Theoretical task 10

Recommendations: all solutions should be short, mathematically strict (unless qualitative explanation is needed), precise with respect to the stated question and clearly written. Solutions may be submitted in any readable format, including images.

1. *How would the steps of K-Means algorithm change, if in minimization criterion the euclidean distance is replaced by Manhattan distance (L1 distance)? What about its computational complexity?*

2. (a) *Show that K-Means criterion and the following criterion are equivalent:*

$$Q(C) = \sum_k \sum_{x_i, x_j \in C_k} \frac{x_i^\top x_j}{|C_k|},$$

   where C_k is cluster with label k

   (b) *Given that result, what technique can be used in K-Means algorithm?*

3. *Consider Lance-Williams Formula for hierarchical clustering:*

$$\rho(C_i \cup C_j, C_k) = a_i \cdot \rho(C_i, C_k) + a_j \cdot \rho(C_j, C_k) + b \cdot \rho(C_i, C_j) + c \cdot |\rho(C_i, C_k) - \rho(C_j, C_k)|$$

   where $\rho(\cdot, \cdot)$ is some distance between clusters. Show that

   (a) *for single linkage $a_i = a_j = \frac{1}{2}$, $b = 0$ and $c = -\frac{1}{2}$*

   (b) *for complete linkage $a_i = a_j = \frac{1}{2}$, $b = 0$ and $c = \frac{1}{2}$*

   (c) *for average linkage $a_i = \frac{|C_i|}{|C_i|+|C_j|}$, $a_j = \frac{|C_j|}{|C_i|+|C_j|}$ and $b = c = 0$*

4. *Provide an example of objects in 2d and their partition on 2 clusters such that*

   - *silhouette score is rather poor*
   - *based on human judgement, partition has clearly revealed shaped of clusters*

   In other words, show that silhouette score may provide misleading values for arbitrary-shaped clusters (for instance after density based clustering)