CS550: Massive Data Mining and Learning
Homework 4

Due 11:59pm Friday, Apr 29, 2022

Submission Instructions

Honor Code Students may discuss 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)SK ______________________________

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

Answer to Question 1

$$LHS = cost(S, T)$$

As we know that,

$$S = S_1 \cup S_2 \cup S_3 .... \cup S_l$$

$$\begin{aligned} cost(S, T) &= \sum_{x \in S} d(x, T)^2 \\ &= \sum_{i=1}^{l} \sum_{x \in S_i} d(x, T)^2 \\ &= \sum_{i=1}^{l} \sum_{x \in S_i} [\min_{z \in T} [d(x, z)]]^2 \end{aligned} \tag{1}$$

Because of triangle inequality, we have,

$$d(x, z) \leq d(x, y) + d(y, z) \tag{2}$$

Therefore,

$$\begin{aligned} \min_{z \in T} [d(x, z)] &\leq \min_{z \in T} [d(x, y) + d(y, z)] \\ &\leq d(x, y) + \min_{z \in T} [d(y, z)] \end{aligned} \tag{3}$$

Putting eq(3) in eq.(1), we have,

$$cost(S, T) \leq \sum_{i=1}^{l} \sum_{x \in S_i} [d(x, y) + \min_{z \in T} [d(y, z)]]^2 \tag{4}$$

Applying inequality, $(a + b)^2 <= 2a^2 + 2b^2$,

$$\begin{aligned} cost(S, T) &\leq 2 \sum_{i=1}^{l} \sum_{x \in S_i} [d(x, y)]^2 + 2 \sum_{i=1}^{l} \sum_{x \in S_i} \min_{z \in T} [d(y, z)]^2 \\ &\leq 2 \sum_{i=1}^{l} \sum_{x \in S_i} [d(x, y)]^2 + 2 \sum_{i=1}^{l} \sum_{x \in S_i} d(y, T)^2 \end{aligned} \tag{5}$$

Here, for every $x \in S_i$, let $y = t_{ij}$. This implies that y is the centroid that $x \in S_i$ is assigned to. Therefore it follows that,

$$\sum_{x \in S_i} d(x, y)^2 = \sum_{x \in S_i} d(x, T_i)^2 = cost(S_i, T_i)$$

We know that y takes the values in $\hat{S} = t_{ij}$, and the number of times y takes a particular outcome t_{ij} is proportional to the no. of times $x \in S_i$ is assigned to cluster center t_{ij}.
So Second term becomes,

$$\begin{aligned}\sum_{i=1}^{l} \sum_{x \in S_i} d(y,T)^2 &= \sum_{y \in \hat{S}} |S_{ij}|.d(y,T)^2 \\ &= cost_w(\hat{S}, T)\end{aligned}$$

Substituting these two values in equation, we get,

$$cost(S,T) \leq 2cost_w(\hat{S}, T) + 2\sum_{i=1}^{l} cost(S_i, T_i) \tag{6}$$

Hence proved.

Answer to Question 2

The question describes an algorithm ALG which guarantees an upper bound such that for each individual term $cost(S_i, T_i)$,

$$cost(S_i, T_i) \leq \alpha cost(S_i, T_i^*) \leq \alpha cost(S_i, T^*)$$

where T_i^* is the optimal clustering for $S_i (1 \leq i \leq l)$

The first of the inequality arises from the fact that the algorithm ALG returns a set T_i that is α -approximate of T_i^*. The second of the inequality stems from the fact that the optimal clustering set for S_i is T_i. Therefore, it must have a cost that is lower than any other candidate T' including T^*.
Summing over i,

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

As we know that,

$$S = S_1 \cup S_2 \cup S_3 .... \cup S_l$$

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

Hence proved.

Answer to Question 3

Consider the following facts,
Fact 1: Let $\hat{T}^*$ be the optimum clustering for the subset $\hat{S}$.

$$
\begin{aligned}
const_w(\hat{S}, T) &\leq \alpha const_w(\hat{S}, \hat{T}^*) \\
&\leq \alpha const_w(\hat{S}, T^*) \qquad (7)
\end{aligned}
$$

Fact 2: For any $x \in S_{ij}$, where 1 ¡= i ¡= l, l¡= j ¡= k.

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

Summing over all values of i, j and x,

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

From Question 2,

$$const_w(\hat{S}, T^*) \leq 2\alpha cost(S, T^*) + 2cost(S, T^*) \qquad (8)$$

From eq.(6), we have,

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

From Question 2, we can rewrite this as,

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

From eq.(7), we can rewrite this as,

$$cost(S, T) \leq 2\alpha const_w(\hat{S}, T^*) + 2\alpha cost(S, T^*) \qquad (9)$$

Now using eq(8) and eq(9),

$$
\begin{aligned}
&cost(S, T) \leq 2\alpha[2\alpha const_w(S, T^*) + 2cost(S, T^*] + 2cost(S, T^*) \\
&cost(S, T) \leq (4\alpha^2 + 6\alpha)cost(S, T^*)
\end{aligned}
$$

Hence proved.