PS3.

3.1 (a): i $J=\sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2$, where $\mu_j \triangleq \frac{1}{|C_j|}\sum_{x\in C_j} x$, $j=1,2,\dots,k$

$$C=\arg\min_C J = \arg\min_C \sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2 = \arg\min_C \sum_{j=1}^{k}\sum_{x\in C_j}\Big\|x-\frac{1}{|C_j|}\sum_{x'\in C_j}x'\Big\|^2$$

$$=\arg\min_C \sum_{j=1}^{k}\sum_{x\in C_j}\Big\|\frac{1}{|C_j|}\Big(|C_j|x-\sum_{x'\in C_j}x'\Big)\Big\|^2$$

$$=\arg\min_C \sum_{j=1}^{k}\sum_{x\in C_j}\frac{1}{|C_j|^2}\Big\|\sum_{x'\in C_j}(x-x')\Big\|^2 = \arg\min_C \sum_{j=1}^{k}\frac{1}{|C_j|^2}\sum_{x\in C_j}\Big(\sum_{x'\in C_j}(x-x')\Big)^T\Big(\sum_{x'\in C_j}(x-x')\Big)$$

$$=\arg\min_C \sum_{j=1}^{k}\frac{1}{|C_j|^2}\cdot\sum_{x\in C_j}\sum_{x'\in C_j}(x-x')^T\cdot\sum_{x'\in C_j}(x-x')$$

$$=\arg\min_C \sum_{j=1}^{k}\frac{1}{|C_j|^2}\cdot\sum_{x\in C_j}\sum_{x'\in C_j}(x-x')^T(x-x') = \arg\min_C \sum_{j=1}^{k}\frac{1}{2|C_j|}\cdot\sum_{x\in C_j}\sum_{x'\in C_j}\|x-x'\|^2$$

ii. $\sum_{x\in X}(x-\mu)^T(x-\mu)=\text{const}$,

(b):

$$=\sum_{j=1}^{k}\sum_{x\in C_j}(x-\mu_j+\mu_j-\mu)^T(x-\mu_j+\mu_j-\mu) = \sum_{j=1}^{k}\sum_{x\in C_j}(x-\mu_j)^T(x-\mu_j) + \underbrace{\sum_{j=1}^{k}\sum_{x\in C_j}\left[(x-\mu_j)^T(\mu_j-\mu)+(\mu_j-\mu)^T(x-\mu_j)\right]}_{①}$$

$$+\underbrace{\sum_{j}^{k}\sum_{x\in C_j}(\mu_j-\mu)^T(\mu_j-\mu)}_{②}$$

$$② = \sum_{j}^{k}|C_j|\cdot(\mu_j-\mu)^T(\mu_j-\mu) = \sum_{j}^{k}\sum_{i}^{k}|C_j|\cdot(\mu_j+\mu_i-\mu_i-\mu)^T(\mu_j+\mu_i-\mu_i-\mu)$$

$$=\sum_{x\in C_i}\sum_{j}^{k}\sum_{i}^{k}|C_j|\cdot\{(\mu_j-\mu_i)^T(\mu_j-\mu_i)+(\mu_j-\mu_i)^T(\mu_i-\mu)+(\mu_i-\mu)^T(\mu_j-\mu_i)+(\mu_i-\mu)^T(\mu_i-\mu)\}.$$

$$=\underbrace{\sum_{j}^{k}\sum_{i}^{k}|C_i||C_j|\cdot\|\mu_j-\mu_i\|^2}_{③} + \underbrace{\sum_{j}^{k}\sum_{i}^{k}|C_i||C_j|\left((\mu_j-\mu_i)^T(\mu_i-\mu)+(\mu_i-\mu)^T(\mu_j-\mu_i)\right)}_{④} + \underbrace{\sum_{j}^{k}\sum_{i}^{k}|C_i||C_j|\cdot(\mu_i-\mu)^T(\mu_i-\mu)}_{⑤}.$$

3.1 (b) i. $J(C,\mu) = \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$.

the k-means algorithm has two steps in each epoch.

I: $C^{(i)} = \arg\min_j \| x^{(i)} - \mu_j \|^2$ with μ fixed.

II: $\mu_{c^{(i)}} = \frac{1}{|C^{(i)}|} \sum_{x_j \in C^{(i)}} x_j$, with $C^{(i)}, i=1..m$ fixed.

In step I:

$\| x^{(i)} - \mu_{c_{t+1}^{(i)}} \|^2 \leq \| x^{(i)} - \mu_{c_t^{(i)}} \|^2$, because in each epoch it searches for a nearest centroid for a point.

hence $\sum_{i=1}^{m} \| x^{(i)} - \mu_{c_{t+1}^{(i)}} \|^2 \leq \sum_{i=1}^{m} \| x^{(i)} - \mu_{c_t^{(i)}} \|^2$. $J(C^{t+1}, \mu^*) \leq J(C^t, \mu^*)$, it decreases

In step II: the new centroid of each Cluster C_j. μ_j is adjusted to be the average that has the smallest sum of distance to other within cluster points such that $\sum_{i=1}^{m} 1\{c^{(i)}=j\} \| x^{(i)} - \mu_{c^{(i)}}^{t+1} \|^2 \leq \sum_{i=1}^{m} 1\{c^{(i)}=j\} \| x^{(i)} - \mu_{c^{(i)}}^{t} \|^2$,

Therefor for $j=1,\dots k$, we have

$$\sum_{j=1}^{k} \sum_{i=1}^{m} 1\{c^{(i)}=j\} \| x^{(i)} - \mu_{c^{(i)}}^{t+1} \|^2 \leq \sum_{j=1}^{k} \sum_{i=1}^{m} 1\{c^{(i)}=j\} \| x^{(i)} - \mu_{c^{(i)}}^{t} \|^2,$$

$\Rightarrow \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}}^{t+1} \|^2 \leq \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}}^{t} \|^2$. $J(C^*, \mu^{t+1}) \leq J(C^*, \mu)$, it decreases.

So that, J decreases monotonically in each epoch.

ii. Since J decreases monotonically, J must converge.

However, there is no guarantee that J converges to the global minimum as J is a non-convex function.

3.2 (b).

$$\hat{C} = \frac{1}{m-1} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T.$$

if $\forall u$, $u^T cov(x) u = u^T \hat{C} u > 0$, then $\hat{C}$ is positive definite matrix

and then $\hat{C}$ is non singular.

$$u^T \hat{C} u = \frac{1}{m-1} \sum_{i=1}^{m} u^T (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T u$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} \cancel{(x^{(i)}-\hat{\mu})u} \, \| u^T (x^{(i)} - \hat{\mu}) \|^2 . \geq 0 .$$

if we want $u^T \hat{C} u > 0$, then for $i = 1, \cdots, m$, $x^{(i)}$:

$$m \left\{ \begin{array}{l} u^T(x^{(1)} - \hat{\mu}) \\ u^T(x^{(2)} - \hat{\mu}) \\ \vdots \\ u^T(x^{(m)} - \hat{\mu}) \end{array} \right.$$

there is at least one $u^T(x^{(k)} - \hat{\mu}) \neq 0$.

as $x^{(i)} \in \mathbb{R}^D$.

in a D-dimensional space.

~~the~~ in each direction, there is one vector which is orthogonal to u.

such that if $\boxed{m > D+1}$. ~~it~~ is guaranteed that

there is at least one $u^T(x^{(k)} - \hat{\mu})$ ~~that~~ $\neq 0$, then

$u^T \hat{C} u > 0$, and $\hat{C}$ is non-singular.