CS550: Massive Data Mining and Learning Spring 2019
Problem Set 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):

None

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

None

I acknowledge and accept the Honor Code.

(Signed)_Haoyang Zhang_____

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

Answer to Question 1

Let $T_i(x) = \underset{z \in T}{\operatorname{argmin}}\, d(x,z)$:

$$
\begin{aligned}
cost(S,T) &= \sum_{i=1}^{l} \sum_{x \in S_i} \big(d(x,T)\big)^2 \\
&\leq \sum_{i=1}^{l} \sum_{x \in S_i} \Big(d\big(x,T_i(x)\big) + d\big(T_i(x),T\big)\Big)^2 \\
&\leq \sum_{i=1}^{l} \sum_{x \in S_i} \Big(2\Big(d\big(x,T_i(x)\big)\Big)^2 + 2\big(d\big(T_i(x),T\big)\big)^2\Big) \\
&= \sum_{i=1}^{l} 2\Big(d\big(x,T_i(x)\big)\Big)^2 + \sum_{i=1}^{l} \sum_{j} |S_{ij}| \Big(2\Big(d\big(T_j,T\big)\Big)^2\Big) \\
&= 2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost_w(\hat{S},T)
\end{aligned}
$$

Answer to Question 2

Let $T_i \leftarrow S_i$ indicate that we compute T_i on set S_i. By the definition of ALG, we have:

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

Let $T_i^* = \underset{T_i \leftarrow S_i}{\operatorname{argmin}} cost(S_i, T_i)$, and then we have $\forall T_i \leftarrow S_i, cost(S_i, T_i) \geq cost(S_i, T_i^*)$. Therefore, $cost(S_i, T^*) \geq cost(S_i, T_i^*)$:

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

Answer to Question 3

Similarly, we have:

$$cost_w(\hat{S},T) \le \alpha \min_{T \leftarrow \hat{S}} cost_w(\hat{S},T) \ \le \alpha cost_w(\hat{S},T^*)$$

Consider $cost_w(\hat{S},T^*)$:

$$\begin{aligned} cost_w(\hat{S},T^*) &= \sum_{i=1}^{l}\sum_{j}|S_{ij}|\left(d(T_j,T^*)\right)^2 \\ &\le \sum_{i=1}^{l}\sum_{x\in S_i}\left(2\left(d(T_i(x),x\ )\right)^2 + 2\left(d(x,T^*)\right)^2\right) \\ &= 2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost(S,T^*) \end{aligned}$$

Therefore, we have:

$$\begin{aligned} cost(S,T) &\le 2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost_w(\hat{S},T) \\ &\le 2\sum_{i=1}^{l} cost(S_i,T_i) + 2\alpha\left(2\sum_{i=1}^{l} cost(S_i,T_i) + 2cost(S,T^*)\right) \\ &= (2+4\alpha)\sum_{i=1}^{l} cost(S_i,T_i) + 4\alpha cost(S,T^*) \\ &\le \left((2+4\alpha)\alpha + 4\alpha\right)cost(S,T^*) \\ &= (4\alpha^2 + 6\alpha)cost(S,T^*) \end{aligned}$$