Class (Sept. 10, Thursday)

Application of concepts of distance / similarity.

<u>task:</u> certain collection of $\boxed{\text{objects}}$ is given.

find $\boxed{\text{"similar"}}$ objects in one class.

vectors / feature vectors $\in \mathbb{R}^n$

collection vectors which are "close"

Objects $\longleftrightarrow$ vectors

similarity $\longleftrightarrow$ distance

groups $\longleftrightarrow$ clusters (unsupervised / semi-supervised)

$x_1, x_2, \ldots, x_N \in \mathbb{R}^n$

$c_1, c_2, \ldots, c_n \leftarrow$ numbers / cluster assignment.

$x_1, x_2, \dots, x_N \in \mathbb{R}^n$ (Data / vectors)

Total - 3 clusters.

$c_i \in \{1, 2, 3\}$ for $i = 1, 2, \dots, N$

$N = 6$,

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | ← data / vectors / feature vectors |
|---|---|---|---|---|---|---|---|
| $c_i$ | 3 | 2 | 3 | 1 | 2 | 1 | |
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | ← cluster assignment |

$[G_1 = \{4, 5\}, \quad G_2 = \{x_2, x_5\}, \quad G_3 = \{1, 3\}]$

clusters

<u>cluster representative</u>: (Group representative)

N - vectors, k - clusters

N- data points $\quad$ k - clusters

Let $z_1, z_2, \ldots, z_k$ : cluster representatives.

Representatives are "close" to the vectors in a cluster.

$\|x_i - z_{c_i}\|$ should be small.

$x_1, x_2, \ldots, x_N$ - vectors.

$c_1, c_2, \ldots, c_N$ — cluster assignments *

$c_1, \ldots, c_N \in \{1, 2, \ldots, k\}$

$z_1, z_2, \ldots, z_k$ : cluster representatives.

$z_{c_i} \in \mathbb{R}^n$

$N = 100, \; k = 5$

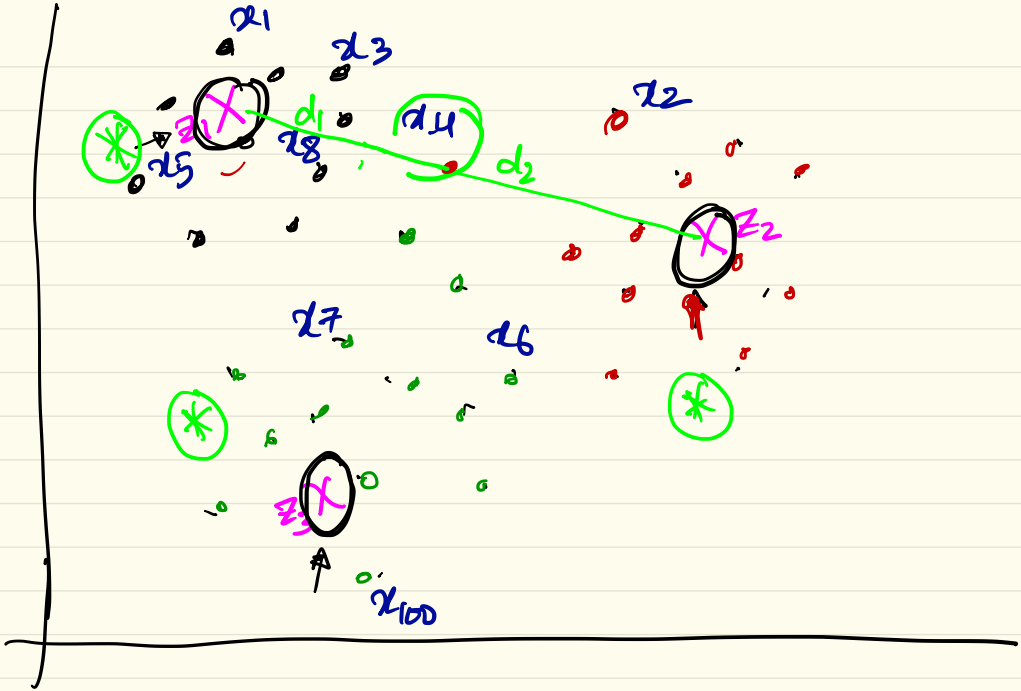
$$\begin{array}{ccccc} x_1, & x_2, & \cdots & , & x_{100} \\ 1 & 5 & 1 & & \\ \| & \| & \| & & \\ c_1, & c_2 & c_3 & & c_{100} \end{array}$$

$x_i \longleftrightarrow c_i \qquad z_{c_i} \qquad c_i = 1, \qquad z_{c_i} = z_1$

---

clustering objective:

$$J^{clust} = \frac{\left( \|x_1 - z_{c_1}\|^2 + \cdots + \|x_N - z_{c_N}\|^2 \right)}{N}$$




$k = 3$

$N = 100$

$z_1, z_2, z_3$

: cluster representatives.

$c_1, c_2, \ldots, c_{100}$

$\in \{1, 2, 3\}$

$d_1 = d_2$

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| $c_1 = 1$ | $c_2 = 2$ | $c_3 = 1$ |
| $z_{c_1} = z_1$ | $z_{c_2} = z_2$ | $z_{c_3} = z_1$ |
| $\underbrace{\|x_1 - z_{c_1}\|}$ | $\underbrace{\|x_2 - z_{c_2}\|}$ | $\underbrace{\|x_3 - z_{c_3}\|}$ |

$$J^{clust} = \frac{\|x_1 - z_{c_1}\|^2 + \|x_2 - z_{c_2}\|^2 + \cdots + \|x_N - z_{c_N}\|^2}{N}$$

$J^{clust}$ = sum of squares $\geq 0$

$= 0$ (best case)

clustering objective: minimize $J^{clust}$.

$k$ : How many number of clusters. $(k < N)$ $\longleftrightarrow$ (1)

$c_1, \ldots, c_N$ : assignment. $\longleftrightarrow$ (*)

$z_1, \ldots, z_k$ : choice of cluster representatives. $\longleftrightarrow$ (2)

Optimal clustering: minimize $J^{clust}$ : decision variables (*)

Suboptimal clustering: