Information and Computation

Recovery guarantees for exemplar-based clustering

Abhinav Nellore, Rachel Ward

PII: S0890-5401(15)00097-8
DOI: http://dx.doi.org/10.1016/j.ic.2015.09.002
Reference: YINCO 4121

To appear in: *Information and Computation*

Received date: 19 April 2014

Recovery guarantees for exemplar-based clustering

Abhinav Nellore

The Johns Hopkins University, anellore@gmail.com

Rachel Ward

The University of Texas at Austin, rward@math.utexas.edu

Abstract

For a certain class of distributions, we prove that the linear programming relaxation of k-medoids clustering—a variant of k-means clustering where means are replaced by exemplars from within the dataset—distinguishes points drawn from nonoverlapping balls with high probability once the number of points drawn and the separation distance between any two balls are sufficiently large. Our results hold in the nontrivial regime where the separation distance is small enough that points drawn from different balls may be closer to each other than points drawn from the same ball; in this case, clustering by thresholding pairwise distances between points can fail. We also exhibit numerical evidence of high-probability recovery in a substantially more permissive regime.

Keywords: exact recovery, k-medoids, linear programming, separated balls

2010 MSC: 91C20, 52A41, 62-07,

1. Introduction

Consider a collection of points in Euclidean space that forms roughly isotropic clusters. The *centroid* of a given cluster is found by averaging the position vectors of its points, while the *medoid*, or exemplar, is the point *from within the collection* that best represents the cluster. To distinguish clusters, it is popular to pursue the k-means objective: partition the points into k clusters such that the average squared distance between a point and its cluster centroid is minimized. This problem is in general NP-hard [1, 2]; practical algorithms like Lloyd's [3] and Hartigan-Wong [4] typically converge to local optima. k-medoids clustering[1] is also in general NP-hard [5, 6], but it does admit a linear programming (LP) relaxation. The objective is to select k points as medoids such that the average squared distance (or other measure of dissimilarity) between a point and its medoid is minimized. **This paper obtains guarantees for exact recovery of the unique globally optimal solution to the k-medoids integer program by its LP relaxation.** Commonly used algorithms that may only converge to local optima include partitioning around medoids (PAM) [7, 8] and affinity propagation [9, 10].

[1] k-medoids clustering is sometimes called k-medians clustering in the literature.