CS550: Massive Data Mining and Learning Spring 2022
Problem Set 4
Due 11:59pm Friday, Apr 29, 2022

Submission Instructions

Honor Code: Students may discuss the homework problems with peers. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students with whom they have discussed the homework problems. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

I acknowledge and accept the Honor Code.

(Signed)____Vaishnavi Manthena_________

If you are not printing this document out, please type your initials above.

Answer to Question 1

To prove: $\boldsymbol{cost(S,T) \leq 2 * cost_w(\widehat{S}\,,T) + 2\sum_{i=1}^{l} cost(S_i,T_i)}$

$cost(S,T) = \sum_{x\in S} d(x,T)^2 = \sum_{x\in S_1} d(x,T)^2 + \sum_{x\in S_2} d(x,T)^2 + \cdots + \sum_{x\in S_l} d(x,T)^2 = \sum_{i=1}^{l}\sum_{x\in S_i} d(x,T)^2$

The above was possible, because:
$the\ union\ of\ all\ S_i's\ is\ the\ total\ dataset\ S, and\ their\ intersection\ is\ null.$

We know that $d(x,T) = \min_{z\in T}\{d(x,z)\}$.

Substituting this to above we get:

$$cost(S,T) = \sum_{i=1}^{l}\sum_{x\in S_i}\left(\min_{z\in T}\{d(x,z)\}\right)^2$$

From the algorithm we know that each point 'x' of S_i is assigned to a cluster with center t_{ij}. Let 'c' denote this center of the cluster that 'x' in S_i is assigned to. Now, since 'c,' 'x,' and 'z' are just points in the Euclidean space and $d(a,b)$ is the Euclidean distance, according the triangle inequality the following statement holds: $d(x,z) \leq d(x,c) + d(c,z)$. Using this:

$$cost(S,T) \leq \sum_{i=1}^{l}\sum_{x\in S_i}\left(\min_{z\in T}\{d(x,c) + d(c,z)\}\right)^2$$

$d(x,c)$ is independent of 'z,' so:

$$cost(S,T) \leq \sum_{i=1}^{l}\sum_{x\in S_i}\left(d(x,c) + \min_{z\in T}\{d(c,z)\}\right)^2$$

From the given hint [$(a+b)^2 \leq 2a^2 + 2b^2$]:

$$cost(S,T) \leq \sum_{i=1}^{l}\sum_{x\in S_i} 2 * d(x,c)^2 + 2 * \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

We get the below inequality by splitting the summation.
Brief intuition about RHS of the first inequality below:

- The 1st term is basically summing up for every 'x' value in S, 2 times the square of Euclidean distance between this point and the center it was assigned to while it participated in a clustering.
- The 2nd term is basically summing up for every 'x' value in S, 2 times the square of Euclidean distance between the center it was assigned to while it participated in a clustering and the closest point in the final result T to this center.

$$cost(S,T) \leq \sum_{i=1}^{l}\sum_{x\in S_i} 2 * d(x,c)^2 + \sum_{i=1}^{l}\sum_{x\in S_i} 2 * \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

$$= 2 * \sum_{i=1}^{l}\sum_{x\in S_i} d(x,c)^2 + 2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

$$= 2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{\mathrm{r\in T_i}} d(x,r)\right)^2 + 2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

$$= 2 * \sum_{i=1}^{l}\sum_{x\in S_i} d(x,T_i)^2 + 2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

$$= 2 * \sum_{i=1}^{l} cost(S_i,T_i) + 2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{z\in T}\{d(c,z)\}\right)^2$$

Simplifying the second term on RHS:

Note: c is the centroid of the cluster that ‘x’ is assigned to while clustering the subset of data that ‘x’ belongs to.

$$2 * \sum_{i=1}^{l}\sum_{x\in S_i} \left(\min_{z\in T}\{d(c,z)\}\right)^2 = 2 * \sum_{i=1}^{l}\left(\sum_{j=1}^{k}\sum_{x\in S_{ij}} \left(\min_{z\in T}\{d(c,z)\}\right)^2\right)$$

$$= 2 * \sum_{i=1}^{l}\left(\sum_{j=1}^{k}\sum_{x\in S_{ij}} d(c,T)^2\right)$$

Since each ‘x’ belonging to the same cluster has the same center ‘c’:

$$= 2 * \sum_{i=1}^{l}\left(\sum_{j=1}^{k}\left|S_{ij}\right| d(c,T)^2\right)$$

Writing the above in terms of the summation over all centers:

$$= 2 * \sum_{c\in\hat{\mathbf{S}}}\left|S_{ij}\right| d(c,T)^2 = 2 * cost_w\left(\hat{S}\,,T\right)$$

Therefore,

$cost(S,T) \leq 2 * cost_w\left(\hat{S}\,,T\right) + 2 * \sum_{i=1}^{l} cost(S_i,T_i)$ **-> EQUATION 1**

Answer to Question 2

$\sum_{i=1}^{l} cost(S_i, T_i) \leq \sum_{i=1}^{l} \alpha * cost(S_i, T_i^*) = \alpha * \sum_{i=1}^{l} cost(S_i, T_i^*)$

The above equation is due to the given output guarantee of ALG. T_i^* is the actual optimal clustering centers for S_i. For S_i, since T_i^* is optimal, $cost(S_i, T_i^*) \leq cost(S_i, T^*)$, since T^* may not equal to T_i^*. So:

$\alpha * \sum_{i=1}^{l} cost(S_i, T_i^*) \leq \alpha * \sum_{i=1}^{l} cost(S_i, T^*) = \alpha * \sum_{i=1}^{l} \sum_{x \in S_i} d(x, T^*)^2$

Now since, $the\ union\ of\ all\ S_i^{\prime}s\ is\ the\ total\ dataset\ S, and\ their\ intersection\ is\ null$:

$\alpha * \sum_{i=1}^{l} \sum_{x \in S_i} d(x, T^*)^2 = \alpha * \sum_{x \in S} d(x, T^*)^2 = \alpha * cost(S, T^*)$

Summary of all steps:
$\sum_{i=1}^{l} cost(S_i, T_i) \leq \alpha * \sum_{i=1}^{l} cost(S_i, T_i^*) \leq \alpha * \sum_{i=1}^{l} \sum_{x \in S_i} d(x, T^*)^2 = \alpha * cost(S, T^*)$

Conclusion:
$\sum_{i=1}^{l} cost(S_i, T_i) \leq \alpha * cost(S, T^*)$ **-> EQUATION 2**

Answer to Question 3

Proof of $\boldsymbol{cost_w(\hat{S}, T) \leq \alpha * cost_w(\hat{S}, T^*)}$:

Notation:
$T^+: optimal\ clustering\ solution\ for\ \hat{S}$
$T^*: optimal\ clustering\ solution\ for\ S$

According to the property of ALG:
$cost_w(\hat{S}, T) \leq \alpha * cost_w(\hat{S}, T^+)$

Now, since T^* may not equal T^+: $cost_w(\hat{S}, T^+) \leq cost_w(\hat{S}, T^*)$.

$cost_w(\hat{S}, T) \leq \alpha * cost_w(\hat{S}, T^+) \leq \alpha * cost_w(\hat{S}, T^*)$

Therefore, $cost_w(\hat{S}, T) \leq \alpha * cost_w(\hat{S}, T^*)$ **-> EQUATION 3**

Proof of $\boldsymbol{cost_w(\hat{S},T^*) \leq 2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost(S,T^*)}$:

In the algorithm consider a particular subset S_i and a particular clustering S_{ij} by running ALG on it. This clustering is associated with a center t_{ij} and some data points of S_i that fall into this cluster. Consider one such datapoint 'x' in S_{ij}.

Consider the quantity $d(t_{ij},T^*)^2$. Intuitively we are analyzing this quantity since this is an integral part of the LHS of inequality to prove. By using the triangle inequality and the hint ($(a+b)^2 \leq 2a^2 + 2b^2$):

$$d(t_{ij},T^*)^2 = \left(\min_{z \in T^*}\{d(t_{ij},z)\}\right)^2 \leq \left(\min_{z \in T^*}\{d(t_{ij},x) + d(x,z)\}\right)^2 = \left(d(t_{ij},x) + \min_{z \in T^*}\{d(x,z)\}\right)^2 \\ \leq 2d(t_{ij},x)^2 + 2d(x,T^*)^2$$

Main result: $d(t_{ij},T^*)^2 \leq 2d(t_{ij},x)^2 + 2d(x,T^*)^2$

The above inequality holds for any arbitrary instance (i, j, values and any 'x' within that cluster). So, summating the components of the above inequality over all instances will still result in a valid inequality.

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(t_{ij},T^*)^2 \leq 2 * \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(t_{ij},x)^2 + 2 * \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(x,T^*)^2$$

<u>Simplifying LHS:</u>
Each $x \epsilon S_{ij}$ has the same center t_{ij}.
$\sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(t_{ij},T^*)^2 = \sum_{i=1}^{l}\sum_{j=1}^{k} |S_{ij}| d(t_{ij},T^*)^2 = \sum_{t_{ij} \in \hat{S}} |S_{ij}| d(t_{ij},T^*)^2 = cost_w(\hat{S},T^*)$

<u>Simplifying first term of RHS:</u>
$2 * \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(t_{ij},x)^2 = 2 * \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(x,T_i)^2 = 2 * \sum_{i=1}^{l}\sum_{x \epsilon S_i} d(x,T_i)^2 = 2\sum_{i=1}^{l} cost(S_i,T_i)$

<u>Simplifying second term of RHS:</u>
$2 * \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x \epsilon S_{ij}} d(x,T^*)^2 = 2 * \sum_{i=1}^{l}\sum_{x \epsilon S_i} d(x,T^*)^2 = 2 * \sum_{x \epsilon S} d(x,T^*)^2 = 2cost(S,T^*)$

Therefore, $cost_w(\hat{S},T^*) \leq 2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost(S,T^*)$ **-> EQUATION 4**

Proof of Final Result

From ‘Equation 1:’

$$cost(S,T) \leq\ 2 * cost_w\left(\hat{S}\ ,T\right) + 2 * \sum_{i=1}^{l} cost(S_i,T_i)$$

From ‘Equation 2,’ using the upper bound for the second term on the RHS of above equation:

$$cost(S,T) \leq\ 2 * cost_w\left(\hat{S}\ ,T\right) + 2 * \alpha * cost(S,T^*)$$

From ‘Equation 3,’ using the upper bound for the first term on the RHS of above equation:

$$cost(S,T) \leq\ 2 * \alpha * cost_w\left(\hat{S},T^*\right) + 2 * \alpha * cost(S,T^*)$$

From ‘Equation 4’, using upper bound on $cost_w\left(\hat{S},T^*\right)$ gives:

$$\begin{aligned} cost(S,T) \leq&\ 2 * \alpha * (2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost(S,T^*)) + 2 * \alpha * cost(S,T^*) \\ &= 4\alpha\left(\sum_{i=1}^{l} cost(S_i,T_i) + cost(S,T^*)\right) +\ 2 * \alpha * cost(S,T^*) \end{aligned}$$

Using ‘Equation 2’ again:
$cost(S,T) \leq 4\alpha\left(\alpha * cost(S,T^*) + cost(S,T^*)\right) +\ 2\alpha cost(S,T^*) = 4\alpha^2 cost(S,T^*) + 6\alpha cost(S,T^*)$
Therefore, $\boldsymbol{cost(S,T)} \leq \left(\boldsymbol{4\alpha^2} + \boldsymbol{6\alpha}\right) * \boldsymbol{cost(S,T^*)}$.