CS550: Massive Data Mining and Learning
Homework 4

Twisha Gaurang Naik (tn268)

Due 11:59pm Wednesday, Apr 29, 2020
Only one late period is allowed for this homework (11:59pm Apr 30)

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): Prakruti Joshi (phj15), Keya Desai (kd706)

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

I acknowledge and accept the Honor Code.

(Signed) Twisha Gaurang Naik (tn268)

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

Answer to Question 1

To prove:

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

Proof:

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

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

By triangle inequality, we have:

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

Thus, we get:

$$\underset{z \in T}{min}[d(x,z)] \leq \underset{z \in T}{min}[d(x,y) + d(y,z)] = d(x,y) + \underset{z \in T}{min}[d(y,z)] \tag{2}$$

Substituting equation 2 in equation 1, we get:

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

Applying the inequality, $(a+b)^2 \leq 2a^2 + 2b^2$, to this equation:

$$\begin{aligned} cost(S,T) &\leq 2 \sum_{i=1}^{l} \sum_{x \in S_i} d(x,y)^2 + 2 \sum_{i=1}^{l} \sum_{x \in S_i} \underset{z \in T}{min}[d(y,z)]^2 \\ &\leq 2 \sum_{i=1}^{l} \sum_{x \in S_i} d(x,y)^2 + 2 \sum_{i=1}^{l} \sum_{x \in S_i} d(y,T)^2 \end{aligned} \tag{3}$$

- **First Term:**
  For every $x \in S_i$, let y = t_{ij} i.e. y is the centroid assigned to $x \in S_i$.
  Thus,
  $$\sum_{x \in S_i} d(x,y)^2 = \sum_{x \in S_i} d(x,T_i)^2 = cost(S_i,T_i)$$

- **Second term:**
  y takes the values in $\hat{S} = t_{ij}$, and the number of times that y takes a particular value t_{ij} is proportional to the number of times $x \in S_i$ is assigned to cluster center t_{ij}. Thus,

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

Substituting these results in equation 3,

$$cost(S,T) \leq 2 \cdot \sum_{i=1}^{l} cost(S_i, T_i) + 2cost_w(\hat{S}, T) \tag{4}$$

Hence, proved.

Answer to Question 2

To prove:

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

Proof:
The algorithm ALG described in the question guarantees an upper bound such that for each individual term cost(S_i, T_i),

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

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

- The first inequality is derived 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,

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

Hence, proved.

Answer to Question 3

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

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

Proofs:

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

$$\begin{aligned} cost_w(\hat{S}, T) &\leq \alpha \cdot cost_w(\hat{S}, \hat{T}^*) \\ &\leq \alpha \cdot cost_w(\hat{S}, T^*) \end{aligned} \tag{5}$$

- **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:

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

- **Main Proof**
  From Question (1) we know,

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

  From Question (2), we can rewrite this as,

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

  Using Fact 1,

$$cost(S, T) \leq 2\alpha \cdot cost_w(\hat{S}, T^*) + 2\alpha cost(S, T^*) \tag{6}$$

  Now, Fact 2 says,

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

  Replacing the first term using part (b),

$$cost_w(\hat{S}, T^*) \leq 2\alpha cost(S, T^*) + 2cost(S, T^*) \tag{7}$$

Using equation 6 and 7,

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

Hence, proved.