COS 435, Spring 2016 - Problem Set 6

Due 11:59 pm Wednesday April 13, 2016 by DropBox submission

FOR THIS ASSIGNMENT, WE ASK THAT YOU SCAN ANY HANDWRITTEN WORK AND SUBMIT BY DROPBOX

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently. For each problem, list the students with whom you discussed general methods of solving the problem.

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- Penalized 10% of the earned score if submitted by 10am Thursday (4/14/16).
- Penalized 25% of the earned score if submitted by 4:30 pm Friday (4/15/16).
- Penalized 50% if submitted later than 4:30 pm Friday (4/15).

Submission

Submit your solutions as a PDF file using the Computer Science Department DropBox submission system for COS435 at
https://dropbox.cs.princeton.edu/COS435_S2016/HW6 Name your file HW5.pdf. If you have not used this facility before, consult the instructions at
https://csguide.cs.princeton.edu/academic/csdropbox - student
Note that you are automatically enrolled in CS DropBox using the registrar's COS435 enrollment list.

CHANGE OF POLICY: You may hand write your solutions as long as they are legible. In this case, you must scan your writing to produce a PDF file for submission through DropBox.

Problem 1:

Slide #16 of Part 2 of the slides for clustering, posted under April 4, presents an iterative improvement algorithm for divisive partitioning. This problem addresses recalculating the total relative cut cost (slides #10 and #11) incrementally for use with that algorithm.

Let U denote the set of objects to be clustered. Assume that for any objects v and w, sim(v,w)=sim(w,v) (we have been assuming this in class). Also assume that for any object v, sim(v,v)=0. Let C_p be an arbitrary cluster containing object x, C_q be an arbitrary cluster that does not contain x. (The set notation $C_p - \{x\}$ denotes C_p with x removed, and $C_q \cup \{x\}$ denotes C_q with x added.)

The following relationship holds for incremental changes to the intracost of a cluster when removing or adding an object x.

$$\text{intracost}(C_p) - \text{intracost}(C_p - \{x\}) = \sum_{v_i \text{ in } C_p - \{x\}} \text{sim}(v_i, x)$$

$$= \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x) \qquad \text{since sim}(x,x) = 0$$

From this relationship we derive the incremental cost changes for intracost:

$$\text{intracost}(C_p - \{x\}) = \text{intracost}(C_p) - \sum_{v_i \text{ in } C_p} \text{sim}(v_i, x)$$

$$\text{intracost}(C_q \cup \{x\}) = \text{intracost}(C_q) + \sum_{v_i \text{ in } C_q} \text{sim}(v_i, x)$$

Your task is to derive incremental cost equations for cutcost. The ultimate goal is to minimize the computation time used by the iterative improvement algorithm.

Part a:

i. Give an equation for

$$\text{cutcost}(C_p) - \text{cutcost}(C_p - \{x\})$$

when x is an object in C_p. Your equation should be in terms of similarities between x and other objects.

Using your equation, derive equations for

ii. $\text{cutcost}(C_p - \{x\})$ as an incremental change to $\text{cutcost}(C_p)$;

iii. $\text{cutcost}(C_q \cup \{x\})$ as an incremental change to $\text{cutcost}(C_q)$.

Part b: Given the equations for the incremental changes in intracost and cutcost, what is the computational time complexity of the step:

move v_j to that cluster, if any, such that move gives maximum decrease in cost

of the iterative improvement algorithm on slide #25? Specify the data structures you are using and how they are used to achieve the time complexity.

Problem 2:

For this problem we consider the eight points in the plane shown in the Figure on the next page, where ***s*** is the separation between the two horizontal rows, and the distance between neighboring points on a row is 1. The i^{th} point on the top row is vertically aligned with the i^{th} point on the bottom row. Also, for this problem we **use the L_1 distance** as the similarity measure (smaller is more similar)**.** The L_1 distance between two points (x_1, y_1) and (x_2, y_2) is $|x_1-x_2| + |y_1-y_2|$. L_1 is also called the Manhattan distance because a path is composed of east-west and north-south segments (at least in midtown).

Consider the points split into two clusters {a,b,c,d} and {e,f,g,h}. Using symmetry, the conductance can be easily calculated: the calculation for point a is the same as for d, and the calculation for point b is the same as for c.

Here is the computation of cutcost({a,b,c,d}):
sum of similarities of point a with e,f,g,h = s + (s+1) + (s+2) + (s+3) = 4s+6
sum of similarities of point b with e,f,g,h = (s+1) + s + (s+1) + (s+2) = 4s+4
Since, by symmetry, c yields same value as b and d yields same value as a, we get
cutcost = 2*(4s+6) + 2*(4s+4) = 16s + 20

Here is the computation of (2* intracost).
L_1 distance a to b,c,d: 1+2+3 = 6
L_1 distance b to a,c,d: 1+1+2 = 4
Since c yields same value as b and d yields same value as a, we get
(2*intracost) = 2*6+2*4 = 20
(Note we computed (2*intracost) directly because we considered each pair twice, e.g. (a,b) and (b,a).)

This gives
conductance ({a,b,c,d}) = conductance ({e,f,g,h}) =
cutcost({a,b,c,d}/ s_degree({a,b,c,d}) = (16s+20)/(16s+ 40)

Part a: Consider the points split into two clusters {a,b,e,f} and {c,d,g,h}. What is the conductance of {a,b,e,f} as a function of s? Use symmetry to simplify the computation.

Part b: What is the minimum ***integer*** value such that for all values of s at least this large, conductance({a,b,c,d}) > conductance({a,b,e,f}), and thus the partition {a,b,c,d} {e,f,g,h} is preferred over {a,b,e,f} {c,d,g,h}. (Note we maximize conductance because we are using a distance measure.)

Figure for Problem 2: