View Reviews

Paper ID	342
Paper Title	Provably noise-robust k-means clustering

Reviewer #1

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

The paper presents algorithms for the k-means with noisy input points. Multiple results are given regarding separation of the true clusters and the noise under the assumption that the true clusters are contained in unit balls whose centers are separated by distance at least \delta. NP-hardness is shown for k=1 (as opposed to k=2 for the standard k-means).

Strengths:
-- Multiple theoretical results, rigorous formulation and treatment of the problem and new SDP/LP-based algorithms.

Weaknesses:
-- Unrealistic assumptions about the data, algorithms assume that clusters form unit balls separated with their centers separated by distances at least \delta > 2 * \alpha where \alpha depends on the specific setting.
-- Mediocre efficiency of the proposed algorithms -- can only be scaled up to 10^3 points which is too small for any interesting applications.
-- Multiple typos and somewhat sloppy theorem statements.

Major concern:
-- The separation requirement of 2(1 + \sqrt{k}) in the noiseless case doesn't seem to make sense. In the noiseless case one can separate for \delta > 4 trivially -- take an arbitrary point and form a ball of radius 2 around it. All points within the ball B_i containing this point will be cut away and no other balls will be touched. Repeat until balls are identified.

Comments:
-- In Theorem 11 in the supplement r is missing in a few places, check

2. Please provide an overall score for the submission.

Weak Reject: Borderline, tending to reject

3. Please enter a 1-2 sentence summary of your review explaining your overall score.

Multiple theoretical results regarding noisy formulations of k-means. Quite unrealistic assumptions about the structure of the data, non-scalable algorithms, some results don't quite pass a sanity check, sloppy writing.

5. Please rate your confidence in the score assigned.

High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.

Reviewer #2

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

This paper studies a setting of robust k-means clustering, where there is an inlier cluster consisting of points from the inliers. The authors suggest to consider a regularized objective (3) and prove that the optimization problem is NP-hard in general cases. Then a SDP based algorithm is proposed to solve the raised optimization problem. The SDP formula (6) mostly follows from existing papers. The main brightness of the paper is a comprehensive analysis about the optimality of the produced solution. This is significant, as the studied problem is NP-hard.

The reader has few questions:

1. The second paragraph of Introduction seems aimless. How the proposed method can resolve the under-specification issue of clustering ?
2. I think it is necessary to give some details about how to compute the distance between two clusterings, i.e., \Delta (C, C'). In general, it is an NP-hard problem to evaluate \Delta (C, C').
3. In Theorem 7 and Theorem 8, I am curious about what does the term ``k-means SDP'' refer to, the 0-1 SDP or the relaxed SDP ?
4. Regarding the term ``noise-robust'', I think it is more accurate to use ``outlier'' instead of ``noise''. Usually, noise is used to standard for the white Gaussian noise, and what addressed in the paper is actually outlier.

2. Please provide an overall score for the submission.

Accept: Good paper

3. Please enter a 1-2 sentence summary of your review explaining your overall score.

This is a theoretical paper contains some valuable results.

5. Please rate your confidence in the score assigned.

Medium: Reviewer has understood the main points in the paper, but skipped the proofs and technical details.

Reviewer #3

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

The paper considers the problem of robust k-means clustering. Here we assume that the dataset can be subdivided into two parts: the majority of the data is a dataset which admits a good clustering, either deterministically or because of some generative model, and a small fraction of the data is arbitrary. The goal of the paper is to recover the clustering of the well-clusterable part of the data, given this corrupted dataset. They propose a modification of the a previously proposed k-means SDP objective based on a regularization-based paradigm for robustification, and demonstrate that this algorithm is able to recover the true clustering in the presence of corrupted data, under suitable conditions on the separation between the clusters. The results are based on improving and generalizing previous techniques for constructing SDP dual certificates.

The results for clustering with only deterministic conditions on the data look quite interesting, though I am not an expert in the area. However, I have concerns about the novelty of the results for clustering under the stochastic ball assumption. In particular, I am confused as to why it is not immediately implied by (and improved upon) by the results in [1, 2]. Both of these papers introduce similar techniques for clustering in the presence of noise, under distributional assumptions on the data. Specifically, if the distribution of each cluster is given by D_i with mean mu_i and covariance Sigma_i, where Sigma_i \preceq \sigma^2 I, these papers are able to achieve clustering with separation at most k^eps sigma for any eps > 0, assuming that each D_i is nice. Unless I am mistaken, it is easily verified that the isotropic ball distribution satisfies the necessary conditions with sigma = 1/sqrt(d), so this would imply that these algorithms better results for this setting than in the paper.

[1] Mixture Models, Robustness, and Sum-of-Squares Proofs. Sam B. Hopkins and Jerry Li. In STOC 2018.

[2] Improved clustering and robust moment estimation via sum-of-squares. Pravesh Kothari, Jacob Steinhardt, David Steurer. STOC 2018.

Minor notes:

- Definition 4 is rather informal; what does “any robustification paradigm” mean?

Update: given that there has been no author feedback, my score remains the same.

2. Please provide an overall score for the submission.

Weak Reject: Borderline, tending to reject

3. Please enter a 1-2 sentence summary of your review explaining your overall score.

Potentially interesting results under worst case assumptions, but results under distributional assumptions appear superceded by other work.

Unfortunately the authors did not respond to any of the questions raised by the reviewers.

5. Please rate your confidence in the score assigned.

High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.