Machine Learning EM algorithm and Multidimensional Scaling

Lecture Notes 7: EM algorithm and Multidimensional Scaling

Professor: Zhihua Zhang *Scribe: Yao Cheng, Yuchen Sha*

1 Expectation-Maximization(EM algorithm)

1.1 Convergence of EM algorithm

Suppose $\theta^{(t)}$ and $\theta^{(t+1)}$ are the parameters from two successive iterations of EM. We will now prove that $L(\theta^{(t)}) \leq L(\theta^{(t+1)})$, that is, $log\, p(x|\theta^{(t)}) \leq log\, p(x|\theta^{(t+1)})$.
We know

$$Q(\theta|\theta^{(t)}) = \textstyle\int (log\, p(x,z|\theta))(p(z|x,\theta^{(t)}))dz.$$

Let $H(\theta|\theta^{(t)}) = log\, p(x|\theta) - Q(\theta|\theta^{(t)})$, then

$$log\, p(x|\theta^{(t+1)}) - log\, p(x|\theta^{(t)}) = Q(\theta|\theta^{(t+1)}) + H(\theta|\theta^{(t+1)}) - (Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)}))$$

Since

$$Q(\theta|\theta^{(t+1)}) = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

We have

$$Q(\theta|\theta^{(t+1)}) \geq Q(\theta|\theta^{(t)})$$

Now we prove

$$H(\theta|\theta^{(t+1)}) \geq H(\theta|\theta^{(t)})$$

Firstly,

$$\begin{aligned}
H(\theta|\theta^{(t)}) &= log\, p(x|\theta) - Q(\theta|\theta^{(t)}) \\
&= log\, p(x|\theta)\int p(z|x,\theta^{(t)})dz - \int (log\, p(x,z|\theta))(p(z|x,\theta^{(t)}))dz \quad (p(z|x,\theta^{(t)})\ is\ a\ distribution\ of\ z) \\
&= \int log\, p(x|\theta)p(z|x,\theta^{(t)})dz - \int (log\, p(x,z|\theta))(p(z|x,\theta^{(t)}))dz \\
&= \int log\, \frac{p(x|\theta)}{p(x,z|\theta)} p(z|x,\theta(t))dz \\
&= -\int log\, p(z|x,\theta)p(z|x,\theta^{(t)})dz
\end{aligned}$$

Thus,

$$\begin{aligned}
H(\theta|\theta^{(t+1)}) - H(\theta|\theta^{(t)}) &= -\int log\, p(z|x,\theta^{(t+1)})p(z|x,\theta^{(t)})dz + \int log\, p(z|x,\theta^{(t)})p(z|x,\theta^{(t)}) \\
&= \int log\, \frac{p(z|x,\theta(t))}{p(z|x,\theta(t+1))} p(z|x,\theta t)dz \\
&= -\int log\, \frac{p(z|x,\theta(t+1))}{p(z|x,\theta(t))} p(z|x,\theta t)dz \\
&\geq -log \int \frac{p(z|x,\theta(t+1))}{p(z|x,\theta(t))} p(z|x,\theta t)dz \\
&= 0.
\end{aligned}$$

Hence,

$$log\, p(x|\theta^{(t+1)}) \geq log\, p(x|\theta^{(t)})$$

EM causes the likelihood to converge monotonically.

1.2 Equivalence of EM algorithm and MLE

To prove the correctness of EM algorithm, we now show the equivalence of EM algorithm and MLE when solving probabilistic PCA.
Based on the previous notes,we have the following results. Please refer to *Lecture Notes 6* for detailed proof if necessary.
1. The estimation of W and τ derived by maximum likelihood estimation(MLE) is:

$$\tau = \frac{1}{p-q} \sum_{j=q+1}^{p} \mathit{\Gamma}_j \tag{1}$$

$$W = \mathit{\Phi}_q(\mathit{\Gamma}_q - \tau I_q)^{\frac{1}{2}} V^T \tag{2}$$

2.The iterative formulas for W and τ derived by EM algorithm is:

$$W^{(t+1)} = (\sum_{i=1}^{n}(x_i - \mu) < z_i^T >)(\sum_{i=1}^{n} < z_i z_i^T >)^{-1} \tag{3}$$

$$\tau^{(t+1)} = \frac{1}{p}[tr(S) - \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^T W^{(t+1)} < z_i >] \tag{4}$$

where

$$< z_i >= M_{(t)}^{-1} W_{(t)}^T (x_i - \mu) \tag{5}$$

$$\begin{aligned}
< z_i z_i^T > &= \tau_{(t)} M_{(t)}^{-1} + < z_i >< z_i >^T \\
&= \tau_{(t)}^2 M_{(t)}^{-1} + M_{(t)}^{-1} W_{(t)}^T (x_i - \mu)(x_i - \mu)^T W(t) M_{(t)}^{-1}
\end{aligned} \tag{6}$$

Suppose W converges to Ŵ, τ converges to $\hat{\tau}$. Substitute $< z_i >$ and $< z_i z_i^T >$ with formula (5)(6) and we get

$$\hat{W} = S\hat{W}(\tau^2 I_q + M^{-1}\hat{W}^T S\hat{W})^{-1} \tag{7}$$

$$\hat{\tau} = \frac{1}{p}(tr(S) - S\hat{W}M^{-1}\hat{W}^T) \tag{8}$$

Here we just omit the hat on W and τ. Then according to equation(7),

$$\begin{gathered}
W(\tau I_q + M^{-1}W^T SW) = SW \\
\tau W + WM^{-1}W^T SW = SW \\
WM^{-1}W^T SW = (S - \tau I_p)W \\
W(\tau I_q + W^T W)^{-1}W^T SW = (S - \tau I_p)W \\
(\tau I_p + WW^T)^{-1}WW^T SW = (S - \tau I_p)W \\
(\tau I_p + WW^T)(S - \tau I_p)W = WW^T SW \\
(\tau S - \tau^2 I_p + WW^T S - \tau WW^T)W = WW^T SW \\
\tau SW - \tau^2 W + WW^T SW - \tau WW^T W = WW^T SW \\
\tau SW = \tau WW^T W + \tau^2 W \\
SW = WW^T W + \tau W \\
SW = W(W^T W + \tau I_q) \\
SW = W(V\Lambda V^T + \tau VV^T) \qquad (Note: WW^T = V\Lambda V^T) \\
SW = WV(\Lambda + \tau I_q)V^T \\
SWV = WV(\Lambda + \tau I_q) \\
SWV\Lambda^{-\frac{1}{2}} = WV\Lambda^{-\frac{1}{2}}(\Lambda + \tau I_q)
\end{gathered}$$

Moreover,

$$\Lambda^{-\frac{1}{2}}V^T W^T WV\Lambda^{-\frac{1}{2}} = I$$

Thus, $\Lambda + \tau I_q$ is composed of eigenvalues of S and $WV\Lambda^{-\frac{1}{2}}$ is composed of a set of orthonormal eigenvectors.
Let $\Phi_q = WV\Lambda^{-\frac{1}{2}}, \Gamma = \Lambda + \tau I_q$. Together we have

$$\begin{aligned}
S\Phi &= \Phi\Gamma_q \\
W &= \Phi(\Gamma_q - \tau I)^{\frac{1}{2}}V^T
\end{aligned}$$

τ can be derived similarly. It can be seen that the formulas are the same as those derived using maximum likelihood estimation.

2 Multidimensional Scaling

Definition 2.1 *A distance matrix D is called Euclidean if there exists a configuration of points in some Euclidean space where interpoint distances are given by D, that is, if for some p, there exist points* $x_1, x_2, ...x_n \in IR^p$ *such that* $d_{rs}^2 = ||x_r - x_s||^2 = (x_r - x_s)^T(x_r - x_s)$.

Theorem 2.1 *Let D be a distance matrix and define B=HAH, where* $A = [a_{rs}](a_{rs} = -\frac{1}{2}d_{rs}^2)$*.Then D is Euclidean iff B is p.s.d.*

Proof:
(a) Firstly, we prove that if D is a distance matrix, then B = HAH as defined is *p.s.d.* Suppose D is the matrix of Euclidean interpoint distances for a configuration $Z = (z_1, z_2, ...z_n)^T$.

Then $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$.

$$\begin{aligned} B &= HAH \\ &= (I_n - \frac{1}{n}I_nI_n^T)A(I_n - \frac{1}{n}I_nI_n^T) \\ &= A - \frac{1}{n}I_nI_n^TA - \frac{1}{n}AI_nI_n^T + \frac{1}{n^2}I_nI_n^TAI_nI_n^T \end{aligned}$$

Hence we have,

$$b_{rs} = a_{rs} - \bar{a}_{.s} - \bar{a}_{r.} + \bar{a}$$

Here $\bar{a}_{.s}$ represents the mean of the sth column of A, $\bar{a}_{r.}$ represents the mean of the rth row of A, $\bar{a}$ represents the mean of A.
Thus,

$$b_{rs} = -\frac{1}{2}||z_r - z_s||^2 + \frac{1}{2n}\sum_{i=1}^{n} d_{ri}^2 + \frac{1}{2n}\sum_{i=1}^{n} d_{is}^2 - \frac{1}{2n^2}\sum_{i,j} d_{i,j}^2$$

We calculate the formula part by part:
(1)

$$-\frac{1}{2}||z_r - z_s||^2 = -\frac{1}{2}z_r^Tz_r - \frac{1}{2}z_sz_s^T + z_r^Tz_s;$$

(2)

$$\frac{1}{2n}\sum_{i=1}^{n} d_{ri}^2 = \frac{1}{2n}\sum_{i=1}^{n}(z_r - z_i)^T(z_r - z_i) = \frac{1}{2}z_r^Tz_r + \frac{1}{2n}\sum_{i}^{n} z_i^Tz_i - z_r^T\bar{z};$$

(3)

$$\frac{1}{2n}\sum_{i=1}^{n} d_{is}^2 = \frac{1}{2n}\sum_{i=1}^{n}(z_i - z_s)^T(z_i - z_s) = \frac{1}{2}z_s^Tz_s + \frac{1}{2n}\sum_{i}^{n} z_i^Tz_i - z_s^T\bar{z};$$

(4)

$$\begin{aligned} \frac{1}{2n^2}\sum_{i,j} d_{i,j}^2 &= \frac{1}{2n^2}\sum_{i,j}^{n}(z_i^Tz_i + z_j^Tz_j - 2z_iz_j) \\ &= \frac{1}{2n}\sum_{i=1}^{n} z_i^Tz_i + \frac{1}{2n}\sum_{j=1}^{n}(z_j^Tz_j) - (\frac{1}{n}\sum_{i=1}^{n} z_i)^T(\frac{1}{n}\sum_{j=1}^{n} z_j) \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i^Tz_i - \bar{z}^T\bar{z}; \end{aligned}$$

Hence,

$$\begin{aligned} b_{rs} &= \bar{z}^Tz + \bar{z}_s^Tz_r - \bar{z}_r^T\bar{z} - z_s^T\bar{z} \\ &= z_s^T(z_r - \bar{z}) + \bar{z}^T(\bar{z} - z_r) \\ &= (z_r - \bar{z})^T(z_s - \bar{z}); \end{aligned}$$

Let $f(z) = z - \bar{z}$, then $b_{rs} = f(z_r)^Tf(z_s)$. Therefore, B is *p.s.d.*
(b) Now we prove the opposite direction. That is, if B is *p.s.d.* of rank p, then a configuration corresponds to B can be constructed as follows:

Let $\lambda_1 > \lambda_2 > ... > \lambda_p$ denote the positive eigenvalues of B with corresponding eigenvectors $Z = (z_{(1)}, z_{(2)}, ...z_{(p)})$ normalized by $z_{(i)}^T z_{(i)} = \lambda_i (i = 1, 2...p)$, then points p_r in IR^p with coordinates $z_r = (z_{(r_1)}), z_{(r_2)}, ...z_{(r_p)})^T$ have interpoint distance given by D. Further, this configuration has center of $\bar{z} = 0$.
Now we show the construction is correct.
Let $\Lambda = diag(\lambda_1, ...\lambda_p), U = [U(1), U(2), ...U(p)] where U(i) = z_{(i)}\lambda_i^{-\frac{1}{2}}$.
Thus,

$$U = Z\Lambda^{-\frac{1}{2}}$$

$$U^T U = \Lambda^{-\frac{1}{2}} Z Z^T \Lambda^{-\frac{1}{2}} = I_p$$

We know that U is composed of a set of orthonormal eigenvectors with U(i) corresponds to λ_i. The diagonal of Λ is composed of eigenvalues of B.
Hence, the spectral decomposition of B is $B = U\Lambda U^T = U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T = ZZ^T$, which means $b_{rs} = z_r^T z_s$.
According to the construction of D, if D is the Euclidean distance matrix of $p_1, p_2, ...p_n$,

$$D = (d_{rs}^2)$$

where $d_{rs}^2 = (z_r - z_s)^T (z_r - z_s) = z_r^T z_r - 2z_r^T z_s + z_s^T z_s = b_{rr} - 2b_{rs} + b_{ss}$.
According to (a),

$$b_{rs} = a_{rs} - \bar{a}_{.s} - \bar{a}_{r.} + \bar{a}$$

we have,

$$\begin{aligned}
d_{rs}^2 &= b_{rr} - 2b_{rs} + b_{ss} \\
&= a_{rr} - \bar{a}_{r.} - \bar{a}_{.r} + \bar{a} + a_{ss} - \bar{a}_{s.} - \bar{a}_{.s} + \bar{a} - 2a_{rs} + 2\bar{a}_{r.} + 2\bar{a}_{.s} - 2\bar{a} \\
&= \bar{a}_{r.} + \bar{a}_{.s} - \bar{a}_{.r} - \bar{a}_{s.} - 2a_{rs} \\
&= -2a_{rs}
\end{aligned}$$

As we can see, the result conforms to the definition of A, which means D is really the Euclidean distance matrix of $p_1, p_2, ...p_n$.
Remarks: Here each row of Z corresponds to a point. Moreover, since $H\mathbf{1}_n = \mathbf{0} = 0 \cdot \mathbf{1}_n$, we have $B\mathbf{1}_n = HAH\mathbf{1}_n = 0 \cdot \mathbf{1}_n$, which means $\mathbf{1}_n$ is an eigenvector of B corresponds to eigenvalue 0. Since eigenvectors corresponds to different eigenvalues are orthogonal, $Z^T\mathbf{1}_n = \mathbf{0}$, which is equivalent to $\sum_{r=1}^n z_r = 0$. Therefore, the mean of all points passes the origin, which guarantees the uniqueness of Z.

Add a point:
Now we consider when a new point comes, how to find the coordinates of that point. We formalize the problem as follows:
Given a distance matrix $D_{n\times n}$ of n points $p_1, p_2, ..p_n$. The distance between a new point p_{n+1} with the n points is $(d_{1,n+1}, d_{2,n+1}, ...d_{n,n+1})$. Find the coordinates of p_{n+1}.
Here we suppose the coordinates of p_{n+1} is $z_{n+1} = (z_{n+1,1}, z_{n+1,2}, ...z_{n+1,p})^T$.
Firstly, we can use D to get B and then get $Z_{n\times p}$.

Then

$$\begin{aligned} d_{i,n+1}^2 &= \sum_{k=1}^{p}(z_{n+1,k} - zi,k)^2 \\ &= \sum_{k=1}^{p} z_{n+1,k}^2 + \sum_{k=1}^{p} z_{i,k}^2 - 2\sum_{k=1}^{p} z_{n+1,k}z_{i,k} \\ &= d_{n+1}^2 + d_i^2 - 2\sum_{k=1}^{p} z_{n+1,k}z_{i,k} \end{aligned}$$

Thus

$$\begin{aligned} \sum_{i=1}^{n} d_{i,n+1}^2 &= nd_{n+1}^2 + \sum_{i=1}^{n} d_i^2 - 2\sum_{k=1}^{p}(\sum_{i=1}^{n} z_{i,k})z_{n+1,k} \\ &= nd_{n+1}^2 + \sum_{i=1}^{n} d_i^2 \qquad (Note: \sum_{i=1}^{n} z_{i,k} = 0) \\ d_{n+1}^2 &= \frac{1}{n}\sum_{i=1}^{n}(d_{i,n+1}^2 - d_i^2) \end{aligned}$$

Substitute d_{n+1}^2 with the formula above, we can get

$$\begin{array}{c} d_{i,n+1}^2 = \frac{1}{n}\sum_{i=1}^{n}(d_{i,n+1}^2 - d_i^2) + d_i^2 - 2\sum_{k=1}^{p} z_{n+1,k}z_{i,k} \\ 2\sum_{k=1}^{p} z_{n+1,k}z_{i,k} = d_i^2 - d_{i,n+1}^2 - \frac{1}{n}\sum_{i=1}^{n}(d_{i,n+1}^2 - d_i^2) \end{array}$$

Let $\alpha_i = d_i^2 - d_{i,n+1}^2$, then

$$\begin{array}{c} 2\sum_{k=1}^{p} z_{n+1,k}z_{i,k} = \alpha_i - \frac{1}{n}\sum_{i=1} n\alpha_i \\ 2Zz_{n+1} = \vec{\alpha} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\vec{\alpha} \end{array}$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, ...\alpha_n)^T$,thus

$$\begin{aligned} (Z^TZ)z_{n+1} &= \frac{1}{2}Z^T(\vec{\alpha} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\vec{\alpha}) \\ \Lambda z_{n+1} &= \frac{1}{2}Z^T(\vec{\alpha} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\vec{\alpha}) \\ z_{n+1} &= \frac{1}{2}\Lambda^{-1}Z^T(\vec{\alpha} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\vec{\alpha}) \\ z_{n+1} &= \frac{1}{2}\Lambda^{-1}Z^TH\vec{\alpha} \end{aligned}$$

Summary:
We can see that MDS is based on distance matrix, and PCO is a special instance of MDS.Compare PCO with PCA, we can find PCA is based on kernel matrix, which is

p.s.d, while PCO is based on distance matrix, which is *n.d.* However, the results of these two methods are in fact the same. While dealing with a real problem, you can choose the one that of lower complexity.