STAT 542: Statistical Learning

Gaussian Mixture Models, EM and MM Algorithms

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course Website: https://teazrq.github.io/stat542/

April 21, 2022

Department of Statistics
University of Illinois at Urbana-Champaign

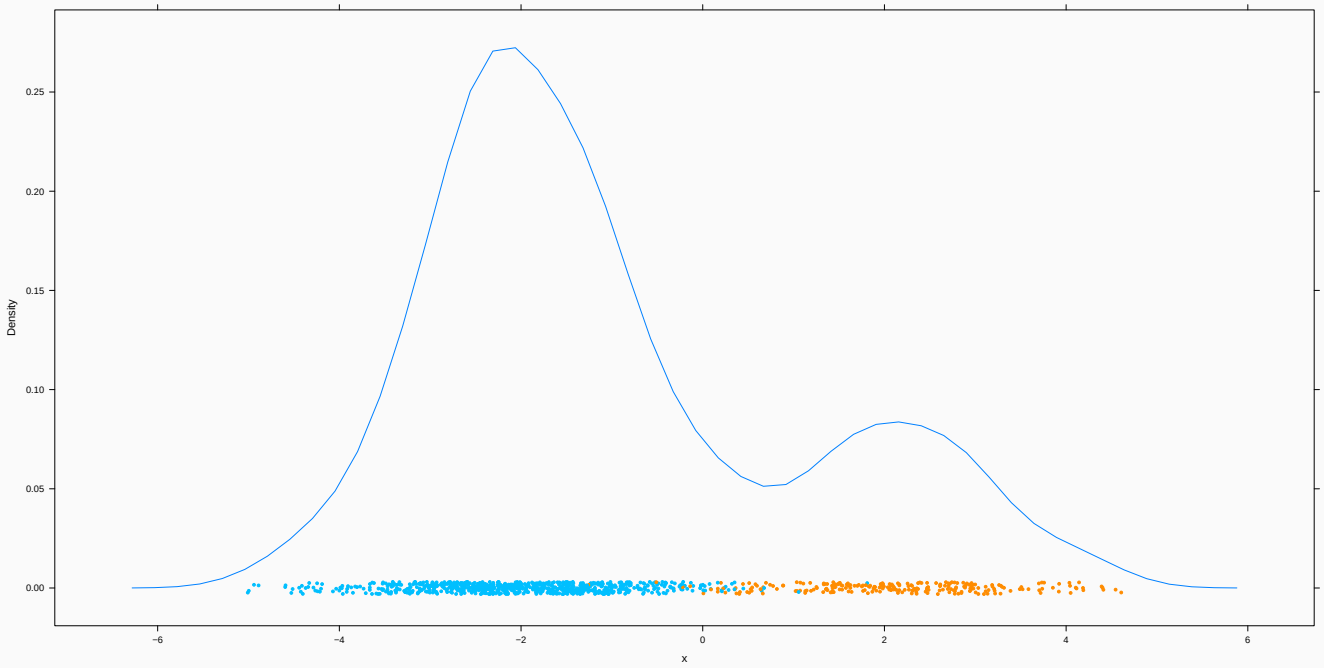
Outline

- Gaussian Mixture Models
- The EM Algorithm
- The MM Algorithm
- Ascend Property of EM

Gaussian Mixture Models

Gaussian Mixture

- Suppose
  - We know that there are two populations, with means μ_1 and μ_2, respectively, and variance $\sigma^2 = 1$.
  - X is the observed outcome (from one of the two populations), and $Z \in \{0, 1\}$ is a hidden variable (not observable) that indicates the population label, with $P(Z = 1) = \pi$.
  - From only the observed data $\{x_i\}_{i=1}^n$, we want to estimate the two population means and the mixing probability: $\boldsymbol{\theta} = (\mu_1, \mu_2, \pi)$.



Gaussian Mixture

- The density of X is a mixture of two Gaussian:

$$\mathbf{p}_X(x) = (1-\pi)\phi_{\mu_1}(x) + \pi\phi_{\mu_2}(x)$$

  where ϕ_μ is the density function of $\mathcal{N}(\mu, 1)$.

- The log-likelihood function based on n observed training data is

$$\ell(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left[(1-\pi)\phi_{\mu_1}(x_i) + \pi\phi_{\mu_2}(x_i)\right]$$

- Of course, we can solve this by gradient descent, however, that is often slow.

- Instead, we can treat the hidden labels Z as a "missing variables" and use the EM algorithm.

Gaussian Mixture: The EM algorithm

- Instead of directly optimizing $\ell(\mathbf{x}|\boldsymbol{\theta})$, we introduce the latent variable Z, and write the joint log-likelihood as

$$\ell(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[(1 - z_i) \log \phi_{\mu_1}(x_i) + z_i \log \phi_{\mu_2}(x_i)\right] + \sum_{i=1}^{n} \left[(1 - z_i) \log(1 - \pi) + z_i \log \pi\right]$$

- EM algorithm: We will then optimize this likelihood function by iteratively updating the unknowns: $\mathbf{z}$ and $\boldsymbol{\theta}$.

— At the E-step (expectation), we treat both $\mathbf{x}$ and (μ_1, μ_2, π) as known, and calculate the conditional probability of each z_i.

— At the M-step (maximization), we treat $\mathbf{x}$ and $\mathbf{z}$ as known, and solve the parameters by maximizing the likelihood.

Gaussian Mixture: E-step

- E-step, if both $\mathbf{x}$ and $\boldsymbol{\theta} = (\mu_1, \mu_2, \pi)$ are known, then the conditional probability of each z_i can be calculated as:

$$\begin{aligned} \mathrm{P}(Z_i = 1 | \boldsymbol{\theta}^{(k)}, \mathbf{x}) &= \frac{\mathbf{p}(Z_i = 1, x_i | \boldsymbol{\theta}^{(k)})}{\mathbf{p}(x_i | \boldsymbol{\theta}^{(k)})} \\ &= \frac{\mathbf{p}(Z_i = 1, x_i | \boldsymbol{\theta}^{(k)})}{\mathbf{p}(Z_i = 1, x_i | \boldsymbol{\theta}^{(k)}) + \mathbf{p}(Z_i = 0, x_i | \boldsymbol{\theta}^{(k)})} \end{aligned}$$

- This is pretty simple since we just need to calculate the densities functions of each x_i under the current parameter $\boldsymbol{\theta}^{(k)}$ for each possible label ($z_i = 0$ or 1).

Gaussian Mixture: E-step

- Lets first set up the initial values and estimate the conditional probabilities for each z_i

```
> # generate the data:
> n = 1000; x1 = rnorm(n, mean=-2)
> x2 = rnorm(n, mean=2); z = (runif(n) <= 0.25)
> x = ifelse(z, x2, x1)
>
> # lets setup some (arbitrary) initial values:
> hat_PI = 0.5
> hat_mu1 = -0.25
> hat_mu2 = 0.25
>
> # E step
> # calculate the conditional distribution of the hidden variable z
> d1 = hat_PI * dnorm(x, mean= hat_mu1)
> d2 = (1-hat_PI) * dnorm(x, mean= hat_mu2)
> ez = d2/(d1 + d2)
```

Gaussian Mixture: M-step

- Now we already have $\mathbf{p}(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(k)})$, we can replace all the z_i values (since they are unknown anyways) in the likelihood function $\ell(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ by their expectations (from the E-step).

$$\widehat{p}_i = \mathbf{p}(Z_i = 1|\mathbf{x}, \boldsymbol{\theta}^{(k)})$$

- After this, things remained in the likelihood only involves $\mathbf{x}$ and $\boldsymbol{\theta}$, so we can solve (the M-step) for the "MLE" of $\boldsymbol{\theta}$ based on this new likelihood function.

- It turns out that these estimators are just weighted means:

$$\widehat{\mu}_1 = \frac{\sum_{i=1}^n (1 - \widehat{p}_i) x_i}{\sum_{i=1}^n (1 - \widehat{p}_i)}, \quad \widehat{\mu}_2 = \frac{\sum_{i=1}^n \widehat{p}_i x_i}{\sum_{i=1}^n \widehat{p}_i} \quad \text{and} \quad \widehat{\pi} = \sum_{i=1}^n \widehat{p}_i / n$$

- We will then iterate the E- and M- steps until convergence.

The EM algorithm: M-step

```
> # M-step
> # based on the conditional distribution, calculate the new MLE of the
    parameters
> PI = mean(ez)
> hat_mu1 = sum( (1-ez) * x ) / sum(1-ez)
> hat_mu2 = sum( ez * x ) / sum(ez)
```

The EM algorithm: Gaussian Mixture

- The algorithm converges pretty fast after a few iterations: $(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\pi})$

```
[1] -1.7424035  0.1277127  0.3877030
[1] -2.2467959  0.9673550  0.3825091
[1] -2.2117884  1.3957797  0.3310913
[1] -2.1518538  1.6386121  0.2993035
[1] -2.1167579  1.7706276  0.2828132
[1] -2.0986018  1.8367258  0.2747542
[1] -2.0897518  1.8682925  0.2709414
[1] -2.0855753  1.8830238  0.2691684
[1] -2.0836364  1.8898248  0.2683511
[1] -2.0827434  1.8929490  0.2679758
[1] -2.0823335  1.8943810  0.2678039
```

Connection with the K-mean

- Suppose in the E-step, we do not use the "soft" label (the probability), instead, we use a "hard" label:

$$\mathbf{1}\{\mathbf{p}(Z_i = 1|\mathbf{x}, \boldsymbol{\theta}^{(k)}) > 0.5\}$$

- This is essentially comparing the two densities and see which one is larger. Notice that the log density of Gaussian is just the Euclidean norm, we are just choosing the "closer" cluster mean (of the previous iteration).
- The M-step is just re-calculating the cluster means based on the new assignments.

The EM algorithm for general purpose

The EM algorithm

- Suppose that we want to maximize the a log-likelihood

$$\log \mathbf{p}(\mathbf{x}|\boldsymbol{\theta})$$

- Under some scenarios this likelihood can be difficult to derive:
  - $\mathbf{X}$ are generated from a mixture of several models, however, we do not know which is the underlying true model for each observation (GMM belongs to this case).
  - $\mathbf{X}$ contains missing values, where we have $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$.
- However, it would be easier if we introduce a latent variable $\mathbf{Z}$, such that the joint likelihood of $\mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ is much easier to derive.

The EM algorithm

- For example, if $\mathbf{Z}$ represents the hidden label, then

$$\mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \mathbf{p}(\mathbf{z}|\boldsymbol{\theta})\mathbf{p}(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$$

  where both probabilities are easier to write out given the underlying model.

- In general, for a discrete case of $\mathbf{Z}$, we need to maximize the log-likelihood

$$\log \mathbf{p}(\mathbf{x}|\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} \mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$$

- For a continuous case, we maximize the log-likelihood

$$\log \mathbf{p}(\mathbf{x}|\boldsymbol{\theta}) = \log \int_{\mathbf{z}} \mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z}$$

The EM algorithm

- This can still be difficult to solve since there is a summation in the $\log$ function.
- The EM (Expectation–Maximization) algorithm is designed to solve this problem (Dempster, Laird, and Rubin, 1977)
- An EM algorithem consists of two steps:
  - E-step: Under the current value of $\boldsymbol{\theta}$, denoted as $\boldsymbol{\theta}^{(k)}$, find $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(k)})$, the distribution of the unobserved variables given the data and $\boldsymbol{\theta}^{(k)}$. Then calculate the conditional expectation:

$$
\begin{aligned}
g(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{x},\boldsymbol{\theta}^{(k)}} \log \mathbf{p}(\mathbf{x}, \mathbf{Z}|\boldsymbol{\theta}) \\
&= \begin{cases} \sum_{\mathbf{z}} \mathbf{p}(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(k)}) \log \mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) & \text{(discrete)} \\ \int_{\mathbf{z}} \mathbf{p}(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(k)}) \log \mathbf{p}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} & \text{(continuous)}. \end{cases}
\end{aligned}
$$

  - M-step: Re-estimate the parameter $\boldsymbol{\theta}$ to maximize $g(\boldsymbol{\theta})$:

$$
\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta})
$$

The EM algorithm

- Three components are involved in an EM algorithm:
  - The outcome variable $\mathbf{X}$, with observed value $\mathbf{x}$
  - The distribution parameters $\boldsymbol{\theta}$
  - The latent variables $\mathbf{Z}$
- The EM algorithm bounces back and forth between two processes:
  - E-step Given the current parameters and the observed data, estimate the (conditional mean of) latent variables
  - M-step Given the observed data and the latent variables, estimate the parameters
- We will see more examples using the Hidden Markov Models (HMM)

The MM algorithm

The MM algorithm

- The Majorize-Maximization (or Majorize-Minimization) algorithm is a general framework for optimization.
- The EM algorithm can be treated as a special case of MM.
- We will show that both MM and EM has the ascend (or descend) property.

General Philosophy

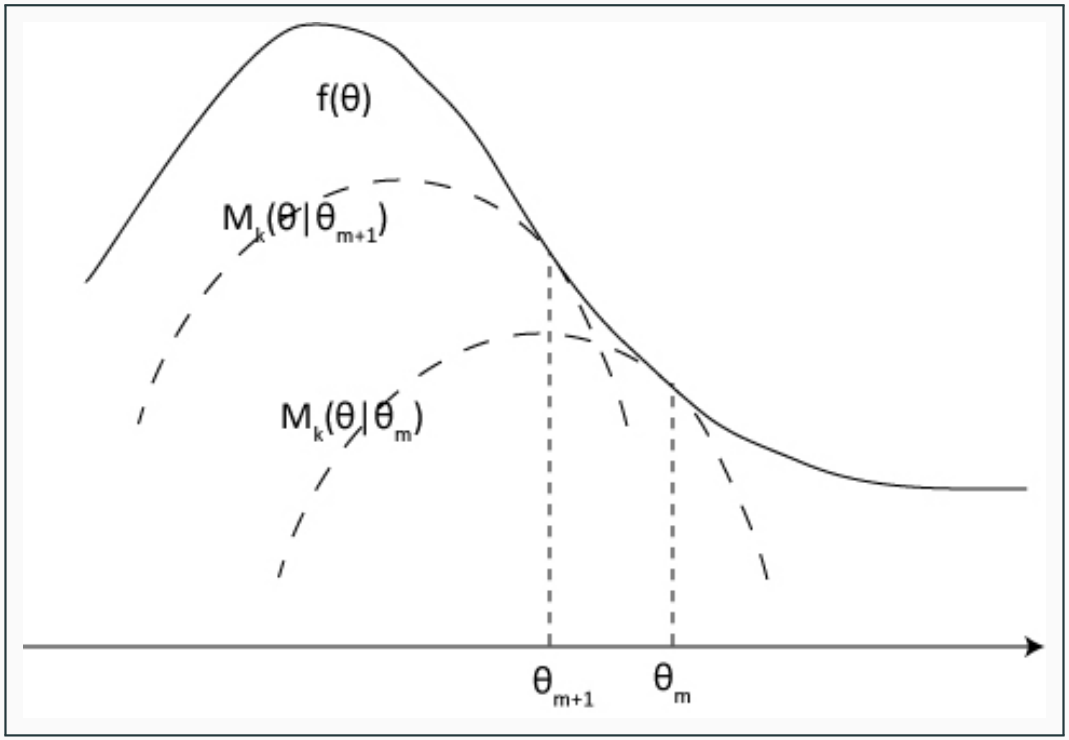
- The MM algorithm is relatively simpler than EM
- Suppose we want to maximize the function $f(\theta)$. We perform the optimization by recursively updating the θ value
  - At each iteration k, let $\theta^{(k)}$ represent the currently parameter value, we first find a function $g(\theta|\theta^{(k)})$ that "majorize $f(\theta)$":

$$\begin{aligned} \text{for all } \theta, \quad & g(\theta|\theta^{(k)}) \leq f(\theta) \\ \text{and} \quad & g(\theta^{(k)}|\theta^{(k)}) = f(\theta^{(k)}) \end{aligned}$$

  - Update θ with

$$\theta^{(k+1)} = \arg\max_{\theta} \; g(\theta|\theta^{(k)})$$

General Philosophy



from Wiki.

Ascend Property

- The MM algorithm is guaranteed to ascend:

$$\begin{aligned} f(\theta^{(k+1)}) &= g(\theta^{(k+1)}|\theta^{(k)}) \\ &\quad + \left[f(\theta^{(k+1)}) - g(\theta^{(k+1)}|\theta^{(k)}) \right] \\ &\geq g(\theta^{(k+1)}|\theta^{(k)}) \\ &\geq g(\theta^{(k)}|\theta^{(k)}) \\ &= f(\theta^{(k)}) \end{aligned}$$

- If we use $-f(\theta^{(k+1)})$, then its guaranteed to descend.

Ascend Property: EM

- The EM algorithm is a special case of MM
- We can show that

$$\ell(\theta^{(k+1)}|\mathbf{x}) = \log \mathbf{p}(\mathbf{x}|\theta^{(k+1)}) \geq \log \mathbf{p}(\mathbf{x}|\theta^{(k)}) = \ell(\theta^{(k)}|\mathbf{x})$$

- The proof relies on Jensen's inequality:

$$\begin{aligned} & \mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \quad \text{if } f \text{ is convex} \\ \text{or} \quad & \mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) \quad \text{if } f \text{ is concave} \end{aligned}$$

Ascend Property: EM

- Since $\mathbf{x}$ is observed and $\mathbf{z}$ is the missing part

$$\ell(\theta|\mathbf{x}) = \log \mathbf{p}(\mathbf{x}|\theta) = \log \frac{\mathbf{p}(\mathbf{x}, \mathbf{z}|\theta)}{\mathbf{p}(\mathbf{z}|\mathbf{x}, \theta)}$$

- The difference of two likelihood functions under $\theta^{(k)}$ and $\theta^{(k+1)}$

$$\begin{aligned}\ell(\theta^{(k+1)}|\mathbf{x}) - \ell(\theta^{(k)}|\mathbf{x}) = &\log \mathbf{p}(\mathbf{x}, \mathbf{z}|\theta^{(k+1)}) - \log \mathbf{p}(\mathbf{x}, \mathbf{z}|\theta^{(k)}) \\ &- \left\{ \log \mathbf{p}(\mathbf{z}|\mathbf{x}, \theta^{(k+1)}) - \log \mathbf{p}(\mathbf{z}|\mathbf{x}, \theta^{(k)}) \right\}\end{aligned}$$

Ascend Property: EM

- We take the expectation of $\mathbf{z}$, conditioning on $\mathbf{x}$ and $\theta^{(k)}$
- For the LHS, since $\theta^{(k+1)} = \arg\max_\theta g(\theta|\mathbf{x}, \theta^{(k)})$, the updated likelihood $\ell(\theta^{(k+1)}|\mathbf{x})$ can only depend on $\mathbf{x}$ and $\theta^{(k+1)}$. Hence, taking the conditional expectation does not change the LHS.
- For the RHS, we have

$$\mathbb{E}\left[\log \mathbf{p}(\mathbf{x}, \mathbf{z}|\theta^{(k+1)}) - \log \mathbf{p}(\mathbf{x}, \mathbf{z}|\theta^{(k)})\middle|\mathbf{x}, \theta^{(k)}\right]$$
$$-\mathbb{E}\left[\log \mathbf{p}(\mathbf{z}|\mathbf{x}, \theta^{(k+1)}) - \log \mathbf{p}(\mathbf{z}|\mathbf{x}, \theta^{(k)})\middle|\mathbf{x}, \theta^{(k)}\right]$$

- The first term is the difference $g(\theta^{(k+1)}|\mathbf{x}, \theta^{(k)}) - g(\theta^{(k)}|\mathbf{x}, \theta^{(k)})$, from the E step in the EM algorithm. Hence this is non-negative.

Ascend Property: EM

- The second term is

$$\mathbb{E}\left[\log\left\{\frac{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k+1)})}{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k)})}\right\}\middle|\mathbf{x},\theta^{(k)}\right]$$

- Since $\log$ is a concave function, we have

$$\mathbb{E}\left[\log\left\{\frac{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k+1)})}{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k)})}\right\}\middle|\mathbf{x},\theta^{(k)}\right] \leq \log\mathbb{E}\left[\frac{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k+1)})}{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k)})}\middle|\mathbf{x},\theta^{(k)}\right]$$

$$= \log\int\frac{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k+1)})}{\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k)})}\mathbf{p}(\mathbf{z}|\mathbf{x},\theta^{(k)})d\mathbf{z}$$

$$= \log(1) = 0$$

- Hence, overall, we have the ascend property

$$\ell(\theta^{(k+1)}|\mathbf{x}) - \ell(\theta^{(k)}|\mathbf{x}) \geq 0$$