Institut für Informatik
Dr. Melanie Schmidt
Clemens Rösner

Cluster Analysis
Winter 2018/19

Problem Set 8

Please hand in your solutions for this problem set via email (roesner@cs.uni-bonn.de) or personally at Room 2.060 until *Tuesday, 11th of December*.

Problem 1

In Lemma 51 of the lecture notes we have shown that if we draw x from S with D^2-sampling based on the previous center set C, then $E[dist^2(S, C \cup \{x\})]$ is small. Is it possible to give a similar small bound for $E[dist^2(S, x)]$? In other words: Is x itself on expectation a good center for S, or is only $C \cup x$ a good center set for S, but x itself can be bad?

Problem 2

Instead of the k-means problem we want to use D^2-sampling for the k-median cost function. So assume that we chose the first point x_1 uniformly at random from P and then iteratively select the next center x_i where each point $p \in P$ is chosen according to the probability distribution $\frac{d(p,C^{i-1})}{\sum_{q \in P} d(q,C^{i-1})}$ where $C^{i-1} = \{x_1, \dots, x_{i-1}\}$ denotes the set which contains the first $i - 1$ chosen points.

- Show similarly to Lemma 50 that for any set $S \subseteq P$ and $x \in S$ chosen uniformly at random we have $E[\sum_{p \in S} d(p, x)] \in O(\sum_{p \in S} d(p, q))$ for all $q \in P$.
- Show similarly to Lemma 51 that for any $C, S \subseteq P$ and $x \in S$ chosen according to the probability distribution where each point $x \in S$ has probability $\frac{d(x,C)}{\sum_{y \in S} d(y,C)}$ we have $E[\sum_{p \in S} d(p, C \cup x)] \in O(\sum_{p \in S} d(p, q))$ for all $q \in P$.

Problem 3

Give worst-case examples for the following variations of D^2-sampling.

- Instead of choosing the first point uniformly at random, pick an arbitrary point.
- Instead of choosing the first point uniformly at random, pick the centroid of P.
- Sample the first point uniformly at random. For iteration 2 up to k, do the following: Sample k points from P according to D^2-sampling (based on P and the so-far chosen centers C^{i-1}), and choose the point which reduces the cost by the largest amount.

Problem 4

Explain why the Johnson-Lindenstrauss Lemma (Theorem 52) can not be used to approximately preserve $dist^2(P, C)$ for every arbitrary set $C \subseteq \mathbb{R}^d$. This means that we would want to have a function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ such that for all $C \subseteq \mathbb{R}^d$ $dist^2(P, C)$ is approximated by $dist^2(f(P), f(C))$. What if instead of wanting to approximate $dist^2(P, C)$ for any arbitrary set $C \subseteq \mathbb{R}^d$ we are given an explicit finite set of center candidates L and wants to approximately preserve the cost function $dist^2(P, C)$ for all sets $C \subseteq L$ of size k.