Problem 2 a

$$\frac{1}{2}\sum_{k=1}^{k}\frac{1}{n_k}\sum_{C(i)=k}\sum_{C(j)=k}\|X_i - X_j\|_2^2$$

add and substract $\bar{X}_k$

$$\sum_{C(i)=k}\sum_{C(j)=k}\|(X_i - \bar{X}_k) - (X_j - \bar{X}_k)\|_2^2$$

by inner product

$$\sum_{C(i)=k}\sum_{C(j)=k}\|X_i - \bar{X}_k\|_2^2 + \|X_j - \bar{X}_k\|_2^2 - 2\langle X_i - \bar{X}_k, X_j - \bar{X}_k\rangle$$

by summation rules

$$\sum_{C(i)=k}\sum_{C(j)=k}\|X_i - \bar{X}_k\|_2^2 + \sum_{C(i)=k}\sum_{C(j)=k}\|X_j - \bar{X}_k\|_2^2 - 2\langle X_i - \bar{X}_k, X_j - \bar{X}_k\rangle$$

$$\frac{1}{2}\sum_{k=1}\frac{1}{n_k}\left[n_k\sum_{C(i)=k}\|X_i - \bar{X}_k\|_2^2 + n_k\sum_{C(j)=k}\|X_j - \bar{X}_k\|_2^2 - 0\right]$$

$$\sum_{C(i)=k} X_i - \bar{X}$$ ← because

$$= \sum_{C(i)=k} X_i - n_k\bar{X}$$

$$= n_k\bar{X} - n_k\bar{X} = 0$$

continued next page

Problem 2 a continued.

$$\frac{1}{2}\sum_{k=1}\frac{1}{n_k}\left[n_k\sum_{C(i)=k}\|X_i-\bar{X}_k\|_2^2+n_k\sum_{C(j)=k}\|X_j-\bar{X}_k\|_2^2\right]$$

$$=\frac{1}{\cancel{2}}\sum_{k=1}\frac{1}{\cancel{n_k}}\left[\cancel{2}\cancel{n_k}\sum_{C(i)=k}\|X_i-\bar{X}_k\|_2^2\right]$$

$$=\sum_{k=1}^{K}\sum_{C(i)=k}\|X_i-\bar{X}_k\|_2^2$$

Problem 2b

$$\sum_{i=1}^{m}\|Z_j-c\|_2^2$$

$$=\sum_{i=1}^{m}\left(\sqrt{(Z_1-c)^2+(Z_2-c)^2+\dots(Z_m-c)^2}\right)^2$$

$$=\sum_{i=1}^{m}(Z_1-c)^2+(Z_2-c)^2+\dots(Z_m-c)^2$$

$$\frac{\partial}{\partial c}=\sum_{i=1}^{m}2(Z_1-c)+2(Z_2-c)+\dots+2(Z_m-c)$$

$$\text{set }\frac{\partial}{\partial c}=0=\sum_{i=1}^{m}2(Z_1+Z_2+\dots Z_m)-\sum_{i=1}^{m}2mc$$

$$2m\sum_{i=1}^{m}Z_i-2m^2c=0$$

$$2m\sum_{i=1}^{m}Z_i=2m^2c$$

$$c=\frac{1}{m}\sum_{i=1}^{m}Z_i$$

Problem 2c

m_k = current centroid of cluster k

$C(i)$ = cluster assignment of i

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} \|X_i - m_k\|^2$$

$$m_k^{t+1} = \frac{1}{n_k} \sum_{C(i)=k} X_i$$ which minimizes $$\sum_{k=1}^{K} \sum_{C(i)=k} \|X_i - \bar{X}_k\|_2^2$$

$$\sum_{k=1}^{K} \sum_{C(i)=k} \|X_i - m_k^t\|_2^2 \ge \sum_{k=1}^{K} \sum_{C(i)=k} \|X_i - m_k^{t+1}\|_2^2$$

because each iteration you assign X_i to closest m_k

Problem 3d

$$d_{single}(G,H) = \min_{i \in G, j \in H} d_{ij}$$

let $d_{gh} = d_{single}(G,H)$

then

$$d_{gh} \leq d_{i_1 j_1} \leq d_{i_1 j_2} \leq \dots \leq d_{i_{n_G} j_{n_H}}$$

if $h(\cdot)$ is monotone increasing function
then if $x \leq x'$, $h(x) \leq h(x')$

$\therefore$

$$h(d_{gh}) \leq h(d_{i_1 j_1}) \leq h(d_{i_2 j_2}) \leq \dots h(d_{i_{n_G} j_{n_H}})$$

$\therefore$

points g, h still closest points

$$d_{complete}(G,H) = \max_{i \in G, j \in H} d_{ij}$$

let $d_{gh} = d_{complete}(G,H)$

$$d_{gh} \geq d_{i_1 j_1} \geq d_{i_2 j_2} \geq \dots \geq d_{i_{n_G} j_{n_H}}$$

by same monotone increasing function rule

$$h(d_{gh}) \geq h(d_{i_1 j_1}) \geq \dots \geq h(d_{i_{n_G} j_{n_H}})$$

$\therefore$ d_{gh} is till the furthest points