Chapter 10.7: Exercises

1.) (a.) Prove (10.12):

$$2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}\left(\frac{|C_k|x_{ij}}{|C_k|}-\frac{\sum_{i\in C_k}x_{ij}}{|C_k|}\right)^2$$

$$=2\sum_{i\in C_k}\sum_{j=1}^{p}\left(\frac{(|C_k|-1)x_{ij}-\sum_{(i\in C_k)\neq i}x_{ij}}{|C_k|}\right)^2$$

$$=2\sum_{i\in C_k}\sum_{j=1}^{p}\left(\frac{\sum_{(i\in C_k)\neq i}(x_{ij}-x_{i'j})}{|C_k|}\right)^2$$

$$=(2)\left(\frac{1}{|C_k|}\right)\left(\frac{1}{|C_k|}\right)\sum_{j=1}^{p}\Big\{(x_{ij}-x_{2j})^2+(x_{ij}-x_{2j})(x_{ij}-x_{3j})+\ldots+(x_{ij}-x_{2j})(x_{ij}-x_{|C_k|j})+\ldots+(x_{ij}-x_{2j})(x_{ij}-x_{|C_k|j})+\ldots+\ldots+(x_{|C_k|j}-x_{ij})^2+(x_{|C_k|j}-x_{ij})(x_{|C_k|j}-x_{2j})+\ldots+(x_{|C_k|j}-x_{ij})(x_{|C_k|j}-x_{|C_k|j})+\ldots\Big\}_{i=1}$$

$$=2\sum_{i\in C_k}\sum_{j=1}^{p}\left(\frac{1}{|C_k|}\right)\left(\frac{1}{|C_k|}\right)\Big[\ldots\Big]$$

$$=2\left(\frac{1}{|C_k|}\right)^2\sum_{j=1}^{p}\Big[(x_{1j}-x_{2j})^2+(x_{1j}-x_{2j})(x_{1j}-x_{3j})+\ldots+(x_{1j}-x_{2j})(x_{1j}-x_{|C_k|j})+\ldots+\Big\}_{i=1}$$

$$(x_{1j}-x_{|C_k|j})(x_{1j}-x_{2j})+\ldots.(x_{1j}-x_{|C_k|j})^2+\ldots\ldots\Big\}_{i=|C_k|}$$

$$(x_{|C_k|j}-x_{1j})^2+(x_{|C_k|j}-x_{1j})(x_{|C_k|j}-x_{2j})+\ldots+(x_{|C_k|j}-x_{1j})(x_{|C_k|j}-x_{(|C_k|-1)j})+\ldots$$

$$+(x_{|C_k|j}-x_{(|C_k|-1),j})(x_{|C_k|j}-x_{1j})+\ldots+(x_{|C_k|j}-x_{(|C_k|-1)j})^2$$

↓BACK↓

$$= (2)\left(\frac{1}{|C_k|}\right)^2 \sum_{j=1}^{p} \Big\{ \sum_{i=1} (x_{1j}^2 - 2x_{1j}x_{2j} + x_{2j}^2) + (x_{1j}^2 - x_{1j}\cdot x_{3j} - \text{[illegible]} - x_{2j}x_{1j} - x_{2j}x_{3j}) + \dots + \Big\}_{i=1}$$

$$(x_{|C_k|j}^2 - 2x_{|C_k|j}\cdot x_{1j} + x_{1j}^2) + (x_{|C_k|j}^2 - x_{|C_k|j}\cdot x_{2j} - x_{1j}\cdot x_{|C_k|j} + x_{2j}x_{ij}) + \dots \Big\}_{i=|C_k|}$$

$$= 2\left(\frac{1}{|C_k|}\right)\left(\frac{1}{|C_k|}\right) \sum_{j=1}^{p} (|C_k|)(|C_k|-1)(x_{1j}^2 + x_{2j}^2 + \dots + x_{|C_k|j}^2 - x_{1j}x_{2j} - x_{1j}x_{3j} - \dots - x_{1j}x_{|C_k|j} - \dots - x_{|C_k|-1,j}\cdot x_{|C_k|j})$$

is used $(|C_k|-1)$ times, etc.

$$= \left(\frac{1}{|C_k|}\right) \sum_{j=1}^{p} (2)\Big[(x_{1j}^2 - 2x_{1j}x_{2j} + x_{2j}^2) + (x_{1j}^2 - 2x_{1j}x_{3j} + x_{3j}^2) + \dots + (x_{11}^2 - 2x_{11}x_{|C_k|1} + x_{|C_k|1}^2) + \dots + \dots\Big]$$

$$= \left(\frac{1}{|C_k|}\right) \sum_{j=1}^{p} (2)\Big[\Big\{(x_{1j} - x_{2j})^2 + \dots + (x_{1j} - x_{|C_k|j})^2 + \dots\Big\}_{i=1} \quad (i-1) \text{ terms}$$

$$\Big\{(x_{2j} - x_{3j})^2 + \dots + (x_{2j} - x_{|C_k|j})^2 + \dots\Big\}_{i=2} \quad (i-2) \text{ terms}$$

$$\vdots$$

$$+ (x_{(|C_k|-1)j} - x_{|C_k|j})^2 \quad 1 \text{ term}$$

$$= \left(\frac{1}{|C_k|}\right) \sum_{j=1}^{p} \sum_{i,i' \in C_k} (x_{ij} - x_{i'j})^2$$

$$\boxed{= \left(\frac{1}{|C_k|}\right) \sum_{i,i' \in C_k} \sum_{j=1} (x_{ij} - x_{i'j})^2} \quad \blacksquare$$

(b) For any iteration of Algorithm 10.1, we begin the iteration with a current assignment.

Then:

* Per Algorithm 10.1 Step 2, Part (a), we compute the centroid for each cluster.
* Given, the above centroids, we then compute γ = the sum of the within-cluster variations, over all clusters.
* Then, per Algorithm 10.1 Step 2, Part (b), we assign each observation to the cluster which is closest to it.
* Thus, after this latest assignment, each observation will either be closer to (or the same distance from) its new cluster centroid. Thus, this either reduces (or keeps the same, if algo has converged) the quantity:

$$\sum_{k=1}^{K} 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

* Then, by identity (10.12) which I proved in (a), this either reduces or keeps constant the quantity:

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

which is precisely the objective (10.11) we wish to minimize.

Thus I have shown that K-Means clustering algorithm decreases (or keeps constant) objective (10.11) at each iteration.