11/9/19 EE5601: Representation Learning

- K-means clustering
  - Problem formulation
  - Lloyd's algorithm
  - Demo

- Review: Goal of K-means algorithm is to cluster $\{\underline{x}_i\}_{i=1}^{N}$ into K clusters $\mathcal{S} = \{S_1, \ldots, S_K\}$ such that the intra-class variance is minimized. Recall, $\underline{x}_i \in \mathbb{R}^d$.

Soln: This is an NP hard problem. We will look at an approximate soln. Specifically, the Lloyd's algorithm.

1 Given, K, $\{\underline{x}_i\}_{i=1}^{N}$, ε

2. Initialize centroids $C^{(0)} = \{\underline{\mu}_1^{(0)}, \ldots \underline{\mu}_K^{(0)}\}$; $\underline{\underline{j=0}}$ iteration

3. Assign $\underline{x}_n$ to cluster 'i' if $\to S_i^{(j)}$ [For all n]

$$i = \arg\min_{k \in \{1,\ldots K\}} \|\underline{x}_n - \underline{\mu}_k^{(j)}\|_2^2 \quad ; \quad d(\underline{x}_n, \underline{\mu}_i)$$

4. Update centroids : $\underline{\mu}_k^{(j+1)} = \frac{1}{N_k} \sum_{i \in S_k^{(j)}} \underline{x}_i$

5. Check for stopping condition.

$$\text{If}\left[\left(\sum_{k=1}^{K} \|\underline{\mu}_k^{(j+1)} - \underline{\mu}_k^{(j)}\|_2^2\right) < \underbrace{\varepsilon}_{\text{threshold}}\right], \text{ output centroids and clusters.}$$

Else, goto step 3.