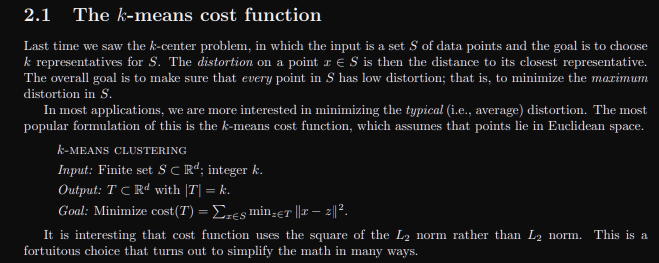
Project Description

Clustering Distortion cost



2.1 The k-means cost function Last time we saw the k-center problem, in which the input is a set S of data points and the goal is to choose k representatives for S. The distortion on a point x ∈ S is then the distance to its closest representative. The overall goal is to make sure that every point in S has low distortion; that is, to minimize the maximum distortion in S. In most applications, we are more interested in minimizing the typical (i.e., average) distortion. The most popular formulation of this is the k-means cost function, which assumes that points lie in Euclidean space. k-means clustering Input: Finite set S ⊂ R d ; integer k. Output: T ⊂ R d with |T| = k. Goal: Minimize cost(T) = P x∈S minz∈T kx − zk 2 . It is interesting that cost function uses the square of the L2 norm rather than L2 norm. This is a fortuitous choice that turns out to simplify the math in many ways.