Prof. Dr. Johannes Blömer

10.12.2015
submission due: 7.1.2016, F1.110, 11:15

Clustering Algorithms

WS 2015/2016

Handout 7

Exercise 1:

(a) Let $D: \; M \times M \to \mathbb{R}^{\geq 0}$ be a metric, $P \subset M$ and $\mathcal{C} = \{C_1, \dots, C_k\}$ a partition of P. Then

$$\frac{1}{2} \cdot \text{cost}_{drad}(\mathcal{C}) \leq \text{cost}_{rad}(\mathcal{C}) \leq \text{cost}_{drad}(\mathcal{C})$$

(b) Now let $M = \mathbb{R}^d$, and let $D := D_{l_2^2}$ be the squared euclidean distance. Then

$$\text{cost}_{drad}(\mathcal{C}) \leq \text{cost}_{diam}(\mathcal{C}) \leq 4 \cdot \text{cost}_{drad}(\mathcal{C})$$

Exercise 2:

(a) We say a clustering $C = (C_1, \dots, C_k)$ is *strongly separated* if for all $i = 1, \dots, k$

$$\max_{x,y \in C_i} D(x,y) < \min_{i \neq j} \min_{x \in C_i,\; y \in C_j} D(x,y).$$

Given that the optimal clustering is stronly separated, prove that agglomerative clustering with complete linkage cost computes an optimal diameter k-clustering in $n-k$ steps.

(b) Give a small example that shows that agglomerative clustering with complete linkage cost might *not* compute an optimal diameter k-clustering in $n-k$ steps if C is *not* strongly separated.

Exercise 3:

Divisive clustering algorithms start with the complete data set as one cluster. In each step, they split a cluster up into two clusters.

Consider the following divisive algorithm to find a partition of P into k clusters $C_1, \dots, C_k$ such that $cost^D_{diam}(C)$ is minimized.

Algorithm 1 GreedyDivisive(X, k):

1: Create one cluster containing the complete dataset
2: **for** $r = n$ downto k **do**
3: Split the cluster with the maximum diameter into two clusters, such that the new clusters have the smallest possible diameter

Show that the clustering obtained by this algorithm can be arbitrarily bad.
Hint: Consider some suitable set $P \subset \mathbb{R}$ with $|P| = 4$ and $k = 3$.

Exercise 4:
Define metric matrix diameter clustering to be the restriction of matrix diameter clustering to symmetric matrices $\Delta \in \mathbb{R}_{\geq 0}^{n \times n}$ that define a metric. That is, for any triple (x, y, z) of indices $\Delta_{xy} + \Delta_{yz} \leq \Delta_{xz}$. Show that unless $P = NP$ metric matrix diameter clustering cannot be approximated with factor $2 - \epsilon, \epsilon > 0$ arbitrary.
Hint: Use the length of a shortest path in a graph as a metric.