CS550: Massive Data Mining and Learning
Homework 4

Due 11:59pm Wednesday, April 29, 2020

Only one late period is allowed for this homework

Submitted by:

Name: Prakruti Joshi

NetID: phj15

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):

Keya Desai (kd706), Twisha Naik (tn268)
On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Honor Code.

(Signed) Prakruti Joshi

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

Answer to Question 1

To prove:

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

Starting with the LHS, using the fact that $S = \bigcup_{i=1}^{l} S_i$:

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

By triangle inequality, we have:

$$d(x,z) \leq d(x,y) + d(y,z) \tag{3}$$

Thus, it follows that:

$$\min_{z \in T}[d(x,z)] \leq \min_{z \in T}[d(x,y) + d(y,z)] = d(x,y) + \min_{z \in T}[d(y,z)] \tag{4}$$

Substituting Eq. 4 in Eq. 2, we get:

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

Applying the inequality, $(a+b)^2 \leq 2a^2 + 2b^2$, to Eq. 5:

$$\begin{aligned} cost(S,T) &\leq 2\sum_{i=1}^{l} \sum_{x \in S_i} d(x,y)^2 + 2\sum_{i=1}^{l} \sum_{x \in S_i} \min_{z \in T}[d(y,z)]^2 \\ &\leq 2\sum_{i=1}^{l} \sum_{x \in S_i} d(x,y)^2 + 2\sum_{i=1}^{l} \sum_{x \in S_i} d(y,T)^2 \end{aligned} \tag{6}$$

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,

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

Consider the second term. We note that y takes the values in $\hat{S} = t_{ij}$, and the number of times that y takes a particular outcome t_{ij} is proportional to the number of times $x \in S_i$ is assigned to cluster center t_{ij}.

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

Putting the two results in Eq. 6, we get the desired result:

$$cost(S,T) \leq 2.\sum_{i=1}^{l} cost(S_i,T_i) + 2cost_w(\hat{S},T) \tag{7}$$

Hence proved.

Answer to Question 2

To prove:

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

Solution:

The algorithm ALG described in the question 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 stems from the fact that the algorithm ALG returns a set T_i that is α-approximate of T_i^*. The second inequality stems from the reasoning that since T_i is the optimal clustering set for S_i. Thus, it must necessarily have a cost that is lower than any other candidate $T^{'}$ including T^*.

Summing over i, we get as follows:

$$\begin{aligned} \sum_{i=1}^{l} cost(S_i,T_i) &\leq \alpha.\sum_{i=1}^{l} cost(S_i,T^*) \\ \implies \sum_{i=1}^{l} cost(S_i,T_i) &\leq \alpha.cost(S,T^*) \qquad (\because S = \bigcup_{i=1}^{l} S_i) \end{aligned}$$

Hence proved.

Answer to Question 3

To prove:
ALGSTR is a $(4\alpha^2 + 6\alpha)$-approximation algorithm for the k-means problem.
Solution:
To prove this, it is enough to show,

$$cost(S, T) \leq (4\alpha^2 + 6\alpha) \cdot cost(S, T^*)$$

Proofs of Facts:

- **Fact 1**
  Let $\hat{T}^*$ be the optimum clustering for the subset $\hat{S}$. Then,

$$\begin{aligned} cost_w(\hat{S}, T) &\leq \alpha \cdot cost_w(\hat{S}, \hat{T}^*) \\ &\leq \alpha \cdot cost_w(\hat{S}, T^*) \end{aligned} \tag{8}$$

- **Fact 2**
  For any $x \in S_{ij}$ where $1 \leq i < l, 1 \leq j \leq k$:

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

  Summing over all values of i, j and x, we get:

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

Using these Facts,

From Q-1 we know,

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

Using the proof from Q-2,

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

Using Fact 1,

$$cost(S, T) \leq 2\alpha \cdot cost_w(\hat{S}, T^*) + 2\alpha cost(S, T^*) \tag{9}$$

Now, from Fact 2 we get,

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

Replacing the first term using Q-2,

$$cost_w(\hat{S}, T^*) \leq 2\alpha cost(S, T^*) + 2cost(S, T^*) \tag{10}$$

Using eq. 9 and 10,

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

Hence proved.