K-Means Image Compression Algorithm Explanation

Given m data points x^i, $i = 1, \ldots, m$, K-means clustering algorithm groups them into k clusters by minimizing the distortion function over $\{r^{ij}, \mu^j\}$

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \|\mathrm{x}^i - \mu^j\|^2, \tag{1}$$

where $r^{ij} = 1$ if x^i belongs to the j-th cluster and $r^{ij} = 0$ otherwise.

1. Mathematical Derivation of using the squared Euclidean distance $\|\mathrm{x}^i - \mu^j\|^2$ as the dissimilarity function, the centroid that minimizes the distortion function J for given assignments r^{ij} are given by

   $$\mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}.$$

   That is, μ^j is the center of j-th cluster.

   Given m data points with dimension n,

   $$M = (x^1, x^2 ... x^m), x^i \in R^n$$

   Assume K clusters with centers:

   $$\mu^1, \mu^2, ... \mu^k, \mu^j \in R^n$$

   Goal is to minimize the distortion function, J: squared sum of difference in distance of each x^i from its cluster center, μ^j. In order to do this we want to chose μ^i to minimize J.
   Take the partial derivative of J w.r.t μ^j and set it to 0.

   $$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \left\| x^i - \mu^j \right\|^2$$

   $$= \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \left(x^i - \mu^j\right)^T \left(x^i - \mu^j\right)$$

   $$\frac{\nabla J}{\nabla \mu^k} = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \nabla \left(x^i - \mu^j\right)^T \left(x_i - \mu^j\right) = 0$$

   $$\nabla \left(x^{iT} x^i - 2\mu^{jT} x^i + \mu_j^T \mu_j\right) = -2x^i + 2\mu^j$$

   $$-2 \sum_{i=1}^{m} r^{ij} x^i + 2\mu^j \sum_{i=1}^{n} r^{ij} = 0$$

   $$\mu^j = \frac{\sum r^{ij} x^i}{\sum r^{ij}}$$

2. Mathematical derivation of what should be the assignment variables r^{ij} be to minimize the distortion function J, when the centroids μ^j are fixed.

   iter = 1
   While iter <= max_iter or convergence hasn't happened(centeroid value in successive iterations keeps changing):
   3) For each datapoint in dataframe:
   - Find the nearest c^j
   - Assign the datapoint to that cluster
   4) For each cluster j: range(1,k):
   - new c^j = mean of all points that have been assigned to that cluster.

3. Why K-means algorithm converges to a local optimum in finite steps.

   There are a finite number of possible clustering for any number of data points N, and clusters K. The possible clustering configurations is K^N. As such, there are only a finite number of clustering arrangements, <= K^N that can be looped through till convergence.
   For each iteration, either:
   1) The new cluster will be the same as the old cluster and the clustering will stop.
   2) The new clustering will produce a lower distortion function than the previous cluster.
   Since the algorithm stops when a clustering is repeated as per 1), clustering arrangements are not repeated when looping through different clustering from the set of K^N arrangements. Since the clustering cannot be repeated, there are only a finite number of clustering arrangements that can be iterated over in order to find the local minimum when the algorithm stops.