CS550: Massive Data Mining and Learning
Homework 4

Christos Mitropoulos - cm1012
email c.mitro@rutgers.edu

Github repository:
https://github.com/CMitropoulos/MassiveDataMining/tree/master/Homework4

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):
I acknowledge and accept the Honor Code.

(Signed) CM

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

Answer to Question 1

Let $T(x) = argmin_{t \in T} d(t, x)$. For any $x \in S_{ij}$ by triangle inequality we get $d(x, T(x)) \leq d(x, T(t_{ij}) \leq d(x, t_{ij}) + d(t_{ij}, T(t_{ij})) = d(x, t_{ij}) + d(t_{ij}, T)$
From the small fact given in the question, $d(x, T)^2 \leq 2d(x, t_{ij})^2 + 2d(t_{ij}, T)^2$
We get the result by summing up over all i, j, x.

Answer to Question 2

Assume T_i^* is the optimal clustering for $S_i (i \in [1, l])$. Then $cost(S_i, T_i) \leq a \times cost(S_i, T_i^*) \leq a \times cost(S_i, T^*)$. Similar like before we get the result by summing over all i.

Answer to Question 3

If $\hat{T}^*$ is the best clustering for $\hat{S}$, then: $cost_w(\hat{S}, T) \leq a \times cost_w(\hat{S}, \hat{T}^*) \leq a \times cost_w(\hat{S}, T^*)$ As we did in part (a) for any $x \in S_{ij}, i \in [1, l], j \in [1, k]$, we get: $d(t_{ij}, T^*)^2 \leq 2d(t_{ij}, x)^2 + 2d(x, T^*)^2$. Summing over all i, j, x provides the result.