k-Means Clustering

Oliver Zhao

1 Description

k-means clustering is a form of vector quantization that separates n observations into k clusters. These clusters are Voronoi cells, where each observation belongs to the cluster with the closest centroid. For learning purposes, we demonstrate the naive algorithm, although usually there are heureistics that allow that clusters to converge quickly. The algorithm converges when the centroids of each cluster no longer changes, which is not always guaranteed, particularly when distances other than the Euclidean distance are used.

2 Algorithm

Description: Consider an initial set of k means $m_1^{(1)}, \ldots, m_k^{1)}$. The algorithm iterations t times through the two steps until convergence.

---

1. **Assignment Step:** Assign each observation x_p to the cluster $S_i^{(t)}$ with the nearest mean, where

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \ \forall j, \ 1 \leq j \leq k\}. \tag{2.1}$$

2. **Update Step:** Recalculate the centroids for observations assigned to each cluster:

$$m_i^{t+1} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \epsilon S_i^{(t)}} x_j. \tag{2.2}$$
