

1. Cluster Means Updating:

Proof: Given the distortion function:

$$SSE(z, c) = \sum_{i=1}^n ||x_i - c_{zi}||_2^2$$

Then for a single cluster, say cluster p , where p could be any integer between 1 and k in the scenario of this question, with its current center denoted as c , the above then equation becomes:

$$SSE = \sum_{x \in p} ||x - c||^2$$

That is just the sum of distance between the points to the center point assigned to them. Obviously, the above equation is minimized when $c = \text{mean}(p)$. Hence, statement is proved. We can further reinforce the proof by evaluating the cost function for the case when $c \neq \text{mean}(p)$, that is:

For any set $p \subset R^d$ and any $z \in R^d$,

$$\text{Cost}(p; c) = \text{cost}(p; \text{mean}(p)) + |p| \cdot ||c - \text{mean}(p)||^2 \quad (1)$$

Clearly, $|p| \cdot ||z - \text{mean}(p)||^2$ is positive. Hence, $c = \text{mean}(p)$ must be satisfied to minimize the cost function.

2. Cluster Assignments Updating:

Proof: Let $c_1^{(t)}, \dots, c_k^{(t)}, p_1^{(t)}, \dots, p_k^{(t)}$ denote the center and clusters at the start of the t^{th} iteration of k-means. By assigning each point to its closet cluster, we obtain the following:

$$\text{Cost}(p_1^{(t+1)}, \dots, p_k^{(t+1)}; c_1^{(t)}, \dots, c_k^{(t)}) \leq \text{Cost}(p_1^{(t)}, \dots, p_k^{(t)}; c_1^{(t)}, \dots, c_k^{(t)}) \quad (2)$$

Combines equations (1) and (2), we have:

$$\text{Cost}(p_1^{(t+1)}, \dots, p_k^{(t+1)}; c_1^{(t+1)}, \dots, c_k^{(t+1)}) \leq \text{Cost}(p_1^{(t+1)}, \dots, p_k^{(t+1)}; c_1^{(t)}, \dots, c_k^{(t)})$$

which is basically how k-means algorithm works fundamentally.