Lec 19 Unsupervised Learning

1. Notation:

↳ $\mathcal{D} = \{ \underline{x}_1, \underline{x}_2 \dots \underline{x}_N \}$, $x_i \in \mathbb{R}^d$.

No labels

e.g. $\mathcal{D}$: set of documents Goal : group by topic

$\underline{x}_i$: histogram of word lengths in documents i

↳ ① Clustering ✓ ② Density estimation ✓ ③ Dimensionality reduction (e.g. PCA) (Not in this course)

2. Clustering: (ch 6.3.3)

↳ We want partitation $\mathcal{D}$ into k disjoint clusters such that the elements in each cluster are close to each other.

↳ Given $\mathcal{D}$, we want output:

① Clusters $S_1, S_2, \dots S_k$, where $S_i \subseteq \mathcal{D}$

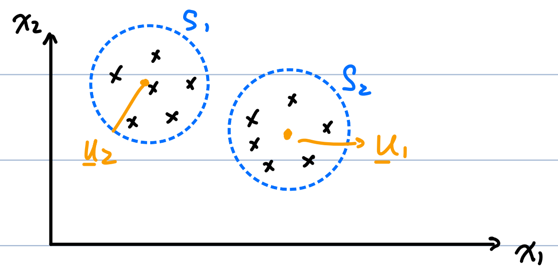
s.t. $S_i \cap S_j = \emptyset \quad \forall\, i, j$

$\bigcup_{i=1}^{k} S_i = \mathcal{D}$

② Clusters center:

$u_1, u_2 \dots u_k$, $u_i \in \mathbb{R}^d$

$d=2$
$k=2.$



↳ Error Measure :

Distance from each point to the clusters center

$E_j = \sum_{\underline{x}_n \in S_j} \| \underline{x}_n - \underline{u}_j \|^2$ // approx. error for cluster S_j

Given $\mathcal{D}$ and k, define

$$E_{in}(S_1, S_2 \ldots S_k, \underline{\mu}_1, \underline{\mu}_2 \ldots \underline{\mu}_k) = \sum_{j=1}^{k} E_j$$
$$= \sum_{n=1}^{N} \| \underline{x}_n - \underline{\mu}(\underline{x}_n) \|^2$$

where $\underline{\mu}(\underline{x}_n) \equiv$ center of cluster to which $\underline{x}_n$ belongs

↳ Learning Problem:

$$\min_{\substack{S_1, S_2 \ldots S_k \\ \mu_1, \mu_2 \ldots \mu_k}} E_{in}$$

↳ Application:

- Classification (e.g. documents, coins)
- Recommendation system

↳ Optimal clustering is NP-hard.

$\therefore$ Need to use heuristic

$\Rightarrow$ k-means clustering

3. k-means Clustering:

↳ An alternating optimizing approach with two subproblems:

↳ Subproblem 1:

Given $S_1 \ldots S_k$, find $\underline{\mu}_1, \underline{\mu}_2 \ldots \underline{\mu}_k$ to minimize E_{in}

$$E_{in} = \sum_{j=1}^{k} E_j$$

$E_j = \sum_{\underline{x}_n \in S_j} \| \underline{x}_n - \underline{\mu}_j \|^2$, depends only on S_j

$\therefore$ only needs to consider $\min_{\underline{\mu}_j} E_j(\underline{\mu}_j) = \min_{\underline{\mu}_j} \sum_{\underline{x}_n \in S_j} \| \underline{x}_n - \underline{\mu}_j \|^2$

$\because$ want $\nabla E_j(\underline{\mu}_j) = 0$

$\therefore \nabla_{\underline{\mu}_j} \sum_{\underline{x}_n \in S_j} \| \underline{x}_n - \underline{\mu}_j \|^2 = 0$

$$\sum_{\underline{x}_n \in S_j} -2(\underline{x}_n - \underline{\mu}_j) = 0$$

$$\sum_{\underline{x}_n \in S_j} \underline{x}_n = \underline{\mu}_j |S_j|$$

$\therefore \underline{\mu}_j = \frac{1}{|S_j|} \sum_{\underline{x}_n \in S_j} \underline{x}_n$ // This is the avg of sample point in S_j

"centroid"

↳ Subproblem 2:

Given $\underline{\mu}_1 \dots \underline{\mu}_K$, find $S_1 \dots S_K$ to minimize E_{in}

i.e. Given $\underline{x}_i \in \mathcal{D}$, with which cluster should it be associated?

- $\underline{\mu}(\underline{x}_i)$: cluster center for $\underline{x}_i$

$$E_{in} = \sum_{n=1}^{N} \| \underline{x}_n - \underline{\mu}(\underline{x}_n) \|^2$$

$\therefore$ The min. of $\| \underline{x}_n - \underline{\mu}(\underline{x}_n) \|$ depends only on $\underline{x}_n$ and $\underline{\mu}(\underline{x}_n)$

$\because \underline{\mu}(\underline{x}_n) = \underset{\underline{\mu} \in \{\mu_1 \dots \mu_K\}}{\arg\min} \| \underline{x}_n - \underline{\mu} \|$

$\therefore$ Assign $\underline{x}_n$ to the nearest cluster center.