CS550: Massive Data Mining and Learning
Homework 4

Due 11:59pm Wednesday, May 8, 2019
Only one late period is allowed for this homework (11:59pm Thursday May 9)

Submission Instructions

Assignment Submission Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

Late Day Policy Each student will have a total of *two* free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

Honor Code Students may discuss and work on homework problems in groups. This is encouraged. 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 who were part of their discussion group. 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):

On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Honor Code.

*(Signed)Akhilesh Harish Mahajan*_____

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

Answer to Question 1

We need to prove that:
$cost(S,T) \leq 2cost_w(S,T) + 2\sum_{i=1}^{l} cost(S_i,T_i)$
Answer:
Let $T(x) = argmin_{t \in T} d(t,x)$
By the triangle inequality, that is, any side is less than or equal to the sum of two other sides.
$\therefore$ For any $x \in S_{ij}, (1 \leq i \leq l, 1 \leq j \leq k)$:
$d(x,T) = d(x,T(x))$
$\leq d(x,T(t_{ij}) + d(t_{ij},T)$
$= d(x,t_{ij}) + d(t_{ij},T)$
and since $(a+b)^2 \leq 2a^2 + 2b^2$
$d(x,T)^2 \leq 2d(x,t_{ij})^2 + 2d(t_{ij},T)^2$
Summing the result over all ij, x gives the result.
Hence proved!

Answer to Question 2

To bound the cost of final clustering on the left, since it is less than the right summation, bounding right hand side terms would be enough.
We assume here that T_i^* is the optimal clustering for $S_i (1 \leq i \leq l)$. Then, cost $(S_i,T_i) \leq \alpha cost(S_i,T_i^*) \leq \alpha cost(S_i,T^*)$.
$\hat{S}$ is union of T_i from 1 to l, and we run algorithm on $\hat{S}$ with w weights to get centers T_i and we return T.
$\therefore$ We can say that $\alpha cost(S_i,T_i^*) \leq \alpha cost(S_i,T^*)$.
Summing up over i gives the result.
If we see each term, the subroutine guarantees that
$cost(S_i,T_i) \leq \alpha cost(S_i,T_i^*) \leq \alpha cost(S_i,T^*)$
where T_i^* is the globally optimum assignment of the cluster centers given S_i. The inequality follows the assumption that the ALG routine returns a set T_i which is an α-approximate of T_i^*. The second inequality follows since T_i^* is the optimal clustering for S_i (must have cost lower than any other T' including T^*).
$\therefore \sum_{i=1}^{l} cost(S_i,T_i) \leq \alpha \sum_{i=1}^{l} cost(S_i,T_i^*) = \alpha cost(S,T^*)$.
The final equality uses the fact that S is $U_{i=1}^{l} S_i$ to collect mulitple summations over S_i over a single sum over S (in the cost function definition).

Answer to Question 3

Let T^* denote the optimal solution for $\hat{S}$.
As T^* must have a larger cost with respect to S (weight w) than $\hat{T}^*$, as $\hat{T}^*$ is the global optimum.

$cost_w(\hat{S}, T) \leq \alpha cost_w(\hat{S}, \hat{T^*}) \leq \alpha cost_w(\hat{S}, T*)$
Now, $cost(S, T) \leq 2cost_w(\hat{S}, T) + 2\sum_{i=1}^{l} cost(S_i, T_i)$
$\leq 2cost_w(\hat{S}, T) + 2\alpha cost(S, T^*) \ldots$ from question 1 and 2.
$\leq 2\alpha cost_w(\hat{S}, T^*) + 2\alpha cost(S, T^*)$
$\leq 2\alpha[2\alpha cost(S, T^*) + 2cost(S, T^*)] + 2\alpha cost(S, T^*)$
$(openingbrackets)$
$\leq 4\alpha^2 cost(S, T^*) + (4\alpha + 2\alpha)[cost(S, T^*)]$
$\leq 4\alpha^2 cost(S, T^*) + 6\alpha cost(S, T^*)$
$\leq (4\alpha^2 + 6\alpha) cost(S, T^*)$