CS550: Massive Data Mining and Learning
Homework 4

Due 11:59pm Friday, Apr 23, 2021

Only one late period is allowed for this homework (11:59pm Apr 24)

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 Canvas. 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):
Rohan Shah (ras513)

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

I acknowledge and accept the Honor Code.

(Signed) RG

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

Answer to Question 1

The term on the left-hand side of the inequality can be rewritten as the following:

$$\text{cost}(S,T) = \sum_{x\in S} d^2(x,T)$$

$\hat{S}$ is simply the union of all T_i and therefore the union of all t_{ij}. Since all $x \in S$ have a corresponding center t_{ij}, and the sum of all $w(t_{ij}) = |S_{ij}|$ is equivalent to $|S|$, the implied summation in the first term of the right-hand side of the inequality can be expanded to iterate over all $x \in S$ instead of all $t_{ij} \in \hat{S}$. Therefore, this term can be rewritten as the following:

$$2\cdot \text{cost}_w(\hat{S},T) = 2\sum_{t_{ij}\in \hat{S}} w(t_{ij})d^2(t_{ij},T) = 2\sum_{t_{ij}\in \hat{S}} |S_{ij}|d^2(t_{ij},T) = 2\sum_{x\in S} d^2(t_{ij},T)$$

For any given S_i, all $x \in S_i$ correspond to a single center $t_{ij} \in T_i$ because the distance between x and t_{ij} is less than the distance between x and any other center $t'_{ij} \in T_i$. In other words, $d(x,T_i) = d(x,t_{ij})$. This can be expanded to all $x \in S$, since every value in S will have a corresponding t_{ij}. Therefore, the second term on the right-hand side of the inequality can be rewritten as the following:

$$2\sum_{i=1}^{l} \text{cost}(S_i,T_i) = 2\sum_{i=1}^{l}\sum_{x\in S_i} d^2(x,T_i) = 2\sum_{x\in S} d^2(x,t_{ij})$$

The summation expression of the final rewritten expressions of the two terms on the right-hand side of the goal inequality can be extracted and applied to the suggested inequality:

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

$d(x,T)$, or the least distance from x to a center of T, is less than or equal to the sum of the distance from x to its corresponding t_{ij} and the distance from this t_{ij} to the closest center of T. In other words, $d(x,T) \leq d(x,t_{ij}) + d(t_{ij},T)$. The above inequality can therefore be simplified:

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

The terms in this inequality correspond to each of the terms in the goal inequality based on the above rewritten expressions. Since all of these expressions sum over all $x \in S$, the above inequality applies to all of S. Therefore:

$$\sum_{x\in S} d^2(x,T) \leq 2\sum_{x\in S} d^2(t_{ij},T) + 2\sum_{x\in S} d^2(x,t_{ij})$$

$$\therefore \text{cost}(S,T) \leq 2\cdot \text{cost}_w(\hat{S},T) + 2\sum_{i=1}^{l} \text{cost}(S_i,T_i)$$

Answer to Question 2

Step 2 of ALGSTR utilizes ALG for all S_i partitions of S. The given condition of the returned set of ALG can be applied for all $|T'| = k$:

$$\text{cost}(S_i, T_i) \leq \alpha \cdot \text{cost}(S_i, T')$$

T^* is included in the set of all possible T's. Therefore, the above inequality's terms can be expressed in the following inequality of summations:

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

As proven in **Question 1**, the term on the right-hand side of the above inequality can be rewritten as follows:

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

Note that the final expression is slightly different than what was proven in **Question 1** but still accurate, since T^* is the same for all S_i partitions of S. This final expression is also equivalent to $\alpha \cdot \text{cost}(S, T^*)$. Therefore:

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

Answer to Question 3

To prove the first useful fact, we must consider the fact that T is produced as a result of using ALG on $\hat{S}$. Therefore, for all $|T'| = k$, including the optimal T^*:

$$\text{cost}(\hat{S}, T) \leq \alpha \cdot \text{cost}(\hat{S}, T')$$

To prove the second useful fact, we can simply reuse much of the logic from the proof in **Question 1**. Based on the rewritten versions of the terms in **Question 1**, the terms on the right-hand side of the given inequality for this fact can be rewritten as follows:

$$2 \sum_{i=1}^{l} \text{cost}(S_i, T_i) = 2 \sum_{x \in S} d^2(x, t_{ij}) \qquad \text{cost}(S, T^*) = \sum_{x \in S} d^2(x, T^*)$$

The inner functions of the summations on the right-hand sides of these equations can be extracted, and the given hint inequality from **Question 1** can be applied as follows:

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

As per the triangle inequality, $d(t_{ij}, T^*) \leq d(x, t_{ij}) + d(x, T^*)$. Additionally, again pulling from **Question 1**, the inner function that results from rewriting $\text{cost}_w(\hat{S}, T)$ is $d^2(t_{ij}, T^*)$. Therefore, inserting these functions back into their summations results in the following inequality:

$$\sum_{x \in S} d^2(t_{ij}, T^*) \leq 2 \sum_{x \in S} d^2(x, t_{ij}) + 2 \sum_{x \in S} d^2(x, T^*)$$

$$\therefore \text{cost}_w(\hat{S}, T^*) \leq 2 \sum_{i=1}^{l} \text{cost}(S_i, T_i) + \text{cost}(S, T^*)$$

For this proof, let us re-use the given inequality from **Question 1**:

$$\text{cost}(S, T) \leq 2 \cdot \text{cost}_w(\hat{S}, T) + 2 \sum_{i=1}^{l} \text{cost}(S_i, T_i)$$

As per the first fact, $2 \cdot \text{cost}_w(\hat{S}, T) \leq 2\alpha \text{cost}_w(\hat{S}, T^*)$. Applying the second fact results in the following:

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

The term on the right-hand side of this inequality can be expressed as the following:

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

The proof from **Question 2** can be used to justify the following:

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

The proof from **Question 2** can also be applied to the second term on the right side of the inequality from **Question 1** as follows:

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

Therefore, the inequality from **Question 1** can be rewritten as follows:

$$\text{cost}(S, T) \leq (4\alpha^2 + 4\alpha) \text{cost}(S, T^*) + 2\alpha \text{cost}(S, T^*)$$

$$\therefore \text{cost}(S, T) \leq (4\alpha^2 + 6\alpha) \text{cost}(S, T^*)$$