1. (a). $\frac{\partial}{\partial\beta_t} = \frac{\partial}{\partial\beta_t}\left[(e^{\beta_t}-e^{-\beta_t})\sum_n w_t(n)\mathbb{I}[y_n \neq h_t(x_n)] + e^{-\beta_t}\sum_n w_t(n)\right]$

$\varepsilon_t = \sum_n w_t(n)\,\mathbb{I}[y_n \neq h_t(x_n)]$

$\sum_n w_t(n) = 1$

$\frac{\partial}{\partial\beta_t} = \frac{\partial}{\partial\beta_t}\left[(e^{\beta_t}-e^{-\beta_t})\varepsilon_t + e^{-\beta_t}\right]$

$= (e^{\beta_t}+e^{-\beta_t})\varepsilon_t - e^{-\beta_t} = 0$

$(e^{\beta_t}+e^{-\beta_t})\varepsilon_t = e^{-\beta_t}$

$(e^{2\beta_t}+1)\varepsilon_t = 1$

$e^{2\beta_t} = \frac{1}{\varepsilon_t} - 1 = \frac{1-\varepsilon_t}{\varepsilon_t}$

$2\beta_t = \log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

$\beta_t^* = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

(b). We want to find minimum ε_t. For hard SVM, if the datasets are linearly separable, then the minimum $\varepsilon_t = 0$. So according to the equation of β_t above,

$\beta_{t_1} = \frac{1}{2}\log\left(\frac{1-0}{0}\right) \to \infty$

2. (a). Since $K=3$, and there are 4 data points, we can assign $X_1=1$ and $X_2=2$ to the first cluster, $X_3=5$ to the second cluster, and $X_4=7$ to the third cluster. Thus, $u_1=\frac{1+2}{2}=1.5$. $u_2=5$. $u_3=7$. Objective:

$$(1-1.5)^2+(2-1.5)^2+(5-5)^2+(7-7)^2=0.5$$

(b). Suboptimal cluster assignment: Assign $X_1=1$ to cluster 1, $X_2=2$ to cluster 2, and $X_3=5$ and $X_4=7$ to cluster 3. Thus, $u_1=1$. $u_2=2$. $u_3=\frac{5+7}{2}=6$. The objective in this case is:

$$(1-1)^2+(2-2)^2+(5-6)^2+(7-6)^2=2 \quad 2>0.5 \quad \therefore \text{Suboptimal}$$

According to Lloyd's algorithm, after setting $\{u_k\}$, J will be minimized over $\{r_{nk}\}$, so the points will be reassigned to the closest cluster. However, in the assignment described above, all the points are already closest to the centroids in the cluster. X_1 closest to u_1, X_2 closest to u_2, and X_3 and X_4 closest to u_3. So the assignment will not change and J stays the same. So it does not converge to global minimum, and the algorithm will not improve J from 2 to the optimal 0.5.

3. (a). $\nabla_{\mu_j} l(\theta) = \frac{\partial}{\partial \mu_j}\left(\sum_k \sum_n \gamma_{nk} \log w_k + \sum_k \left\{ \sum_n \gamma_{nk} \log N(x_n \mid \mu_k, \Sigma_k) \right\}\right)$

$= \frac{\partial}{\partial \mu_j}\left\{ \sum_k \sum_n \log \left[w_k \frac{1}{|2\pi \Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mu_k - x_n)^T \Sigma_k^{-1} (\mu_k - x_n)} \right] \gamma_{nk} \right\}$

$= \frac{\partial}{\partial \mu_j}\left\{ \sum_k \sum_n \left[\log w_k - \frac{1}{2}\log(2\pi \Sigma_k) - \frac{1}{2}(\mu_k - x_n)^T \Sigma_k^{-1} (\mu_k - x_n) \right] \gamma_{nk} \right\}$

$= \sum_n -\frac{1}{2} \times 2 \Sigma_j^{-1} (\mu_j - x_n) \gamma_n$

$= \sum_n - \Sigma_j^{-1} (\mu_j - x_n) \gamma_{nj}$

(b). $\sum_n - \Sigma_j^{-1} (\mu_j - x_n) \gamma_{nj} = 0$

$\sum_n \Sigma_j^{-1} (\mu_j - x_n) \gamma_{nj} = 0$

$\sum_n (\mu_j - x_n) \gamma_{nj} = \sum_n \mu_j \gamma_{nj} - \sum_n x_n \gamma_{nj} = 0$

$\therefore \mu_j = \frac{1}{\sum_n \gamma_{nj}} \sum_n \gamma_{nj} x_n$

(c).

$$W_1 = \frac{\sum_n \gamma_{n1}}{\sum_k \sum_n \gamma_{nk}} = \frac{0.2+0.2+0.8+0.9+0.9}{0.2+0.2+0.8+0.9+0.9+0.8+0.8+0.2+0.1+0.1} = 0.6$$

$$W_2 = 1 - W_1 = 0.4$$

$$\mu_1 = \frac{\sum_n \gamma_{n1} x_n}{\sum_n \gamma_{n1}} = \frac{0.2\times5+0.2\times15+0.8\times25+0.9\times30+0.9\times40}{0.2+0.2+0.8+0.9+0.9} = 29$$

$$\mu_2 = \frac{\sum_n \gamma_{n2} x_n}{\sum_n \gamma_{n2}} = \frac{0.8\times5+0.8\times15+0.2\times25+0.1\times30+0.1\times40}{0.8+0.8+0.2+0.1+0.1} = 14$$