Prof. Dr. Johannes Blömer 29.10.2015
submission due: 12.11.2014, F1.110, 11:15

Clustering Algorithms

WS 2015/2016

Handout 2

Exercise 1:
Let $D: M \times M \to \mathbb{R}_{\geq 0}$ be a distance function, $P \subset M$, $|P| < \infty$, and $k \geq 2$.
Prove the following statements.

(a) $\text{cost}_k^D(P) \leq \text{cost}_{k-1}^D(P)$

(b) Let C be a set of optimal k-medians of P, i.e., $D(P, C) = \text{cost}_k^D(P)$.
Let $\{P_1, \ldots, P_k\}$ be the partition induced by C, i.e. $P = \cup_{i=1}^k P_i$.
Then for all $i = 1, 2, \ldots, k$, we have

$$D(P \setminus P_i, C \setminus \{c_i\}) = \text{cost}_{k-1}^D(P \setminus P_i).$$

Exercise 2:
Let $X \subseteq \mathbb{R}^d$ be a set of n points, where each point x has assigned a weight $w(x) \in \mathbb{R}^{\geq 0}$. The cost of center m is defined by

$$cost(m) = \sum_{x \in X} w(x) D_{l_2^2}(x, m).$$

Show that

$$cost(m) = cost(X, c(X)) + W \cdot D_{\ell_2^2}(c(X), m),$$

where $W = \sum_{x \in X} w(x)$ and $c(X) = \frac{1}{W} \sum_{x \in X} w(x) \cdot x$.

Exercise 3:
In the lecture we saw that started with poor initial centroids the k-means algorithms may compute solutions that are arbitrarily worse than an optimal solution. However, for the construction we used initial centroids that are not contained in the input set.
Show that even if started with initial centroids from the input set, the k-means algorithm may compute arbitrarily poor solutions.