Institut für Informatik
Dr. Melanie Schmidt
Clemens Rösner

Cluster Analysis
Winter 2018/19

Problem Set 11

Please hand in your solutions for this problem set via email (roesner@cs.uni-bonn.de) or personally at room 2.060 until *Tuesday, 15th of January.*

Problem 1

Let c be a constant such that for any radius $R > 0$ and any $x \in \mathbb{R}^d$, the ball $B(x, R)$ can always be covered by $2^{c \cdot d}$ balls of radius $R/2$. Propose a simple algorithm that computes $2^{\mathcal{O}(d)}$ balls of radius $R/2$ that cover all points from a given finite set of points P in a given ball $B(x, R)$. Use an algorithm from the lecture!

Problem 2

Let $P_1, P_2 \subset \mathbb{R}^d$ be two disjoint set of points. Assume that S_1 with $w_1 : S_1 \to \mathbb{R}$ and S_2 with $w_2 : S_1 \to \mathbb{R}$ are (k, ϵ)-coresets for P_1 and P_2, respectively. Show that $S_1 \cup S_2$ with $w_1 + w_2 : S_1 \cup S_2 \to \mathbb{R}$ is a (k, ϵ)-coreset for $P_1 \cup P_2$.

Problem 3

Assume that instead of the unconstrained k-means problem, we look at coresets for the k-means problem with an additional constraint. We again say that a set S with $w : S \to \mathbb{R}$ is a (k, ϵ)-coresets for a set $P \subset \mathbb{R}^d$ if for every feasible set C of k centers the weighted cost of the best constrained k-means solution on S with centers C and the cost of the best constrained k-means solution on P with centers C differ by at most a factor of $(1 + \epsilon)$. Show that the statement from task 2 is not true for: k-means clustering with outliers, capacitated k-means clustering and fair k-means clustering.

Problem 4

Similar to the distributed partition based k-means problem from problem set 9, we assume that the information of our points is shared among multiple places. Again we assume that each point is known in exactly one place. This time we want to compute a small (k, ϵ)-coreset for our whole set of points. To do so we assume that we have an algorithm which can, given a possibly weighted set P and an $\varepsilon > 0$, compute a (k, ε)-coreset of of P of size $(1/\varepsilon)^d k \log n$ where $n = |P|$ if P is unweighted and $n = \sum_{x \in P} w(x)$ if P is weighted (so it's slightly more general and slightly smaller than the coreset in the lecture). Let us assume that the dimension d of our instance is small (say $d = 2$) and that the number of different places t is much larger than d. Compare the following two approaches to the problem in terms of necessary communication and size of the resulting coreset.

- Let every place compute a coreset, communicate these coresets to a central server and then merge these sets as shown in task 2 to obtain the final coreset.
- Again let every place compute and communicate a coreset, merge these coresets and then compute a coreset of the merged set to obtain the final coreset.

Take into account, what choice for ε you use in each of the steps, given that we want to obtain a (k, ϵ)-coreset for some fixed $\epsilon > 0$ at the end of the process.