ORIE 5380 - Project - Andrea Lodi

A company delivering goods has to set up some depots to help the distribution to a set $C = \{1, \ldots, n\}$ of clients by assigning to each depot clients that are not too far from each other. To do that an $n \times 2$ matrix X is given, where each row i represents the coordinates of client i. and one needs to partition the set of clients C into subsets, each one to be assigned to a depot. We want to minimize the number of depots constructed. The constraints require that

- each client has to be assigned to a unique subset (i.e., a depot),
- each subset has a number of clients between $n_{\min}$ and $n_{\max}$, and
- the distance between any pair of clients in the same subset cannot be larger than $d_{\max}$.

Questions:

1. Define a natural Integer Linear Programming model with polynomially many constraints and variables for the problem and discuss the strength or weakness of some of its constraints. Implement and solve your integer programming formulation on the 3 provided problem instances, each given as a matrix containing $n = 25$ client coordinates. Your implementation can either be in Julia (using JuMP and Gurobi) or Python (using the `gurobipy` package for example). For each problem instance, please use the following parameters:
   - Instance 1: $n_{\min} = 2$, $n_{\max} = 8$, $d_{\max} = 10$
   - Instance 2 and 3: $n_{\min} = 4$, $n_{\max} = 6$, $d_{\max} = 9$
2. For each of the 3 instances, how do the results change as you adjust $n_{\min}$, in terms of feasibility and objective value? Try doing the same with $n_{\max}$ and $d_{\max}$ and explain what happens.
3. Propose a formulation (with polynomially many constraints and variables) of the above problem with the goal of minimizing the average pairwise distance (where "pairs" are clients in the same cluster). Find these clusters for all datasets with feasible solutions. Are they different from your answers in part 1? Note that you need to take into account the possibility of label switching when comparing cluster assignments.

   Clarifications: The objective is no longer to minimize the number of clusters. Any clustering that was feasible for question 1 is feasible here. Any clustering that was infeasible for question 1 is infeasible here.
   In your instances you have 25 clients. There are $(25 \cdot 24)/2 = 300$ pairs of clients.
   Let T be the set of these 300 pairs. For every pair, either your clustering put both those clients in the same cluster or it placed them in separate clusters. Let T_1 be the set of pairs of clients who were both in the same cluster. T_1 should be a strict subset of T. For each pair in T_1, there is a distance between the two points in that pair. **Your integer programming formulation should find the clustering that minimizes the average of those pairwise distances.** The formulation needs to be linear.
   The challenge with this problem is that the number of pairs in T_1 (which is the denominator in your average) is clustering dependent. However, there is a way to overcome this obstacle.

4. **Bonus question**: Suppose there are 25 clients, all located on a 2D plane. For 15 of them, we know the exact location. For the remaining 10, we are given 3 possible locations, each of which could be realized with probability $\frac{1}{3}$. $n_{\min} = 0$. You have to open clusters in 2 phases. Clusters opened in phase 1 have a cost of 1, clusters opened in phase 2 have a cost of 2. The exact locations of the 10 clients are revealed to you after the phase 1 clusters are chosen. Every cluster opened in phase 1 must have either:
   (a) An associated client among the 15 which is assigned to it. **OR**
   (b) Be associated to one of the 30 possible locations for the 10 uncertain clients.

All remaining clients may be assigned after phase 2 clusters are opened.

As in previous questions, no 2 clients in the same cluster can be more than $d_{\max}$ apart. Moreover, each cluster can have at most $n_{\max}$ clients. There is one additional restriction: If a cluster was opened in phase 1 and assigned to one of the 30 possible locations, then no client in the cluster may be more than $d_{\max}$ away from that possible location.

Write (and explain thoroughly) an integer programming formulation with polynomially many constraints and variables to solve this problem. You may find Chapter 17 of the textbook a useful read. You do not need to implement this in code.

Deliverables:

- Implementation: the Julia (or Python) script(s) or Jupyter notebook(s) containing your code to parts 1-3.
- Writeup: a LaTeX PDF file for the problem formulation and answers to the remaining questions. In your answers to 1 and 3 also please include a scatter plot with clients colored according to their assigned clusters.

Please submit a single zip file containing the implementation and writeup files.