SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Data partitioning with k-means clustering

BM Bui-Xuan

Let $\mathcal{S}$ be a set of n points in a 2D plane.

k-means clustering :

The k-means clustering problem consists of computing a partition of $\mathcal{S}$ into k parties in such a way that the total sum of distances between each element of a party to the barycenter of that party is minimized. Heuristics for k-means clustering aim to give a partition with a small total sum of distances (without being the optimal one).

In this lab session, the goal is to propose such a heuristics, with $k = 5$ and $\mathcal{S}$ given in the input file `input.points` in the GUI canvas.

Goal : obtain the smallest score possible.

Constrained k-means clustering :

The constrained k-means clustering problem with a limited budget is defined as follows. Given k an integer, B a real number called *budget*, and $s_1, s_2, \ldots, s_k \in \mathcal{S}$ elements of $\mathcal{S}$ called *centers*, the budgeted k-means clustering problem consists of finding k pairwise disjoint subsets $S_1, S_2, \ldots, S_k \subseteq \mathcal{S}$ such that :

— for every $1 \leqslant i \leqslant k$, we have that $s_i \in S_i$;

— for every $1 \leqslant i \leqslant k$, the total sum of distances between each element of S_i to the barycenter of S_i is at most B ;

— the number of elements in $S_1 \cup S_2 \cup \cdots \cup S_k$ is maximized.

In the lab session, the goal is to propose a heuristics to the budgeted k-means clustering problem, with $k = 5$, $B = 10101$, $\mathcal{S}$ given in input file `input.points`, and $s_1, s_2, \ldots, s_5$ the five first points in that input file.

Goal : obtain the largest score possible.