)

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Alireza Shokrani 99106255

Amirhossein Akbari 99105901

Theory Question 1

In your own words, explain how the MM algorithm can deal with nonconvex optimization objective functions by considering simpler convex objective functions.
Answer:
Non-convex functions that we work with usually have local maximums. MM algorithm tend to use the data of the function in a small interval to approximate the behavior of it locally. In that small interval, we can compute the value of function and its derivative to make another convex function (i.e. a 2nd order polynomial) that behaves alike the original function and by computing the maximum point of that alternative function we're guaranteed to do better in making our way to the local maximum as we're looking at things locally. The algorithm continues this process to the point that we are enough close to a local maximum.

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 (9). 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.
Answer:
The given phrase inside the sum, can just be validated by using the conditional probability formula. Here we used conditional form instead of the joint distribution because calculating the joint form is not so easy and the whole point of using hidden variables was to make the problem easier by assuming that we already know from which of the mixture components our data point is coming. The reason why the sum over these values gives us the probability distribution of Y is the total probability theorem. We calculate the marginal distribution by integrating out the hidden variables.

Theory Question 3

Read about variational inference (or variational bayesian methods) and compare it with the procedure we used for the EM algorithm.
Answer:
Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. There are two main uses for these methods:
1- To provide an analytical approximation to the posterior probability of the unobserved variables.
2- To derive a lower bound for the marginal likelihood (sometimes called the evidence) of the observed data.
In this project we used EM algorithm, a simple member of this family. In the E-step we tried the first use and in the M-step we tried the second use mentioned above. The main difference between EM algorithm and other approaches is that in the M-step of the EM algorithm we tend to give a single best guess and update the model's parameters with that using MLE. But not all other algorithms do that. In fact Variational Bayesian Methods have a more generalized and Bayesian approach and try to give an approximation of the whole posterior distribution (MAP). This is done by minimizing an integral that comes from the distance measure (i.e. KL-divergence). If we use KL-divergence, the integral will consists of two distributions. In E and M step we're actually trying to minimize those two distributions (q in E-step and $\boldsymbol{\theta}$ in M-step).

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.
   Answer:
   The model consists of K Gaussian distributions that each one of them has a mean vector $\boldsymbol{\mu}_k$ and variance matrix $\boldsymbol{\Sigma}$. So the purpose of the EM algorithm here is to find $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ for $k \in \{1, \dots, K\}$ and k numbers showing as π_k which are the coefficients for each Gaussian distribution in the mixture model.

2. Compute complete dataset likelihood.
   Answer:

$$p(\{\mathbf{x_i}\}_{i=1}^N, \{\mathbf{z_i}\}_{i=1}^N; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x_i}, \mathbf{z_i}; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x_i}|\mathbf{z_i}; \boldsymbol{\theta}) p(\mathbf{z_i}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k p(\mathbf{x_i}; \theta_k))^{z_{i,k}}$$

$$p(\mathbf{x_i}; \theta_k) = \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{i,k}}$$

3. Find closed-form solution for parameters using EM algorithm.
   Answer:
   E-step:

$$p(\mathbf{z_i}, \mathbf{x_i}; \theta^{(t-1)}) = \frac{p(\mathbf{x_i}, \mathbf{z_i}; \theta^{(t-1)})}{p(\mathbf{x_i})} = \frac{p(\mathbf{x_i}|\mathbf{z_i}; \theta^{(t-1)}) p(\mathbf{z_i})}{p(\mathbf{x_i})}$$

$$= \frac{\prod_{k=1}^{K} (\pi_k p_k(\mathbf{x_i}; \theta_k^{(t-1)}))^{z_{i,k}}}{\sum_{k=1}^{K} \pi_k p_k(\mathbf{x_i}; \theta_k^{(t-1)})}$$

$$p_k(\mathbf{x_i}; \theta_k^{(t-1)}) = \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k^{(t-1)}, {\boldsymbol{\Sigma}_k}^{(t-1)})$$

$$q_i^{(t)}(\mathbf{z_i}) \longleftarrow \frac{\prod_{k=1}^{K}(\pi_k \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k{}^{(t-1)}))^{z_{i,k}}}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k{}^{(t-1)})}$$

M-step:

$$\begin{aligned}
\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] &= \mathbb{E}_q\left[\ln \prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k p(\mathbf{x_i}; \boldsymbol{\theta}_k))^{z_{i,k}}\right] \\
&= \mathbb{E}_q\left[\sum_{i=1}^{N}\sum_{k=1}^{K} z_{i,k}(\ln(\pi_k) + \ln p(\mathbf{x_i}; \boldsymbol{\theta}_k))\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_q\left[z_{i,k}\right](\ln(\pi_k) + \ln p(\mathbf{x_i}; \boldsymbol{\theta}_k)) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} q_{i,k}(\ln(\pi_k) + \ln p(\mathbf{x_i}; \boldsymbol{\theta}_k)) \\
p(\mathbf{x_i}; \boldsymbol{\theta}_k) &= \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}$$

$$\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{i,k}(\ln(\pi_k) + \ln \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}_k|}} - \frac{1}{2}(\mathbf{x_i} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x_i} - \boldsymbol{\mu}_k))$$

Updating $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\boldsymbol{\mu}_k^{(t)} &= \arg\max_{\boldsymbol{\mu}_k} \mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] \\
\frac{d}{d\boldsymbol{\mu}_k}\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] &= 0 \\
\sum_{i=1}^{N} q_{i,k}\left(\frac{\mathbf{x_i} - \boldsymbol{\mu}_k}{\sigma^2}\right) &= 0 \\
\boldsymbol{\mu}_k^{(t)} &= \frac{\sum_{i=1}^{N} q_{i,k}\mathbf{x_i}}{\sum_{i=1}^{N} q_{i,k}}
\end{aligned}$$

Updating $\boldsymbol{\Sigma}_k$:

$$\begin{aligned}
\boldsymbol{\Sigma}_k^{(t)} &= \arg\max_{\boldsymbol{\Sigma}_k} \mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] \\
\frac{d}{d\boldsymbol{\Sigma}_k}\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] &= 0 \\
\boldsymbol{\Sigma}_k^{(t)} &= \frac{\sum_{i=1}^{N} q_{i,k}(\mathbf{x_i} - \boldsymbol{\mu}_k^{(t)})(\mathbf{x_i} - \boldsymbol{\mu}_k^{(t)})^\top}{\sum_{i=1}^{N} q_{i,k}}
\end{aligned}$$

Updating π_k:

$$\pi_k^{(t)} = \frac{1}{N}\sum_{i=1}^{N} q_{i,k}$$

Theory Question 5

Compute estimate of parameters for Categorical Mixture Models for N observed data $\{\mathbf{x_i}\}_{i=1}^{N}$. (We will be using some the results from the previous question to answer this question)

1. Determine model parameters and initialize them.
   Answer:
   The model consists of K Categorical distributions that each one of them has a probability vector $\boldsymbol{\theta}_k$ that each element of it is the probability of the output to be 1 in that place and 0 in other entries. So the purpose of the EM algorithm here is to find $\boldsymbol{\theta}_k$ and k numbers showing as π_k which are the coefficients for each Categorical distribution in the mixture model.

2. Compute complete dataset likelihood.
   Answer:

$$p(\{\mathbf{x_i}\}_{i=1}^N, \{\mathbf{z_i}\}_{i=1}^N; \boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k p(\mathbf{x_i}; \boldsymbol{\theta}_k))^{z_{i,k}}$$
$$p(\mathbf{x_i}; \boldsymbol{\theta}_k) = \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k)$$
$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}(\pi_k \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k))^{z_{i,k}}$$

3. Find closed-form solution for parameters using EM algorithm.
   Answer:
   E-step:

$$p(\mathbf{z_i}, \mathbf{x_i}; \boldsymbol{\theta}^{(t-1)}) = \frac{\prod_{k=1}^{K}(\pi_k p_k(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)}))^{z_{i,k}}}{\sum_{k=1}^{K}\pi_k p_k(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)})}$$
$$p_k(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)}) = \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)})$$
$$q_i^{(t)}(\mathbf{z_i}) \longleftarrow \frac{\prod_{k=1}^{K}(\pi_k \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)}))^{z_{i,k}}}{\sum_{k=1}^{K}\pi_k \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k^{(t-1)})}$$

   M-step:

$$\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{i,k}(\ln(\pi_k) + \ln p(\mathbf{x_i}; \boldsymbol{\theta}_k))$$
$$p(\mathbf{x_i}; \boldsymbol{\theta}_k) = \mathrm{Cat}(\mathbf{x_i}; \boldsymbol{\theta}_k) = \prod_{m=1}^{M}(\theta_{k,m})^{x_{i,m}}$$
$$\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{i,k}(\ln(\pi_k) + \sum_{m=1}^{M} x_{i,m}\ln(\theta_{k,m}))$$

   Updating $\boldsymbol{\theta}_k$:

$$\boldsymbol{\theta}_k^{(t)} = \arg\max_{\boldsymbol{\theta}_k} \mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right]$$
$$\frac{d}{d\boldsymbol{\theta}_k}\mathbb{E}_q\left[\ln p(\mathcal{D}; \boldsymbol{\theta})\right] = 0$$
$$\boldsymbol{\theta}_k^{(t)} = \frac{\sum_{i=1}^{N} q_{i,k}\mathbf{x_i}}{\sum_{i=1}^{N} q_{i,k}}$$

   Updating π_k:

$$\pi_k^{(t)} = \frac{1}{N}\sum_{i=1}^{N} q_{i,k}$$

Simulation Questions

The answer to this section is in the Jupiter Notebook.