2021FALL MLHW5 Answers

1. Estimate Rate Parameter of Exponential Mixture Model(2%)

This is a toy example of Expectation-Maximization algorithm (EM) to estimate the rate parameter of two exponential distributions.

Derive updated form of EM algorithm

Given dataset $\boldsymbol{X} = \{x_1, x_2, x_3\} = \{0, 2, 4\} \in [0, \inf)$,and we would like to cluster them into 2 clusters.
We would like to apply the Expectation-Maximization algorithm(EM) to find the maximum likelihood estimation of parameters θ. That is to say, we want to estimate rate parameter and prior probability $\theta = (\pi_k, \tau_k)_{k=1}^2$ that maximize the likelihood

$$p(\boldsymbol{X}; \theta) = \prod_{i=1}^{3} p(x_i; \theta) = \prod_{i=1}^{3} \sum_{k=1}^{2} \pi_k f_\tau(x_i)$$

where $\pi_1 + \pi_2 = 1$, k be an index over the exponential distributions, i be an index over the data points and f_τ is probability density function of exponential distribution

$$f_\tau(x) = \begin{cases} \frac{1}{\tau} e^{-\frac{x}{\tau}}, x \geq 0 \\ 0, otherwise \end{cases}$$

(a) Given current parameter estimates $\theta^{[t]} = (\tau_k^{[t]}, \pi_k^{[t]})_{k=1}^2$, where t is an index over the E-M iterations, write down the Expectation Step$(E-step)$. That is to say, derive posterior probability of latent variables $\mathbb{P}[z_i = k | x_i; \theta^{[t]}]$ and expectation of log-likelihood function $Q(\theta|\theta^t)$.

Sol:

For generalization,we consider N data points and K clusters case of exponential mixture model to finish our proof and solution to this problem is special case of $N = 3, K = 2$.

Posterior prob. dist. of latent variables z_i based on current parameters $\theta(t)$

$$\mathbb{P}[z_i = k | x_i; \theta^{[t]}] = \frac{p(x_i, z_i = k; \theta^{(t)})}{\sum_{j=1}^{K} p(x_i, z_i = j; \theta^{(t)})} = \delta_{ik}^{[t]}$$

Log-likelihood functions can be defined as following

$$\log p(\boldsymbol{X} : \theta) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} \pi_k f_\tau(x_i))$$

Log-likelihood of parameter θ given data $x_i \in [0, \infty)$ and latent variable z_i

$$\log p(x_i, z_i = k; \boldsymbol{\theta}) = \log \pi_k f_{\tau_k}(x_i) = \log\left(\frac{\pi_k}{\tau_k}\right) - \frac{x_i}{\tau_k}$$

Expectation of log-likelihood function

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[t]}) &= \sum_{i=1}^{N} \mathbb{E}_{z_i|x_i;\theta^{[t]}}[\log p(x_i, z_i; \boldsymbol{\theta})] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{[t]} (\log(\frac{\pi_k}{\tau_k}) - \frac{x_i}{\tau_k}) \end{aligned}$$

(b) Given current parameter $\theta^{[t]}$,write down the Maximization Step$(M-step)$.That is to say, derive estimated rate parameter $\tau_k^{[t+1]}$ and the prior probability $\pi_k^{[t+1]}$

Sol:

For generalization,we consider N data points and K clusters case of exponential mixture model to finish our proof and solution to this problem is special case of $N = 3, K = 2$.

Maximization Step (M-step): Choose

$$\theta^{[t+1]} = \arg\max_{\theta} Q(\theta|\theta^{[t]}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{[t]} (log(\frac{\pi_k}{\tau_k}) - \frac{x_i}{\tau_k})$$

Partial derivative over τ_k

$$\frac{\partial}{\partial \tau_k} logQ(\theta|\theta^{(t)}) = \sum_{i=1}^{N} \delta_{ik}^{[t]} (\frac{x_i}{\tau_k^2} - \frac{1}{\tau_k})$$

Setting derivate equals zero

$$\tau_k^{[t+1]} = \frac{\sum_{i=1}^{N} \delta_{ik}^{[t]} \cdot x_i}{\sum_{i=1}^{N} \delta_{ik}^{[t]}}$$

Partial derivative over π_k with Lagrange multiplier constrained in $\sum_{k=1}^{K} \pi_k = 1$

$$\nabla_{\pi_k} \left(logQ(\theta|\theta^{(t)} - \lambda(\sum_{k=1}^{K} \pi_k - 1) \right) = \sum_{i=1}^{N} \frac{\delta_{ik}^{[t]}}{\pi_k} - \lambda$$

Setting derivate equals zero

$$\pi_k^{[t+1]} = \lambda^{-1} \sum_{i=1}^{N} \delta_{ik}^{[t]}$$

The constraint $\sum_{k=1}^{K} \pi_k = 1$ implies $\lambda = (\sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{[t]}) = N$

$$\pi_k^{[t+1]} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{[t]}$$

Go through the First iteration of EM algorithm

$\boldsymbol{X} = \{x_1, x_2, x_3\} = \{0, 2, 4\}$ be our three data points, presumed to have each been generated from one of two exponential distributions. Besides, let's initialize the rate parameter and the prior over two exponential distributions to some reasonable values to make calculation easy
$\tau_1^{[0]} = 1, \tau_2^{[0]} = 2$
$\pi_1^{[0]} = \pi_2^{[0]} = 0.5$

(c) Please go through the first iteration of EM algorithm and either write down your derivation or provide your code for computation in report. Besides, construct a table of your answers as following.

i	**1**	**2**	**3**
x_i	0	2	4
$p(x_i, z_i = 1; \theta^{(t)})$			
$p(x_i, z_i = 2; \theta^{(t)})$			
$\sum_{j=1}^{2} p(x_i, z_i = j; \theta^{(t)})$			
$\delta_{i1}^{[t]} = \mathbb{P}[z_i = 1 \vert x_i; \theta^{[t]}]$			
$\delta_{i2}^{[t]} = \mathbb{P}[z_i = 2 \vert x_i; \theta^{[t]}]$			

Sol:

i	**1**	**2**	**3**
x_i	0	2	4
$p(x_i, z_i = 1; \theta^{(t)})$	$\frac{1}{2}$	$\frac{e^{-2}}{2}$	$\frac{e^{-4}}{2}$
$p(x_i, z_i = 2; \theta^{(t)})$	$\frac{1}{2}$	$\frac{e^{-1}}{2}$	$\frac{e^{-2}}{2}$
$\sum_{j=1}^{2} p(x_i, z_i = j; \theta^{(t)})$	1	$\frac{e^{-2}+e^{-1}}{2}$	$\frac{e^{-4}+e^{-2}}{2}$
$\delta_{i1}^{[t]} = \mathbb{P}[z_i = 1 \vert x_i; \theta^{[t]}]$	0.5	$\frac{e^{-2}}{e^{-2}+e^{-1}} \approx 0.26$	$\frac{e^{-4}}{e^{-4}+e^{-2}} \approx 0.12$
$\delta_{i1}^{[t]} = \mathbb{P}[z_i = 2 \vert x_i; \theta^{[t]}]$	0.5	$\frac{e^{-1}}{e^{-2}+e^{-1}} \approx 0.73$	$\frac{e^{-2}}{e^{-4}+e^{-2}} \approx 0.88$

(d)Please derive the estimated rate parameter $\tau_1^{[1]}, \tau_2^{[1]}, \pi_1^{[1]}, \pi_2^{[1]}$ with table from (c) and either write down your derivation or provide your code for computation in report

Sol:

$$\tau_1^{[1]} = \frac{0.5 * 0 + 0.26 * 2 + 0.12 * 4}{0.5 + 0.26 + 0.12} \approx 1.065$$

$$\tau_2^{[1]} = \frac{0.5 * 0 + 0.73 * 2 + 0.88 * 4}{0.5 + 0.73 + 0.88} \approx 2.36$$

$$\pi_1^{[1]} = \frac{0.5 + 0.26 + 0.12}{3} \approx 0.2933$$

$$\pi_2^{[1]} = \frac{0.5 + 0.73 + 0.88}{3} \approx 0.703$$

2. Principle Component Analysis (1%)

Given 10 samples in 3D:
$(1,2,3)$, $(4,8,5)$, $(3,12,9)$, $(1,8,5)$, $(5,14,2)$, $(7,4,1)$, $(9,8,9)$, $(3,8,1)$, $(11,5,6)$, $(10,11,7)$

(a) What are the principal axes? Please write down your derivation or provide your code for computation in report

Sol:

$\mu = \begin{bmatrix} 5.4 & 8 & 4.8 \end{bmatrix}$

$$\Sigma = \begin{bmatrix} 12.04 & 0.5 & 3.28 \\ 0.5 & 12.2 & 2.9 \\ 3.28 & 2.9 & 8.16 \end{bmatrix} = Q\Lambda Q^T$$

$$Q = \begin{bmatrix} -0.6165 & -0.6781 & 0.3998 \\ -0.5888 & 0.7343 & 0.3375 \\ -0.5225 & -0.0272 & -0.8521 \end{bmatrix} \quad \Lambda = \begin{bmatrix} 15.2974 & 0 & 0 \\ 0 & 11.6305 & 0 \\ 0 & 0 & 5.472 \end{bmatrix}$$

1st = $\begin{bmatrix} -0.6165 & -0.5888 & -0.5225 \end{bmatrix}^T$
2nd = $\begin{bmatrix} -0.6781 & 0.7343 & -0.0272 \end{bmatrix}^T$
3rd = $\begin{bmatrix} 0.3998 & 0.3375 & -0.8521 \end{bmatrix}^T$

(b) Please compute the principal components for each sample and either write down your derivation or provide your code for computation in report

$$Principal\ Component = \begin{bmatrix} -0.6165 & -0.6781 & 0.3998 \\ -0.5888 & 0.7343 & 0.3375 \\ -0.5225 & -0.0272 & -0.8521 \end{bmatrix}^T (x_i - \mu)$$

1. $\begin{bmatrix} 7.1865 & 1.3732 & 2.2510 \end{bmatrix}$
2. $\begin{bmatrix} 0.7587 & -0.9439 & 0.7302 \end{bmatrix}$
3. $\begin{bmatrix} -3.0703 & -4.4505 & 3.1883 \end{bmatrix}$
4. $\begin{bmatrix} 2.6084 & -2.9785 & 1.9297 \end{bmatrix}$
5. $\begin{bmatrix} -1.8229 & -4.7540 & -4.2515 \end{bmatrix}$
6. $\begin{bmatrix} 3.3545 & 3.9189 & -2.5275 \end{bmatrix}$
7. $\begin{bmatrix} -4.4146 & 2.5560 & 2.1395 \end{bmatrix}$
8. $\begin{bmatrix} 3.4656 & -1.7313 & -2.2784 \end{bmatrix}$
9. $\begin{bmatrix} -2.3135 & 6.0337 & -0.2038 \end{bmatrix}$
10. $\begin{bmatrix} -5.7524 & 0.9764 & -0.9773 \end{bmatrix}$

(c) What is the average reconstruction error if reduce dimension to 2D? Here the reconstruction error is defined as the squared loss. Please write down your derivation or provide your code for computation in report

將(b)小題的答案，捨去第三個dimension的值(第三個dimension為0)

$$W = \begin{bmatrix} -0.6165 & -0.6781 & 0.3998 \\ -0.5888 & 0.7343 & 0.3375 \\ -0.5225 & -0.0272 & -0.8521 \end{bmatrix}^T$$

$x_1' = [7.1865, 1.3732, 0]$

$Reconsturction\ r_i = X'W + \mu$

$[1.90009072 \quad 2.75992709 \quad 1.08178971]$
$[4.29198496 \quad 8.24651657 \quad 4.37774211]$
$[4.27485905 \quad 13.0763358 \quad 6.28310968]$
$[1.77163801 \quad 8.65147726 \quad 3.35553912]$
$[3.29997625 \quad 12.5647067 \quad 5.62297154]$
$[5.98934216 \quad 3.14672348 \quad 3.15384320]$
$[9.85550052 \quad 8.72228056 \quad 7.17681721]$
$[2.08893199 \quad 7.23080501 \quad 2.94160433]$
$[10.9184895 \quad 4.93118246 \quad 6.17370944]$
$[9.60918683 \quad 10.6700449 \quad 7.83287366]$

$average\ error = \frac{1}{10}\sum_{10}^{i=1} \|x_i - r_i\|^2$

$e = [\,5.0671832 \quad 0.53323052 \quad 10.16525752 \quad 3.72409942 \quad 18.07607017 \quad 6.38855061 \quad 4.57756584 \quad 5.19153323 \quad 0.04155478$

$average\ error = 5.472$

3. Constrained Mahalanobis Distance Minimization Problem (1%)

(a) Let $A \in R^{m\times n}$, show that AA^T and A^TA are both symmetric and positive semi-definite, and share the same non-zero eigenvalues.

Sol:

Symmetric:
1. $(AA^T)^T = (A^T)^TA^T = AA^T$ 2. $(A^TA)^T = A^T(A^T)^T = A^TA$

Positive Semi-definite:
$\forall v \in R^m \setminus 0, \forall u \in R^n \setminus 0$
1. $v^T(AA^T)v = (A^Tv)^T(A^Tv) = \|A^Tv\|^2 \geq 0$
2. $u^T(A^TA)u = (Au)^T(Au) = \|Au\|^2 \geq 0$

Same non-zero eigenvalues:
Let v be a nonzero eigenvector of A^TA with eigenvalue $\lambda \neq 0$. This means $(A^TA)v = \lambda v$
Multiply both sides on the left by A, yields $AA^T(Av) = \lambda(Av)$
Which is the statement that the vector Av is an eigenvector of AA^T with eigenvalue λ
In similar way, we also can show that all non-zero eigen values of AA^T are eigen values of A^TA

(b) Let $\Sigma \in R^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in R^m$. Please construct a set of points $x_1, \ldots, x_n \in R^m$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T = \Sigma, \quad \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$$

Sol:

$WLOG$, let $\mu = 0$. Since Σ is symmetric positive semi-definite matrix, we can perform eigen decomposition as follows:

$$\Sigma = UDU^T = \Sigma_i(d_iu_iu_i^T).$$

Let $n = 2m$ and construct a set of points $x_1, \ldots, x_m, \ldots, x_{2m}$
where $x_i = \sqrt{d_i}u_i$ and $x_{m+i} = -\sqrt{d_i}u_i \ \forall\ 1 \leq i \leq m$. Then,

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \mu = 0$$
$$\frac{1}{n}\sum_{i=1}^{n} x_ix_i^T = \sum_{i=1}^{n}(d_iu_iu_i^T) = UDU^T = \Sigma$$

(c)Let $1 \leq k \leq m$, solve the following optimization problem (and justify with proof):

$$\begin{aligned} &minimize \quad Trace(\Phi^T\Sigma\Phi) \\ &subject\ to \quad \Phi^T\Phi = I_k \\ &variables \quad \Phi \in R^{m\times k} \end{aligned}$$

Sol:

$$Trace(\Phi^T\Sigma\Phi) = Trace(\Sigma\Phi\Phi^T)$$
$$\frac{\partial\, Trace(\Sigma\Phi\Phi^T)}{\partial\Phi} = \Sigma\Phi$$

Using Lagrange Multiplier λ

$$Trace(\Sigma\Phi\Phi^T) - \lambda(\Phi^T\Phi - I) = 0$$

Taking derivatives $\partial\Phi$ on both sides

$$\Sigma\Phi = \lambda\Phi$$

The result indicates that Φ is an eigenvector of Σ
To obtain the least value of λ, we choose the Φ correspond to the k-least eigenvalue.

$$\Phi = [u_{m-k+1}, \ldots, u_m]$$

Where $U = [u_1, \ldots, u_m]$ is orthogonal matrix of eigenvectors (of Σ), $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ is diagonal matrix of the associated eigenvalues arranged in non-ascending order
$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$