STATS 230: Computational Statistics
Expectation Maximization (EM)

Babak Shahbaba

Department of Statistics, UCI

Outline

- In this lecture, we discuss Expectation-Maximization (EM), which is an iterative optimization method dealing with missing or latent data
- In such cases, given the observed data, x, and unobserved data, z, we assume that the hypothetical complete data would have been $y = (x, z)$
- Very often, the inclusion of the unobserved data z is a "data augmentation" strategy to make computation convenient
- That is, the original model involves observable variables X; we augment the data by assuming that there have been unobservable (latent) variables Z to simplify the computational problem
- Throughout this lecture, we will use finite mixture models as an illustrative example, where the mixture membership indicator, Z, is assumed to be a latent variable

Expectation-Maximization (EM)

- After data augmentation, instead of estimating model parameters, θ, based on the log-likelihood $\ell(\theta|x)$ given the observed data, we estimate θ based on the "complete" log-likelihood $\ell_c(\theta|y)$ and the assumed conditional distribution of latent variables given the observed data: $P(z|x, \theta)$

- At each iteration, the EM algorithm involves two steps: Expectation (E-step) and Maximization (M-step)

- The expectation step involves finding the function $Q(\theta)$ by integrating $\ell_c(\theta|y)$ over the conditional distribution of $z|x, \theta$

- Note that the conditional distribution depends on the unknown parameters so at each iteration, we use the previous value of θ to fully specify the conditional distribution

- The maximization steps at each iteration involves maximizing $Q(\theta)$ to update θ

Expectation-Maximization (EM)

- If we could observe the latent variables, we could write the *complete log-likelihood* based on a sample of iid data points as follows:

$$\ell_c(\theta) = \sum_i \log(P(x_i, z_i|\theta))$$

- The EM algorithm is an iterative procedure involving two steps:
  - E step: at iteration t, given the observed data and the previous value of θ, we find the expectation of ℓ_c,

$$\begin{aligned} Q(\theta) &= E[\ell_c(\theta)|x, \theta^{(t-1)}] \\ &= \int \ell_c(\theta) P(z|x, \theta^{(t-1)}) dz \end{aligned}$$

  - M step: we maximize Q with respect to θ to find $\theta^{(t)}$

$$\theta^{(t)} = \arg\max_\theta Q(\theta)$$

Convergence

- Because we have $f(z|x,\theta) = f(y|\theta)/f(x|\theta)$, we can write

$$\log f(x|\theta) = \log f(y|\theta) - \log f(z|x,\theta)$$

- Therefore,

$$E[\log f(x|\theta)|x,\theta^{(t-1)}] = E[\log f(y|\theta)|x,\theta^{(t-1)}] - E[\log f(z|x,\theta)|x,\theta^{(t-1)}]$$

  where the expectation is with respect to $z|x,\theta^{(t-1)}$

- Setting $H(\theta) = E[\log f(z|x,\theta)|x,\theta^{(t-1)}]$, we have

$$\log f(x|\theta) = Q(\theta) - H(\theta)$$

Convergence

- Using Jensen's inequality, we can show that $H(\theta)$ maximizes at $\theta^{(t-1)}$,

$$
\begin{aligned}
H(\theta^{(t-1)}) - H(\theta) &= E[\log f(z|x,\theta^{(t-1)}) - \log f(z|x,\theta)|x,\theta^{(t-1)}] \\
&= \int -\log[\frac{f(z|x,\theta)}{f(z|x,\theta^{(t-1)})}] f(z|x,\theta^{(t-1)}) dz \\
&\geq -\log \int f(z|x,\theta) dz \\
&= 0
\end{aligned}
$$

- Any $\theta \neq \theta^{(t-1)}$ makes $H(\theta)$ smaller than $H(\theta^{(t-1)})$ especially if we choose the optimum value $\theta^{(t)}$; then we have

$$\log f(x|\theta^{(t)}) - \log f(x|\theta^{(t-1)}) \geq 0$$

since at $\theta^{(t)}$, we increase Q and reduce H, with strict inequality when $Q(\theta^{(t)}) > Q(\theta^{(t-1)})$

- See Givens and Hoeting (2013) for more details

Monte Carlo EM (MCEM)

- In some cases, finding the expectation in the E step analytically might be difficult
- For such problems, Wei and Tanner (1990) proposed to use Monte Carlo approximation instead
  - Simulate datasets $\mathbf{z}_1, \ldots, \mathbf{z}_m \sim P(\mathbf{z}|\mathbf{x}, \theta^{(t-1)})$
  - Calculate $\hat{Q}(\theta) = \frac{1}{m}\sum_{j=1}^{m} \ell_c(\theta|\mathbf{y}_j)$, where each $\mathbf{y}_j$ represent a complete dataset $(\mathbf{x}, \mathbf{z}_j)$
  - Maximize $\hat{Q}(\theta)$ to update θ

Illustrative example: finite mixture models

- As an illustrative example, we consider clustering of data using a "finite" mixture of Gaussians
- This is also known as "soft K-means" clustering
- A simple version of this method assumes that all Gaussians have the same covariance matrix that is diagonal
- We can extend this method by allowing for different different covariance matrices, but still keeping the Gaussians *spherical*
- Finally, we can generalize this method by allowing non-spherical Gaussians in the mixture
- See Murphy (2012) for more details

Finite mixture models

- We present a mixture of K base distributions, $P_1, \ldots, P_K$, as follows:

$$P(x_i|\pi, \theta) = \sum_{k=1}^{K} \pi_k P_k(x_i|\theta_k)$$

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

- We mainly focus on mixture of Gaussians,

$$P(x_i|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

Finite mixture models

- We could of course maximize the log-likelihood function,

$$\sum_i \log(\sum_k \pi_k N(x_i | \mu_k, \Sigma_k))$$

subject to the constraints that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

- However, in such cases, it is easier to use the data augmentation approach by introducing latent indicators and estimate the parameters using the expectation-maximization algorithm

Latent indicators

- First, we assume that there are latent (hidden, unobserved) variables, z_i, which assign the observation i to one of the component of the mixture.

- That is, if $z_i = k$, the i^{th} observation has the P_k distribution,

$$x_i | z_i = k \sim N(\mu_k, \Sigma_k)$$

- The log-likelihood for this hypothetical "complete" data is

$$\ell_c(\theta) = \sum_i \log P(x_i, z_i | \theta)$$

- As mentioned above, we also need the conditional distribution of z given x; using Bayes' theorem, we have

$$P(z_i = k | x_i, \theta) = \frac{\pi_k P(x_i | \theta_k)}{\sum_{k'=1}^{K} \pi_{k'} P(x_i | \theta_{k'})}$$

E step

- In the E step, we have (Murphy, 2012)

$$
\begin{aligned}
Q(\theta) &= E[\sum_i \log P(x_i, z_i|\theta)] \\
&= \sum_i E[\log(\prod_k (\pi_k P(x_i|\theta))^{I(z_i=k)})] \\
&= \sum_i \sum_k E[I(z_i = k)] \log(\pi_k P(x_i|\theta)) \\
&= \sum_i \sum_k p_{ik} \log \pi_k + \sum_i \sum_k p_{ik} \log(P(x_i|\theta))
\end{aligned}
$$

where $p_{ik} = E[I(z_i = k)] = P(z_i = k|x_i, \theta^{(t-1)})$ is called *responsibility* and is specified based on the previous value of model parameters,

$$
p_{ik} = \frac{\pi_k^{(t-1)} P(x_i|\theta_k^{(t-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t-1)} P(x_i|\theta_{k'}^{(t-1)})}
$$

M step

- In the M step, we maximize $Q(\theta)$ with respect to θ
- In this example, $\theta = (\pi, \mu, \Sigma)$
  - For π

$$\pi_k^{(t)} = \frac{1}{n} \sum_i p_{ik}$$

  - for μ_k and Σ_k

$$\begin{aligned} \mu_k^{(t)} &= \frac{\sum_i p_{ik} x_i}{\sum_i p_{ik}} \\ \Sigma_k^{(t)} &= \frac{\sum_i p_{ik}(x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})^\top}{\sum_i p_{ik}} \end{aligned}$$

which are the weighted version of MLE for a single Gaussian distribution.

EM for a mixture of two univariate Gaussians

- For the special case where the distribution is a mixture of two Gaussians, we have

$$\pi_1 = 1 - \pi, \quad \pi_2 = \pi$$

- We specify z as follows:

$$x_i \sim \begin{cases} N(\mu_1, \sigma_1^2) & \text{if } z_i = 0, \\ N(\mu_2, \sigma_2^2) & \text{if } z_i = 1. \end{cases}$$

- Then,

$$\begin{aligned} \ell_c(\theta) &= \sum_i [(1 - z_i) \log(N(x_i | \mu_1, \sigma_1^2)) + z_i \log(N(x_i | \mu_2, \sigma_2^2))] \\ &+ \sum_i [(1 - z_i) \log(1 - \pi) + z_i \log(\pi)] \end{aligned}$$

Mixture of two univariate Gaussians– the E step

- We start with an initial guess for model parameters: $\theta^{(0)}$ for model parameters $\theta = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$

- At iteration t, we have

$$p_i = E(z_i|x, \theta^{(t-1)}) \quad = \quad P(z_i = 1|x, \theta^{(t-1)}) =$$

$$\frac{\pi^{(t-1)} N(x_i|\mu_2^{(t-1)}, [\sigma_2^2]^{(t-1)})}{(1-\pi^{(t-1)}) N(x_i|\mu_1^{(t-1)}, [\sigma_1^2]^{(t-1)}) + \pi^{(t-1)} N(x_i|\mu_2^{(t-1)}, [\sigma_2^2]^{(t-1)})}$$

Mixture of two univariate Gaussians– the M step

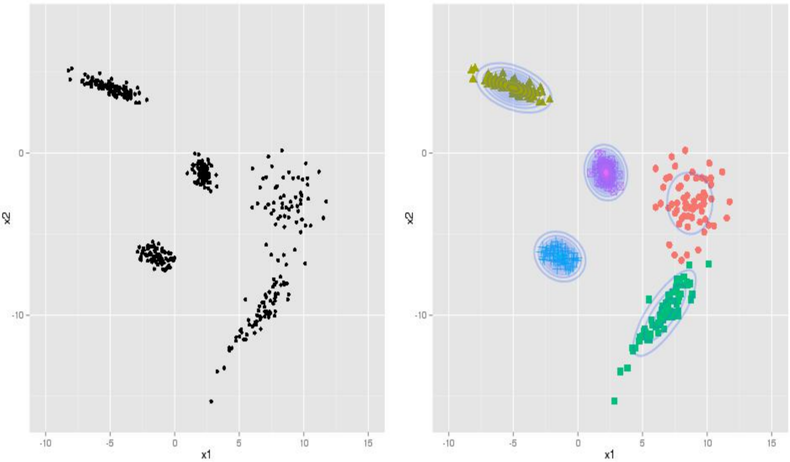
- Now, given the values of p_i, we obtain $Q(\theta)$ and maximize it with respect to θ to obtain a new estimates $\theta^{(t)}$
- In this case, the new estimates (which are simply the weighted mean and variance) are as follows:

$$\pi^{(t)} = \frac{\sum_i p_i}{n}$$

$$\mu_1^{(t)} = \frac{\sum_i (1 - p_i) x_i}{\sum_i (1 - p_i)}, \qquad \mu_2^{(t)} = \frac{\sum_i p_i x_i}{\sum_i p_i}$$

$$[\sigma_1^2]^{(t)} = \frac{\sum_i (1 - p_i)(x_i - \mu_1^{(t)})^2}{\sum_i (1 - p_i)}, \qquad [\sigma_2^2]^{(t)} = \frac{\sum_i p_i (x_i - \mu_2^{(t)})^2}{\sum_i p_i}$$

Example: Mixture of 5 Gaussians



Example: Mixture of 3 Gaussians