Problem: Given N vectors $x_1, x_2 \cdots x_N \in \mathbb{R}^n$ and $k > 0$, partition / group these N vectors into k groups such that vectors "close" are in the same group.

distance between vectors — norms

$d(x_1, x_2)$

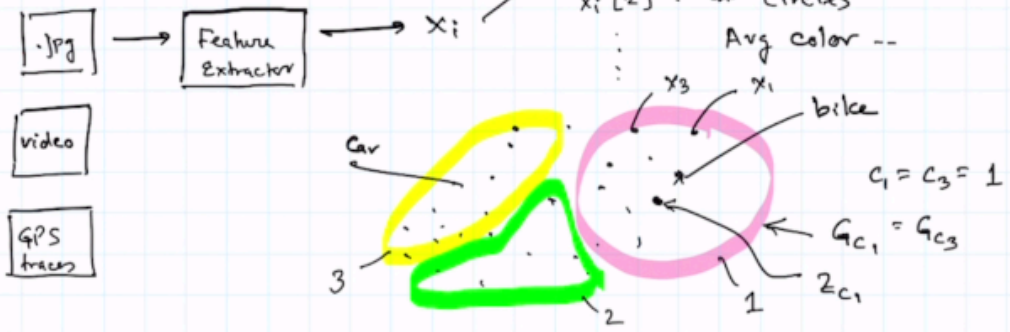
$f: \mathbb{R}^n \to \mathbb{R}_{\geq 0}$

homogeneous $f(ax) = af(x)$

Δ inequality $f(x+y) \leq f(x) + f(y)$

Definite $f(x) = 0 \iff x = 0$

Applications: Object recognition



Notations $x_1, x_2 \cdots x_N \in \mathbb{R}^n$ given set of points

k : Number of clusters

$c_1, c_2 \cdots c_N \in \{1, \cdots, k\}$ Cluster that x_i belongs to

G_{c_i} : Set of x_j's in the same cluster as x_i

$z_{c_i} \in \mathbb{R}^n$ representative point for G_{c_i}

Restating the objective: Choose $c_1, c_2 \cdots c_k$ and the representatives such that

$$J_{clust} = \frac{1}{N} \sum_{i=1}^{k} |x_i - z_{c_i}|^2$$

Two-Step Algorithm : Step 1

Suppose the representatives are given

$z_1 \cdots z_k$ fixed

How to assign the x_i's to each representative / group?

$$\min J_{clust} = \min_{j \in \{1, \cdots, k\}} \frac{1}{N} \sum_{i=1}^{N} |x_i - z_j|^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} \min_{j \in \{1, \cdots k\}} |x_i - z_j|^2$$

$$c_i = \arg\min_j |x_i - z_j|^2$$

Assign x_i to the z_j that is closest to x_i

Step 2: Suppose the groups are fixed.

How to assign representatives for each group?

$$J_{clust} = \frac{1}{k} (J_1 + J_2 \cdots + J_k)$$

$$=$$

$$\min_{z_1 \cdots z_k} J_{clust} = \min_{z_1 \cdots k} \frac{1}{k} \sum_{j=1}^{k} \left(\frac{1}{|G_j|} \sum_{x_i \in G_j} |x_i - z_j|^2 \right)$$

k

$$\min_{z_1 \dots z_k} J_{chat} = \min_{z_1 \dots k} \frac{1}{k} \sum_{j=1}^{k} \left(\frac{1}{|G_j|} \sum_{x_i \in G_j} |x_i - z_j|^2 \right)$$

$$= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{|G_j|} \min_{z_j} \sum_{x_i \in G_j} |x_i - z_j|^2$$

$$T = \sum_{i \in G_j} (x_i^2 + z_j^2 - 2 x_i z_j)$$

$$\frac{\partial T}{\partial z_j} = 0 + \sum_{i \in G_j} 2 z_j - 2 \sum_{i \in G_j} x_i = 0$$

$$|G_j| z_j = \sum_{i \in G_j} x_i$$

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$

z_j = mean of G_j