Lecture 3 — The k-medoid clustering problem

3.1 Problem formulation

Let's go back to working in a metric space $(\mathcal{X}, \rho)$. Here's the problem.

> k-MEDOID CLUSTERING
> *Input:* Finite set $S \subset \mathcal{X}$; integer k.
> *Output:* $T \subset S$ with $|T| = k$.
> *Goal:* Minimize $\text{cost}(T) = \sum_{x \in S} \rho(x, T)$.

This is quite similar to k-means, except that T is forced to be a subset of S (rather than $\mathcal{X}$) and the cost function involves distance rather than squared distance. It is known that this problem is hard to approximate within a factor better than $1 + 1/e$.

3.2 A linear programming relaxation

For convenience index the points in S by $1, 2, \ldots, n$, with interpoint distances $\rho(i,j), 1 \leq i, j \leq n$. Then the k-medoid problem is solved exactly by the following integer program.

$$\begin{aligned} \min \sum_{i,j} x_{ij} \rho(i,j) & \\ \sum_j y_j \leq k & \\ \sum_j x_{ij} = 1 & \\ x_{ij} \leq y_j & \\ x_{ij}, y_j \in \{0,1\} & \end{aligned}$$

where the variables $\{x_{ij}, y_j\}$ have the following interpretation.

$$\begin{aligned} y_j &= \mathbf{1}(\text{point } j \text{ is used as a medoid}) \\ x_{ij} &= \mathbf{1}(j \text{ is the medoid serving point } i) \end{aligned}$$

An integer program cannot in general be solved efficiently, so we turn it into a linear program by relaxing the last two constraints:

$$0 \leq x_{ij}, y_j \leq 1$$

The resulting LP can be solved in polynomial time.

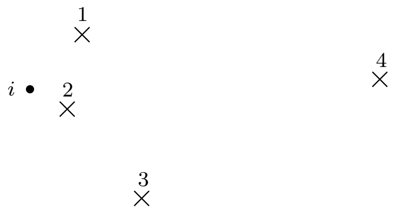
3.2.1 Rounding the LP solution

Suppose the optimal solution to the k-medoid instance has cost $\textsc{opt}$. Since this solution is feasible for the linear program, the optimal LP solution has some cost $\textsc{opt}_{LP} \leq \textsc{opt}$. Say this solution consists of variables $\{x_{ij}, y_j\}$. The difficulty, of course, is that these values might be fractional (such as $y_1 = 0.2$, $y_2 = 0.5$, and so on). We'll show that it is possible to *round* this fractional solution into one that has $2k$ medoids and has cost at most $4\textsc{opt}_{LP}$.

In the LP solution, point i incurs a cost

$$C_i = \sum_j x_{ij} \rho(i, j).$$

This might be spread out over several centers j: those with $x_{ij} > 0$. For instance, it might be the case that $x_{i1}, x_{i2}, x_{i3}, x_{i4} > 0$ (see figure below), and since $\sum_j x_{ij} = 1$, we can think of the x_{ij}'s as a probability distribution over centers for i. Under this distribution, C_i is the *expected* distance of a center from i.



The total cost is $\textsc{opt}_{LP} = \sum_i C_i$. We will find a set of $2k$ medoids in S such that each point i is within distance at most $4C_i$ of these medoids. The total cost will then be at most $4\textsc{opt}_{LP}$.

The hardest points to cover are those with the smallest values of C_i, so let's start by focusing on those. Pick the smallest C_i. If we include i as a medoid, we can use it cover any $i' \in S$ whose distance from i is at most $4C_i$; denote the set of such points by $B(i, 4C_i)$. This is because i has the smallest C_i value, and thus $\rho(i, i') \leq 4C_i \leq 4C_{i'}$.

But this is overly conservative. We can in fact use i to cover any point i' such that $B(i, 2C_i) \cap B(i', 2C_{i'}) \neq \emptyset$. To see this, notice that since the two balls intersect, they have some point q in common, and thus $\rho(i, i') \leq \rho(i, q) + \rho(i', q) \leq 2C_i + 2C_{i'} \leq 4C_{i'}$. Therefore, define the *extended neighborhood* of i as follows.

$$\overline{V}_i = \{i' \in S : B(i, 2C_i) \cap B(i', 2C_{i'}) \neq \emptyset\}.$$

Now we can state the algorithm simply.

`solve the LP and compute the values` C_i
$T \leftarrow \{\,\}$
`while` $S \neq \{\,\}$:
  `pick the` $i \in S$ `with smallest` C_i
  $T \leftarrow T \cup \{i\}$
  $S \leftarrow S \setminus \overline{V}_i$

We will show the following.

Theorem 1. *$\mathrm{cost}(T) \leq 4\textsc{opt}_{LP}$ and $|T| \leq 2k$.*

3.2.2 Analysis

First we'll show that $\text{cost}(T) \leq 4\text{OPT}_{LP}$. This is an immediate consequence of the following lemma.

Lemma 2. *Pick any $q \in S$, and suppose i is the first point selected (to be in T) for which $q \in \overline{V}_i$. Then: (a) $C_i \leq C_q$ and (b) $\rho(q,i) \leq 4C_q$.*

Proof. At the moment when i is selected, both i and q are available in S. Therefore $C_i \leq C_q$. For (b), the condition $q \in \overline{V}_i$ implies that there is some point s in both $B(i, 2C_i)$ and $B(q, 2C_q)$. Thus

$$\rho(q,i) \leq \rho(i,s) + \rho(q,s) \leq 2C_i + 2C_q \leq 4C_q.$$

□

Next we need to bound the size of T. The argument will go like this: we'll show that for each point i selected to be in T, the neighborhood $B(i, 2C_i)$ contains at least "half a medoid": more precisely, the sum of y_j for $j \in B(i, 2C_i)$ is at least $1/2$. However, these neighborhoods are all disjoint (for different $i \in T$) and the total sum of y values is at most k. Therefore there can be at most $2k$ such points $i \in T$.

Lemma 3. *Pick any $i \in T$. Then*

$$\sum_{j \in B(i,2C_i)} y_j \geq \sum_{j \in B(i,2C_i)} x_{ij} \geq \frac{1}{2}.$$

Proof. The first inequality follows from the constraint $x_{ij} \leq y_j$ in the LP. To see the second inequality, define a random variable $Z \in \mathbb{R}$ that takes value $\rho(i,j)$ with probability x_{ij}. As we saw above, $\mathbb{E}Z = \sum_j x_{ij}\rho(i,j) = C_i$. By Markov's inequality,

$$\sum_{j \in B(i,2C_i)} x_{ij} = \mathbb{P}[Z \leq 2C_i] = 1 - \mathbb{P}[Z > 2\mathbb{E}Z] \geq \frac{1}{2}.$$

□

The rest is immediate, since for any $i, i' \in T$, we know $B(i, 2C_i) \cap B(i', 2C_{i'}) = \emptyset$.

Problem 1. Is there a linear or convex programming relaxation for the k-means problem in which the centers are not constrained to be data points?