Talk Abstract

In this talk, I will describe sampling based algorithms for clustering problems, in particular, the $k$-means clustering problem. Given a dataset $X \subseteq \mathbb{R}^d$ and an integer $k$ as input, the objective in the $k$-means problem is to output a set $C \subseteq \mathbb{R}^d$ of $k$ centers such that the $k$-means cost function, $\phi(C, X) = \sum_{x \in X} \min_{c \in C} ||c - x||^2$, is minimised. Given a set $C$ of $k$ centers, the clusters in the $k$-means problem are obtained by assigning each point to its closest center. The $k$-means problem is known to be NP-hard, and a number of approximation algorithms are known for the $k$-means problem. However, in practice, Lloyd’s heuristic (also known as the $k$-means algorithm) is usually used to solve the $k$-means problem. Lloyd’s heuristic is a local search technique described as follows: Starting with an arbitrary set of $k$ centers, make local improvements until no further improvement is possible. The method of choosing the initial $k$ centers for Lloyd’s heuristic is known as seeding. Arthur and Vassilvitskii showed that a simple adaptive seeding technique, known as the $k$-means++ seeding, gives $O(\log k)$-approximation for $k$-means in expectation. The $k$-means++ seeding uses $D^2$-sampling to sample one center in each of the $k$ iterations, where $D^2$-sampling with respect to any center set $C$ chooses a point with probability proportional to the squared Euclidean distance from the point to its nearest center in $C$. $D^2$-sampling is known to be very useful in designing algorithms for clustering problems. Our work in this context may be seen as an attempt to understand the scope and limitations of $D^2$-sampling based algorithms for the $k$-means problem. Next, we describe the main results we cover in this talk.

• We refute a conjecture by Brunsch and Röglin (Theoretical Computer Science, 2013) on the behaviour of $k$-means++ seeding on low-dimensional instances.

• A number of $(\alpha, \beta)$-pseudo-approximation results for the $k$-means problem using $D^2$-sampling were known. We extend these results for any $\alpha$ where $0 < \alpha \leq 1$, and compute almost matching upper and lower bounds on $\beta$.

• Ding and Xu (ACM-SIAM Symposium on Discrete Algorithms, 2015) gave a $(1+\varepsilon)$-approximation for the constrained $k$-means problem. We design a much faster $D^2$-sampling based $(1+\varepsilon)$-approximation algorithm for the constrained $k$-means problem. Our results also generalise for the constrained $k$-median problem.

• We study $k$-means and correlation clustering problems in a semi-supervised learning framework which has access to a same-cluster query oracle. Our algorithms give $(1+\varepsilon)$-approximation guarantees for these problems. We also provide a conditional almost matching lower bound on the number of queries to the oracle.

• We design uniform and non-uniform sampling algorithms in the streaming setting. Our algorithms optimise the number of random bits and space usage.