Institut für Informatik
Dr. Melanie Schmidt
Clemens Rösner

UNIVERSITÄT BONN

Cluster Analysis
Winter 2018/19

Problem Set 7

Please hand in your solutions for this problem set via email (roesner@cs.uni-bonn.de) or personally at Room 2.060 until *Tuesday, 4th of December.*

Problem 1

We want to compare optimal solutions for the Euclidean k-center and the k-means problem. Assume we are given an instance of one problem together with its optimal solution. We want to use this solution for the other problem and compare it to the value of the optimal solution for that problem.

- How high can the factor between a solution for the k-center problem, obtained from an optimal k-means solution, and the optimal k-center solution be?
- How high can the factor between a solution for the k-means problem, obtained from an optimal k-center solution, and the optimal k-means solution be?

Problem 2

We want to adapt Lloyd's algorithm to the k-center program. So instead of the centroid $\mu(C_i)$, we let $\nu(C_i)$ denote the best possible center for C_i. We say that a point $p \in P$ is the best possible center for a set C_i if the following is true. Let $|C_i| = m$, let $C_i = \{c_1, \ldots, c_m\}$ be ordered such that $d(c_i, p) \geq d(c_j, p)$ for all $i \leq j$ and let $C_i(p) = (d(c_1, p), \ldots, d(c_m, p))$ denote ordered vector of distances between a point in C_i and p. We then call $p \in P$ a best possible center for C_i, if $C_i(p)$ is lexicographically smaller or equal to $C_i(q)$ for all $q \in P$. What can we say about this new algorithm? Does it converge to a local optimal solution? How bad of an approximation can such a local optimum be?

Problem 3

We want to have a look at the general idea behind the analysis of $D2$-sampling by Aggarwal, Deshpande and Kannan. Let us assume that we have a series X_i, $i \in \mathbb{N}$ of random binary events. Each event has either 0 (failure) or 1 (success) as its result. Let $p_i \in [0, 1]$ denote the success probability of X_i, i.e. $P[X_i = 1] = p_i$.

- What is the expected number of successes in the first n events?

Assume that we know $p_i \geq p$ for some constant $p \in [0, 1]$.

- Give an upper bound depending on p on the expected number of events needed to obtain at least n successes.
- Give an upper bound for the probability that we need more than twice as much events to have n successes.