Institut für Informatik
Dr. Melanie Schmidt
Clemens Rösner

UNIVERSITÄT BONN

Cluster Analysis
Winter 2018/19

Problem Set 9

Please hand in your solutions for this problem set via email (roesner@cs.uni-bonn.de) or personally at Room 2.060 until *Tuesday, 18th of December*.

Problem 1

We want to look at the distributed partition-based k-means problem. In the distributed partition-based k-means problem the data arises at different places. Specifically we assume that the knowledge of P is spread over the different places, where each place knows only some of the points in P and every point in P is known by exactly one place. The different places then need to communicate with each other in order to compute a good k-means solution on the whole set of points P. We want to minimize the amount of communication necessary to be able to compute an approximate solution. Assume that the number of points at each place is not very large but the points have a high dimension. How can we restrict the necessary amount of information each place has to communicate?

Problem 2

Let $x_1, \dots, x_n \in \mathbb{R}^d$ be a set of points, and let $V_k = \arg\min_{V \subseteq \mathbb{R}^d, dim(V)=k} \sum_{i=1}^n dist^2(x_i, V)$ be a linear subspace of $\mathbb{R}^d$ minimizing the sum of the squared distances between $\{x_1, \dots, x_n\}$ and the subspace among all k-dimensional subspaces.

- Show that $\sum_{i=1}^n dist^2(x_i, V_k)$ is at most as expensive as the best k-means solution on $\{x_1, \dots, x_n\}$.

Problem 3

Let V be any linear subspace of $\mathbb{R}^d$ and for every point $p \in \mathbb{R}^d$ let $V(p) = \arg\min_{q \in V} d(p, q)$ be the point in V closest to p.

- Show that for any point $p \in \mathbb{R}^d \setminus V$, and any point $q \in V$ with $q \neq V(p)$ the triangle $\Delta pqV(p)$, that has the three points p, q and $V(p)$ as its vertices, has a right angle at $V(P)$.
- Show that for any set of point $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ the best k-means solution on X is at least as expensive as the best k-means solution on $\{V(x_1), \dots, V(x_n)\}$ in V.