K-Means Conjecture

February 2024

1 Introduction

We consider an approximation algorithm for obtaining cluster centers in an adaptive way in 1D K-Means clustering (Lloyd-Max quantization). In this algorithm, we first obtaining cluster centers for K number of clusters. Then, for obtaining cluster centers for $\tilde{K} < K$ number of clusters, we cluster the cluster centers (instead of the data) previously obtained in a weighted way. We believe these new cluster centers are close to the optimal ones in the sense that the objective value is bounded by a constant times the optimal objective value of K-Means.

2 Statement of Conjecture

In the following analysis, we assume that the argmin operator returns a single element with ties broken arbitrarily. Consider a dataset $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}$. The K-Means objective for $K > 1$ clusters is given by

$$f_K(\mathcal{X}) \coloneqq \min_{c_1,\dots,c_K} \sum_{i=1}^{n} \min_k (x_i - c_k)^2. \tag{1}$$

Let $c_1^*, \dots, c_K^*$ denote the optimal cluster centers and $s(i)$ denote the optimal cluster assignment:

$$c_1^*, \dots, c_K^* = \operatorname*{argmin}_{c_1,\dots,c_K} \sum_{i=1}^{n} \min_k \|x_i - c_k\|^2 \tag{2}$$

$$s(i) = \operatorname*{argmin}_k \|x_i - c_k^*\|^2. \tag{3}$$

Let n_j denote the number of datapoints assigned to the jth cluster:

$$n_j = |\{i : s(i) = j\}|. \tag{4}$$

We now consider the adaptive K-Means procedure to get cluster centers for $\tilde{K} < K$ clusters. The $\tilde{K}$ centers obtained by adaptive K-Means (starting from

K clusters) is given by:

$$\tilde{c}_1, \ldots, \tilde{c}_{\tilde{K}} = \underset{c_1, \ldots, c_{\tilde{K}}}{\operatorname{argmin}} \sum_{j=1}^{K} \min_k n_j \left\| c_j^* - c_k \right\|^2 \tag{5}$$

The K-means objective value evaluated on $\{\tilde{c}_1, \ldots, \tilde{c}_{\tilde{K}}\}$ is given by

$$\mathrm{Obj}_{K \to \tilde{K}}(\mathcal{X}) \coloneqq \sum_{i=1}^{n} \min_k (x_i - \tilde{c}_k)^2. \tag{6}$$

We conjecture that the following bound holds for any dataset $\mathcal{X}$ and for all $K > 1$ and $\tilde{K} < K$:

$$\mathrm{Obj}_{K \to \tilde{K}}(\mathcal{X}) \leq 2 f_{\tilde{K}}(\mathcal{X}). \tag{7}$$

3 An example of achievability

Consider $\mathcal{X} = \{0, 1 + \epsilon, 2 - \epsilon, 3\}$, where $\epsilon > 0$ is infinitesimally small. Consider $K = 3$ and $\tilde{K} = 2$. The optimal cluster centers for $K = 3$ clusters are $\{0, 3/2, 3\}$. The cluster centers for $\tilde{K} = 2$ clusters obtained adaptively are $\{0, 2\}$. The optimal cluster centers for $\tilde{K} = 2$ clusters are $\{1/2, 5/2\}$. In this case, $\mathrm{Obj}_{K \to \tilde{K}}(\mathcal{X}) = 2 f_{\tilde{K}}(\mathcal{X}) = 2$.