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)____KK____

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

Answer to Question 1

To prove:

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

Given: S = $S_1 \cup S_2 \cup \ . \ . \ . \ S_l$

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

$$= \sum_{i=1}^{l} \sum_{x \epsilon S_i} d(x,T)^2$$

$$= \sum_{i=1}^{l} \sum_{x \epsilon S_i} [min_{z \epsilon T}[d(x, z)]]^2 \qquad \text{– eq(1. 1)}$$

Using the triangle inequality, we get:

$$d(x,z) \leq d(x,y) + d(y,z)$$

so , for $min_{z \epsilon T} d(x, z)$:

$$min_{z \epsilon T} d(x, z) \leq min_{z \epsilon T}[d(x, y) + d(y,z)]$$

$$\leq d(x,y) + min_{z \epsilon T} d(y, z) \qquad \text{– eq(1. 2)}$$

Substituting eq(1. 2) in eq(1. 1), we get:

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

Applying the inequality,

$(a + b)^2 \leq 2a^2 + 2b^2$ in above Equation:

$$\text{cost(S,T)} \leq 2 \sum_{i=1}^{l} \sum_{x \epsilon S_i} d(x,y)^2 + 2 \sum_{i=1}^{l} \sum_{x \epsilon S_i} min_{z \epsilon T} d(y,z)^2$$

$$\leq 2 \sum_{i=1}^{l} \sum_{x \epsilon S_i} d(x,y)^2 + 2 \sum_{i=1}^{l} \sum_{x \epsilon S_i} d(y, T)^2 \qquad \text{– eq(1. 3)}$$

For every $x \epsilon S_i$, let $y = t_{ij}$. So, y will be the centroid that $x \epsilon S_i$ is assigned to, during the clustering.

Hence we get,

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

For the second term in eq(1.3), y takes values in $\widehat{S} = t_{ij}$ and the number of times that y takes a particular outcome t_{ij} is proportional to the number of times $x \epsilon S_i$ is assigned to the cluster center t_{ij}.

So,

$$\sum_{i=1}^{l} \sum_{x \epsilon S_i} d(y, T)^2 = \sum_{y \epsilon S} \left|S_{ij}\right| \cdot d(y, T)^2 = cost_w(\widehat{S}, T)$$

Substituting above results in eq(1.3), we get:

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

$$\Rightarrow cost(S,T) \le 2 \cdot \sum_{i=1}^{l} cost(S_i, T_i) + 2\, cost_w(\widehat{S}, T) \qquad \text{– eq(1.4)}$$

Hence proved.

Answer to Question 2

Task: To Prove:

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

Algorithm ALG guarantees an upper bound for each term $cost\left(S_i, T_i\right)$ like so:

$$cost\left(S_i, T_i\right) \le \alpha cost(S_i, T_i^*) \le \alpha cost(S_i, T\ ^*)$$

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

The second term of the inequality $\alpha cost(S_i, T_i^*)$ is deduced from the following logic: Algorithm ALG returns T_i that is α-approximate of T_i^*

The third term $\alpha cost(S_i, T\ ^*)$ is deduced from: T_i is the optimal clustering set for S_i. So it must have a cost that is lower than any other clustering set T' including $T\ ^*$

Applying summation over i to first and third term, we get:

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

Since $\sum_{i=1}^{l} S_i = S$

We get, $\sum_{i=1}^{l} cost\left(S_i, T_i\right) \le \alpha cost(S, T\ ^*)$ – eq(2.1)

Hence proved.

Answer to Question 3

Task: Prove the following:

$$cost(S,T) \le (4\alpha^2 + 6\alpha).\, cost(S, T\ ^*)$$

We first prove the two facts provided in the homework to do the actual proof.

Proof a: To prove: $cost_\omega(\widehat{S}, T) \le \alpha.\, cost_\omega(\widehat{S}, T\ ^*)$

Let $\widehat{T}\ ^*$ be the optimum clustering for $\widehat{S}$

=> $cost_{\omega}(\hat{S}, T) \leq \alpha. cost_{\omega}(\hat{S}, \hat{T} *)$

=> $cost_{\omega}(\hat{S}, T) \leq \alpha. cost_{\omega}(\hat{S}, T *)$ –eq(3.1)

Hence Proved Proof(a)

Proof b: To prove: $cost_{\omega}\left(\hat{S}, T *\right) \leq 2\sum_{i=1}^{l} cost\left(S_{i}, T_{i}\right) + 2. cost(S, T *)$

For any $x \in S_{ij}$ where $1 \leq i < l, 1 \leq j \leq k$:

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

Sum over all i,j,x we get:

$$cost_{\omega}\left(\hat{S}, T *\right) \leq 2\sum_{i=1}^{l} cost\left(S_{i}, T_{i}\right) + 2. cost(S, T *) \quad \text{–eq(3.2)}$$

Hence Proved Proof(b)

From Proof of Q1 i.e., eq(1.4), we know:

$$cost(S, T) \leq 2. cost_{\omega}(\hat{S}, T) + 2\sum_{i=1}^{l} cost\left(S_{i}, T_{i}\right)$$

Using proof of Q2 i.e., eq(2.1), we can rewrite the above equation as:

$$cost(S, T) \leq 2. cost_{\omega}(\hat{S}, T) + 2\alpha cost(S, T *)$$

Using Proof a i.e., eq(3.1), we can rewrite the above equation like so:

$$cost(S, T) \leq 2\alpha cost_{\omega}(\hat{S}, T *) + 2\alpha cost(S, T *) \quad \text{— eq(3.3)}$$

SUbstituting proof of Q2 i.e., eq(2.1) in eq(3.2), we get:

$$cost_{\omega}(\hat{S}, T *) \leq 2\alpha cost(S, T *) + 2cost(S, T *) \quad \text{— eq(3.4)}$$

Using eq(3.3) and eq(3.4)

$$cost(S, T) \leq 2\alpha(2\alpha cost(S, T *) + 2cost(S, T *)) + 2\alpha cost(S, T *)$$

=> $cost(S, T) \leq \left(4\alpha^2 + 6\alpha\right)cost(S, T *)$.

Hence Proved