ISyE 6416: Computational Statistics
Spring 2017

Lecture 7: EM algorithm and Gaussian Mixture Model

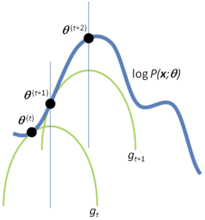
Prof. Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Expectation-Maximization (EM) Algorithm

- an algorithm to a maximum likelihood estimator in **non-ideal** case: missing data, indirect observations
  - missing data
  - clustering (unknown label)
  - hidden-states in HMM
  - latent factors

- replace one difficult likelihood maximization with a sequence of easier maximizations
- in the limit, the answer to the original problem

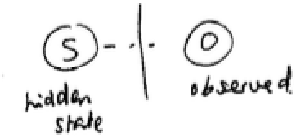


Supplementary Figure 1 Convergence of the EM algorithm. Starting from initial parameters $\theta^{(t)}$, the E-step of the EM algorithm constructs a function g_t that lower-bounds the objective function $\log P(x;\theta)$. In the M-step, $\theta^{(t+1)}$ is computed as the maximum of g_t. In the next E-step, a new lower-bound g_{t+1} is constructed; maximization of g_{t+1} in the next M-step gives $\theta^{(t+2)}$, etc.

Applications of EM

- Data clustering in machine learning
- Natural language processing (Baum-Welch algorithm to fit hidden Markov model)
- Imputing missing data

General set-up



- need the distribution of S (or sometimes jointly S and O), but we only observe directly through O

$$S \sim f(S|\theta)$$

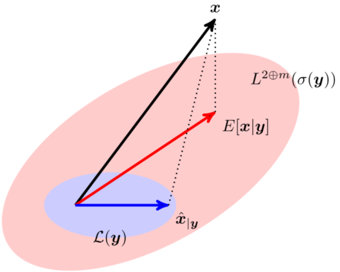
  parameter θ
- Suppose mapping $\chi(o)$ finds all values of S that leads to O
- distribution of O: $g(O|\theta) = \int_{\chi(o)} f(S|\theta) ds$
- EM: $\max_\theta g(O|\theta)$

Deriving EM

- Introduce

$$Q(\theta; \theta') = \mathbb{E}\{\log f(S|\theta)|\theta', O\}$$

- The expectation uses the conditional distribution of S given O and assumed value of parameter θ'



Rhymer's Notes

Intuition

Given O, the "best guess" we could have for S, is its conditional expectation with respect to $S|O, \theta$ (notion of projection); but the computation of expectation involves parameter values. We take a guess, and improve in next round.

E-M algorithm

- **E-step**: compute expectation of the log-likelihood (observed) incomplete data $\mathbf{y}$, missing data $\mathbf{x}$, complete data $(\mathbf{y}, \mathbf{x})$
$$Q(\theta; \theta') = \mathbb{E}\{\log L(\theta|\mathbf{y}, \mathbf{x})|\theta', \mathbf{y}\}$$
- **M-step**: compute maximum likelihood using the expectation in previous step

$$\text{E-step} \Rightarrow \text{M-step} \Rightarrow \text{E-step} \Rightarrow \text{M-step} \Rightarrow$$

- stop until $\|\theta_{k+1} - \theta_k\| < \epsilon$ or $|Q(\theta_{k+1}|\theta_k) - Q(\theta_k|\theta_{k-1})| < \epsilon$

Algorithm 1 (Expectation-Maximization)

<u>**begin initialize**</u> $\boldsymbol{\theta}^0, T, i = 0$

  <u>**do**</u> $i \leftarrow i + 1$

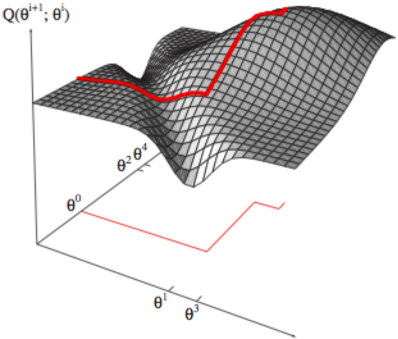
    E step : compute $Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$

    M step : $\boldsymbol{\theta}^{i+1} \leftarrow \arg\max_{\theta} Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i)$

  <u>**until**</u> $Q(\boldsymbol{\theta}^{i+1};\ \boldsymbol{\theta}^i) - Q(\boldsymbol{\theta}^i;\ \boldsymbol{\theta}^{i-1}) \leq T$

 <u>**return**</u> $\hat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{i+1}$

<u>**end**</u>



Example: EM for missing data

$n = 4$, $p = 2$

$$x_1 = (0,2)^T, x_2 = (1,0)^T, x_3 = (2,2)^T, x_4 = (*,4)^T$$

Assume they are i.i.d. samples from Gaussian

$$\mathcal{N}([\mu_1, \mu_2]^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix})$$

Use EM algorithm to impute the missing data *.

Pattern classification, R. O. Duda, P. E. Hart, and D. G. Stork

(Cont.) Example: missing data

- Initialization: $\theta_0 = (0, 0, 1, 1)^T$, i.e., mean $[0, 0]^T$ and covariance I_2.
- E-step

$$\begin{aligned}
Q(\theta|\theta_0) &= \mathbb{E}_{x_{41}}[\log p(x|\theta_0, x_1, x_2, x_3, x_{41}, x_{42})|x_1, x_2, x_3, x_{42}] \\
&= \sum_{i=1}^{3} \log p(x_i|\theta) \\
&+ \int \log(p([x_{41}, 4]^T)|\theta) \cdot p([x_{41}, 4])|\theta_0) dx_{41} \\
&= \sum_{i=1}^{3} \log p(x_i|\theta) - \frac{(1+\mu_1^2)}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2} - \log(2\pi\sigma_1\sigma_2)
\end{aligned}$$

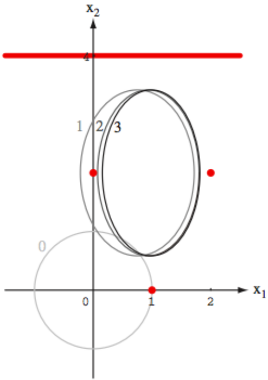
- M-step

$$\theta_1 = \arg\max_\theta Q(\theta|\theta_0)$$

(Cont.) Example: missing data - iterations

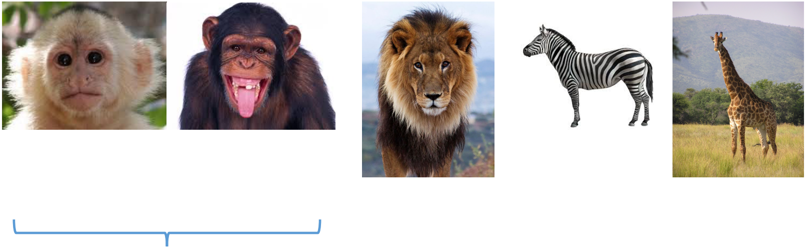
$$\theta_1 = \begin{bmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{bmatrix} \Rightarrow \quad \mu_1 = \begin{bmatrix} 0.75 \\ 2.0 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 0.938 & 0 \\ 0 & 2.0 \end{bmatrix}$$

$$\theta_2 = \begin{bmatrix} 1.0 \\ 2.0 \\ 0.667 \\ 2.0 \end{bmatrix}$$



The absent-minded biologist

197 animals
Distributed into 4 categories



125 18 20 34

Multinomial model of 5 category with parameter θ

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$$

Can we figure out the number of Monkey A based on the data?

(Cont.) The absent-minded biologist

- data $y = (125, 18, 20, 34)$
- now assume $y_1 = y_{11} + y_{12} = 125$
- Likelihood function

$$f(y|\theta) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}(\frac{1}{2})^{y_{11}}(\frac{\theta}{4})^{y_{12}}(\frac{1}{4}-\frac{\theta}{4})^{y_2}(\frac{1}{4}-\frac{\theta}{4})^{y_3}(\frac{\theta}{4})^{y_4}$$

- log-likelihood

$$\ell(\theta|y) \propto ({\color{red}y_{12}} + y_4)\log\theta + (y_2 + y_3)\log(1-\theta)$$

- y_{12} unknown, cannot directly maximize $\ell(\theta|y)$

(Cont.) The absent-minded biologist: set-up EM

$$
\begin{aligned}
Q(\theta|\theta') &= \mathbb{E}_{y_{12}}[(y_{12}+y_4)\log\theta + (y_2+y_3)\log(1-\theta)|y_1,\ldots,y_4,\theta'] \\
&= (\mathbb{E}_{y_{12}}[y_{12}|y_1,\theta'] + y_4)\log\theta + (y_2+y_3)\log(1-\theta)
\end{aligned}
$$

Conditional distribution of y_{12} given y_1: Binomial $(y_1, \frac{\theta'/4}{\theta'/4+1/2})$

$$
\mathbb{E}_{y_{12}}[y_{12}|y_1,\theta'] = \frac{y_1\theta'}{2+\theta'} := y_{12}^{\theta'},
$$

E-step:

$$
Q(\theta|\theta') = (y_{12}^{\theta'} + y_4)\log\theta + (y_2+y_3)\log(1-\theta)
$$

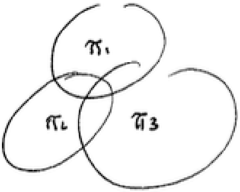
M-step: $\theta_{k+1} = \arg\max Q(\theta|\theta_k) = \frac{y_{12}^{(\theta_k)}+y_4}{y_{12}^{(\theta_k)}+y_2+y_3+y_4}$

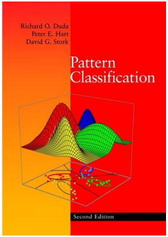
Fitting Gaussian mixture model (GMM)

$$x_i \sim \sum_{i=1}^{C} \pi_i \phi(x_i | \mu_i, \Sigma_i)$$

ϕ: density of multi-variate normal

- parameters $\{\mu_i, \Sigma_i, \pi_i, C\}$
- assume C is known.
- observed data $\{x_1, \ldots, x_n\}$
- complete data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
  y_n: "label" for each sample, missing.





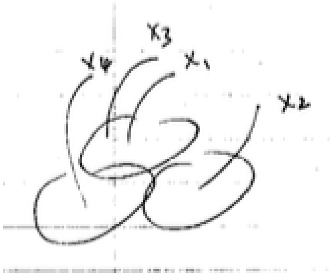
EM for GMM

- If we know the label information y_i, likelihood function can be easily written

$$\pi_{y_i}\phi(x_i|\mu_{y_i},\Sigma_{y_i})$$

- now y_i unknown, compute its expectation with respect to the set of parameters

$$Q(\theta|\theta') = \mathbb{E}[\sum_{i=1}^{n} \log \pi_{y_i} + \log \phi(x_i|\mu_{y_i},\Sigma_{y_i})|x_i,\theta']$$



E-step

- $(\pi_i^{(k)}, \mu_i^{(k)}, \Sigma_i^{(k)})$ parameter values in the kth iteration
- we need $y_i|x_i$, posterior distribution of label, given observation x_i

$$p_{i,c} := p(y_i = c|x_i) \propto \pi_c^{(k)} \phi(x_i|\mu_i^{(k)}, \Sigma_i^{(k)})$$

and $\sum_{c=1}^{C} p(y_i = c|x_i) = 1$

$$\begin{aligned} Q(\theta|\theta_k) &= \sum_{i=1}^{n} \mathbb{E}[\log \pi_{y_i} + \log \phi(x_i|\mu_{y_i}, \Sigma_{y_i}|x_i, \theta_k] \\ &= \sum_{i=1}^{n} \sum_{c=1}^{C} {\color{blue}p_{i,c}} \log {\color{red}\pi_c} + \sum_{i=1}^{n} \sum_{c=1}^{C} {\color{blue}p_{i,c}} \log \phi(x_i|{\color{red}\mu_c}, {\color{red}\Sigma_c}) \end{aligned}$$

Q: where is θ?

M-step

- Maximize $Q(\theta|\theta_k)$ with respect to π_c, μ_c, Σ_c (note that they can be maximized separately)

$$\theta_{k+1} = \arg\max_{\theta} Q(\theta|\theta_k)$$

- note that $\sum_{c=1}^{C} \pi_c = 1$

$$\vdots$$

in the end

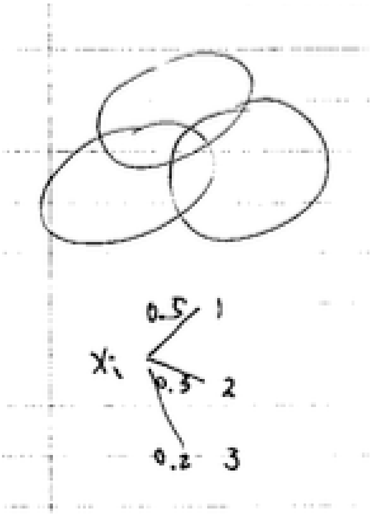
$$\mu_c^{(k+1)} = \frac{\sum_{i=1}^{n} p_{i,c} x_i}{\sum_{i=1}^{n} p_{i,c}}$$

$$\Sigma_c^{(k+1)} = \frac{\sum_{i=1}^{n} p_{i,c}(x_i - \mu_c^{(k+1)})(x_i - \mu_c^{(k+1)})^T}{\sum_{i=1}^{n} p_{i,c}}$$

$$\pi_c^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{i,c}$$

Interpretation

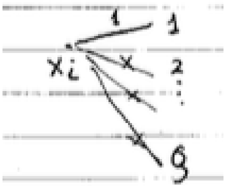
- $p_{i,c}$: probability of each sample belong to computer c
- $\pi_c^{(k+1)}$: count the expected number of samples belong to component c
- soft-assignment: x_i belong to component c with assignment probability $p_{i,c}$
- $\mu_c^{(k+1)}$: "average" centroid using soft assignment
- $\mu_c^{(k+1)}$: "average" covariance using soft assignment

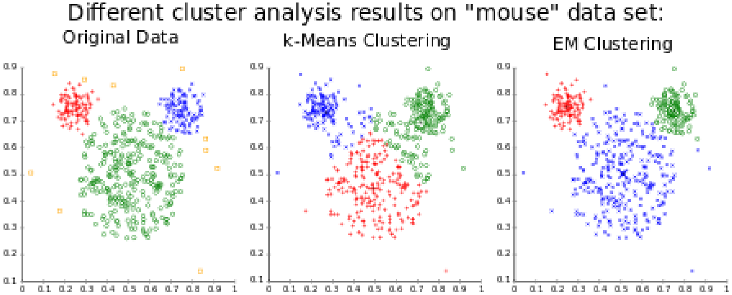


k-means

- "hard" assignment
- EM algorithm can also be used for clustering: in the end, $p_{i,c}$ can be viewed as a soft label for each sample
- convert into hard label:

$$\hat{c}_i = \arg\max_{c=1}^{C} p_{i,c}$$





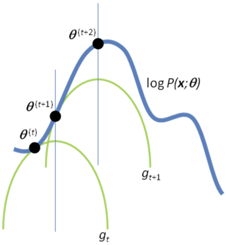
Properties of EM

- EM algorithm converges to local maximum
- Heuristic: escaping the local maximum through a random start
- EM works on improving $Q(\theta|\theta')$ rather than directly improving $\log f(x|\theta)$
- one can show that improvement on $Q(\theta|\theta')$ improves $\log f(x|\theta)$
- EM works well with exponential family
  - E-step: sum of expectations of the sufficient statistics
  - M-step: maximizing a linear function

  usually possible to derive closed-form update

Convergence of EM

- Proof by A. Dempster, N. Larid and D. Rubin in 1977, later generalized by J. Wu in 1983.
- Basic idea: find a sequence of quadratic lower bounds for the likelihood function
- EM monotonically increases the observed data log likelihood



Supplementary Figure 1 Convergence of the EM algorithm. Starting from initial parameters $\theta^{(t)}$, the E-step of the EM algorithm constructs a function g_t that lower-bounds the objective function $\log P(x;\theta)$. In the M-step, $\theta^{(t+1)}$ is computed as the maximum of g_t. In the next E-step, a new lower-bound g_{t+1} is constructed; maximization of g_{t+1} in the next M-step gives $\theta^{(t+2)}$, etc.

$$\ell(\theta_{k+1}) \geq Q(\theta_{k+1};\theta_k) \geq Q(\theta_k;\theta_k) = \ell(\theta_k)$$