Theory Assignment 4

Youzhi Qu USCID: 5045807938

November 2019

1 Optimization over the simplex

1.1

$L(\boldsymbol{q}, \{\lambda_k\}, \lambda) = \sum_{k=1}^{K} a_k \ln q_k - \lambda_k q_k + \lambda(\sum_{k=1}^{K} q_k - 1)$

subject to: $\lambda_k \geq 0, \lambda \gtreqless 0$

for each k, take partial derivative to q_k and set it to 0:

$\frac{\partial L}{\partial q_k} = \frac{a_k}{q_k} - \lambda_k + \lambda = 0$

$q_K^* = \frac{a_k}{\lambda_k - \lambda} \neq 0$

according to complementary slackness we can get:

$\lambda_k q_k^* = 0 \Rightarrow \lambda_k = 0$

since $\lambda_k = 0$ we can get:

$q_k^* = -\frac{a_k}{\lambda}$ (1)

since $\sum_{k=1}^{K} q_k^* = \sum_{k=1}^{K} -\frac{a_k}{\lambda} = 1$, we can get

$\lambda = -\sum_{k=1}^{K} a_k$ (2)

from (1) and (2) we can get :

$q_k^* = \frac{a_k}{\sum_{k'=1}^{K} a_{k'}}$

1.2

$L(\boldsymbol{q}, \{\lambda_k\}, \lambda) = \sum_{k=1}^{K}(q_k b_k - q_k \ln q_k) - \lambda_k q_k + \lambda(\sum_{k=1}^{K} q_k - 1)$

subject to: $\lambda_k \geq 0, \lambda \gtreqless 0$

for each k, take partial derivative to q_k and set it to 0, we can get:

$\frac{\partial L}{\partial q_k} = b_k - \ln q_k - q_k \times \frac{1}{q_k} - \lambda_k + \lambda = 0$

we can get:

$q_k^* = \exp(b_k - 1 - \lambda_k + \lambda)$

according to complementary slackness we can get:

$\lambda_k q_k^* = 0 \Rightarrow \lambda_k = 0$

since $\lambda_k = 0$ we can get:

$q_k^* = \frac{\exp(b_k)}{\exp(1-\lambda)}$ (1)

since $\sum_{k=1}^{K} q_k^* = \sum_{k=1}^{K} \frac{\exp(b_k)}{\exp(1-\lambda)} = 1$:

$\exp(1-\lambda) = \sum_{k=1}^{K} \exp(b_k)$ (2)

from (1) and (2) we can get :

$q_k^* = \frac{\exp(b_k)}{\sum_{k'=1}^{K} \exp(b_{k'})}$

2 Gaussian Mixture Model and EM

2.1

for each $w_k (k = 1, 2...K)$ simplify MLE:

$\arg\max_{\boldsymbol{w}} \sum_n \sum_k \gamma_{nk} \ln w_k$

subject to $w_k \geq 0, \sum_{k=1}^{K} w_k = 1$

according to Problem 1.1, set $a_k = \sum_n \gamma_{nk}$ we can get:

$$w_k^* = \frac{\sum_n \gamma_{nk}}{\sum_{k=1}^{K} \sum_n \gamma_{nk}} = \frac{\sum_n \gamma_{nk}}{N}$$

for each $\mu_k, \sum_k (k = 1, 2...K)$ simplify MLE:

$$\begin{aligned}
\arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} \ln \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \Sigma_k) &= \arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} \ln [\tfrac{1}{(\sqrt{2\pi}\sigma_k)^D} \exp(-\tfrac{\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2})] \\
&= \arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} (-D \ln \sigma_k - \tfrac{\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2} - D\sqrt{2\pi})
\end{aligned}$$

take partial derivative to $\boldsymbol{\mu}_k$ and set it to 0, we can get:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} = \sum_n \gamma_{nk} \frac{(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}{\sigma_k^2} = 0$$

$$\boldsymbol{\mu}_k^* = \frac{\sum_n \gamma_{nk} \boldsymbol{x}_n}{\sum_n \gamma_{nk}}$$

take partial derivative to σ_k and set it to 0, we can get:

$\frac{\partial}{\partial \sigma_k} = \sum_n \gamma_{nk} (-\frac{D}{\sigma_k} - \frac{\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2}{2} \times \frac{-2}{\sigma_k^3}) = \sum_n \gamma_{nk} (-\frac{D}{\sigma_k} + \frac{\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3}) = 0$

$$(\sigma_k^*)^2 = \frac{\sum_n \gamma_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k^*\|^2}{D \sum_n \gamma_{nk}}$$

D is dimension of a data point $\boldsymbol{x}_n$

2.2

for $\arg\max_{\boldsymbol{q}_n \in \Delta} \mathbb{E}_{z_n \sim \boldsymbol{q}_n} [\ln p(\boldsymbol{x}_n, z_n; \theta^{(t)})] - \mathbb{E}_{z_n \sim \boldsymbol{q}_n} [\ln \boldsymbol{q}_n]$

it is same as:

$\arg\max_{\boldsymbol{q}_n \in \Delta} \boldsymbol{q}_n \ln p(\boldsymbol{x}_n, z_n; \theta^{(t)}) - \boldsymbol{q}_n \ln \boldsymbol{q}_n$

according to Problem 1.2, set $b_k = \ln p(\boldsymbol{x}_n, z_n = k; \theta^{(t)})$, we can get:

$$q_{nk}^* = \frac{p(\boldsymbol{x}_n, z_n = k; \theta^{(t)})}{\sum_{k=1}^{K} p(\boldsymbol{x}_n, z_n = k; \theta^{(t)})} = \frac{p(\boldsymbol{x}_n, z_n = k; \theta^{(t)})}{p(\boldsymbol{x}_n; \theta^{(t)})} = p(z_n|\boldsymbol{x}_n; \theta^{(t)})$$

2.3

$\gamma_{nk} = p(z_n = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n, z_n=k)}{p(\boldsymbol{x}_n)} = \frac{w_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)} = \frac{w_k \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2)}{\sum_{j=1}^{K} w_j \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2)}$

assuming $\Sigma_k = \sigma_k^2 I, w_k = \frac{\sum_n \gamma_{nk}}{N}$(mixing proportion of each Gaussian)

as $\sigma^2 \to 0$, the summation of denominator will be dominated by the term with the smallest $\min_{\boldsymbol{\mu}_j} \|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2$.

for that j, let $\sigma^2 \to 0$

$\gamma_{nj} \approx \lim_{\sigma^2 \to 0} \frac{w_j \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2)}{w_j \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2)} \approx 1$

therefore, for that j, $\gamma_{nj} \approx 1$

and for other which $i \neq j$, $\gamma_{ni} \approx \lim_{\sigma^2 \to 0} \frac{0}{w_j \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2)} \approx 0$

we can derive when $\sigma^2 \to 0$ in $\Sigma_k = \sigma^2 I$, GMM reduces to K-means

3 Hidden Markov Model

3.1

$P(X_{T+1} = s|O_{1:T} = o_{1:T}) = \frac{P(X_{T+1}=s, O_{1:T}=o_{1:T})}{P(O_{1:T}=o_{1:T})}$

$= \frac{\sum_{s'} P(X_{T+1}=s, X_T=s', O_{1:T}=o_{1:T})}{\sum_{s''} P(X_T=s'', O_{1:T}=o_{1:T})}$

$= \frac{\sum_{s'} P(X_T=s', O_{1:T}=o_{1:T}) P(X_{T+1}=s|X_T=s', O_{1:T}=o_{1:T})}{\sum_{s''} P(X_T=s'', O_{1:T}=o_{1:T})}$

$= \frac{\sum_{s'} P(X_T=s', O_{1:T}=o_{1:T}) P(X_{T+1}=s|X_T=s')}{\sum_{s''} P(X_T=s'', O_{1:T}=o_{1:T})}$

$= \frac{\sum_{s'} \alpha_{s'}(T) a_{s',s}}{\sum_{s''} \alpha_{s''}(T)}$

3.2

a)

t=1:

$\alpha_A(1) = P(X_1 = A) b_{A,o_1} = 0.7 * 0.4 = 0.28$

$\alpha_B(1) = P(X_1 = B) b_{B,o_1} = 0.3 * 0.7 = 0.21$

t=2:

$\alpha_A(2) = b_{A,o_2} \sum_{s'} a_{s',A} \alpha_{s'}(1) = 0.6 * (0.2 * 0.28 + 0.7 * 0.21) = 0.1218$

$\alpha_B(2) = b_{B,o_2} \sum_{s'} a_{s',B} \alpha_{s'}(1) = 0.3 * (0.7 * 0.28 + 0.2 * 0.21) = 0.0714$

t=3:

$P(X_3 = A|O_{1:2}) = \frac{\sum_{s'} \alpha_{s'}(2) a_{s',A}}{\sum_{s''} \alpha_{s''}(2)} = \frac{\alpha_A(2) a_{A,A} + \alpha_B(2) a_{B,A}}{\alpha_A(2) + \alpha_B(2)} = 0.3848$

$P(X_3 = B|O_{1:2}) = \frac{\sum_{s'} \alpha_{s'}(2) a_{s',B}}{\sum_{s''} \alpha_{s''}(2)} = \frac{\alpha_A(2) a_{A,B} + \alpha_B(2) a_{B,B}}{\alpha_A(2) + \alpha_B(2)} = 0.5152$

$P(X_3 = End|O_{1:2}) = \frac{\sum_{s'} \alpha_{s'}(2) a_{s',End}}{\sum_{s''} \alpha_{s''}(2)} = \frac{\alpha_A(2) a_{A,End} + \alpha_B(2) a_{B,End}}{\alpha_A(2) + \alpha_B(2)} = 0.1$

the most likely state at t = 3 given the observed sequence is B.

b)

t=1:

$\delta_A(1) = \pi_A b_{A,o1} = 0.7 * 0.4 = 0.28$

$\delta_B(1) = \pi_B b_{B,o1} = 0.3 * 0.7 = 0.21$

t=2:

$\delta_A(2) = b_{A,o2} \max_{s'} a_{s',A} \delta_{s'}(1) = 0.6 * \max(0.2 * 0.28, 0.7 * 0.21) = 0.0882$

$\Delta_A(2) = B$

$\delta_B(2) = b_{B,o2} \max_{s'} a_{s',B} \delta_{s'}(1) = 0.3 * \max(0.7 * 0.28, 0.2 * 0.21) = 0.0588$

$\Delta_B(2) = A$

t=3:

$\delta_A(3) = b_{A,o2} \max_{s'} a_{s',A} \delta_{s'}(2) = 0.6 * \max(0.2 * 0.0882, 0.7 * 0.0588) = 0.024696$

$\Delta_A(3) = B$

$\delta_B(3) = b_{B,o2} \max_{s'} a_{s',B} \delta_{s'}(2) = 0.3 * \max(0.7 * 0.0.0882, 0.2 * 0.0588) = 0.018522$

$\Delta_B(3) = A$

the most likely sequence of states is $o_3^* = A, o_2^* = B, o_1^* = A$.

3.3

$$\begin{aligned}
P(X_2 = s'|O_1 = o_1, O_1 = o_2) &= \frac{P(X_2=s', O_1=o_1, O_1=o_2)}{P(O_1=o_1, O_1=o_2)} \\
&= \frac{\sum_s P(X_1=s, X_2=s', O_1=o_1, O_1=o_2)}{P(O_1=o_1)P(O_2=o_2)} \\
&= \frac{\sum_s P(O_1=o_1, O_1=o_2|X_1=s, X_2=s')P(X_1=s, X_2=s')}{P(O_1=o_1)P(O_2=o_2)} \\
&= \frac{P(O_2=o_2|X_2=s') \sum_s P(x_2=s'|x_1=s)P(O_1=o_1|X_1=s)P(X_1=s)}{P(O_1=o_1)P(O_2=o_2)}
\end{aligned}$$

set the model parameters :

$P(X_1 = s) = P(z = s) = w_s$

$P(X_{t+1} = s'|X_t = s) = P(z = s') = w_{s'}$

$P(O_t = o|X_t = s) = \mathcal{N}(O_t = o|\mu_s, \Sigma_s)$

Therefore:

$$\begin{aligned}
P(X_2 = s'|O_1 = o_1, O_1 = o_2) &= \frac{P(O_2=o_2|X_2=s') \sum_s P(x_2=s'|x_1=s)P(O_1=o_1|X_1=s)P(X_1=s)}{P(O_1=o_1)P(O_2=o_2)} \\
&= \frac{\mathcal{N}(O_2=o_2|\mu'_s, \Sigma'_s)w_{s'} \sum_s N(O_1=o_1|\mu_s, \Sigma_s)w_s}{P(O_1=o_1)P(O_2=o_2)} \\
&= \frac{\mathcal{N}(O_2=o_2|\mu'_s, \Sigma'_s)w_{s'} P(O_1=o_1)}{P(O_1=o_1)P(O_2=o_2)} \\
&= \frac{\mathcal{N}(O_2=o_2|\mu'_s, \Sigma'_s)w_{s'}}{P(O_2=o_2)} \\
&= \frac{P(o_2, X_2=s')}{P(O_2=o_2)}
\end{aligned}$$

the result is as same as in problem 2.2 when n=2,k=s':

$p(X_2 = s'|o_2) = q_{n=2,k=s'} = \frac{p(o_2, X_2=s'; \theta^{(t)})}{p(o_2; \theta^{(t)})} = \frac{P(o_2, X_2=s')}{P(O_2=o_2)}$

Therefore, HMM can be reduced to GMM.