05/24/2019 Discussion # 8

k-mean:

$D = \{\underline{x}_1, \ldots, \underline{x}_N\}$

objective function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\underline{x}_n - \underline{\mu}_k\|_2^2$$

$$r_{nk} \in \{0, 1\}, \quad r_{nk} = \begin{cases} 1 & x_n \text{ belong } z_k \\ 0. & \text{otherwise.} \end{cases}$$

Algorithm

$z_1 \ldots z_j$

choose initial 'representatives' $z_1, \ldots, z_k$ for the k groups and repeat:

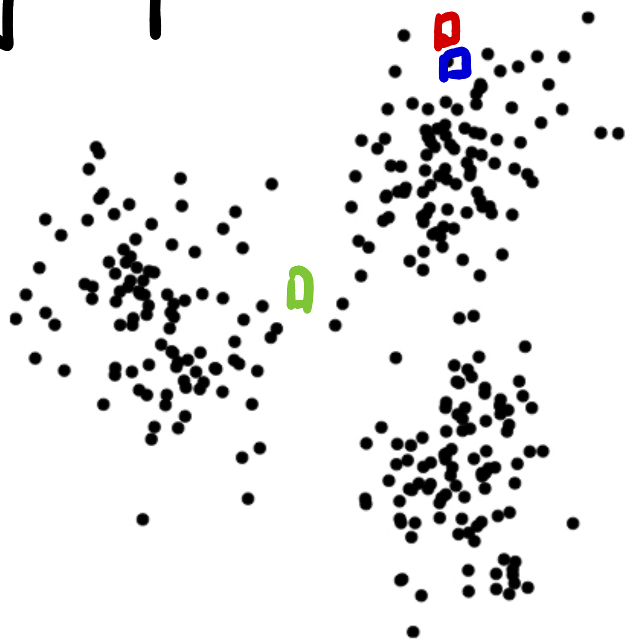
1. assign each vector x_i to the nearest group representative z_j update r_{nk}.
2. set the representative z_j to the mean of the vectors assigned to it

initial point:

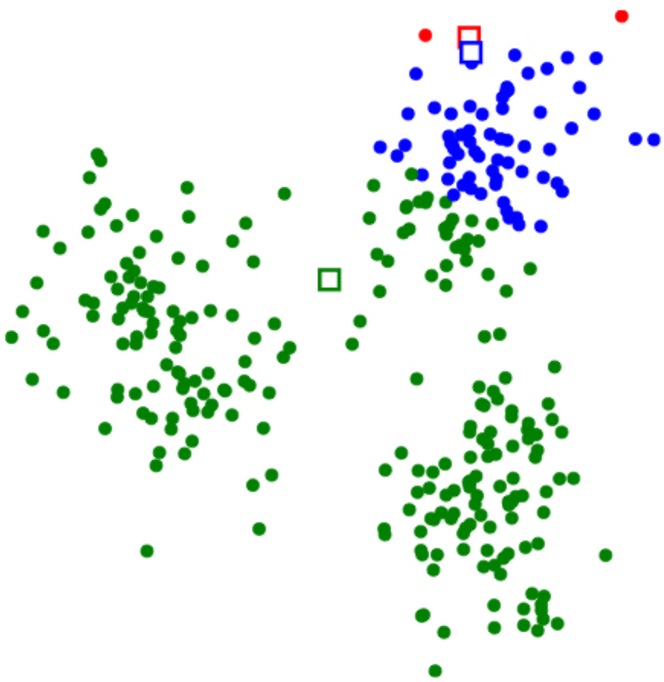
random, or farest apart.

Goal: 3 group.

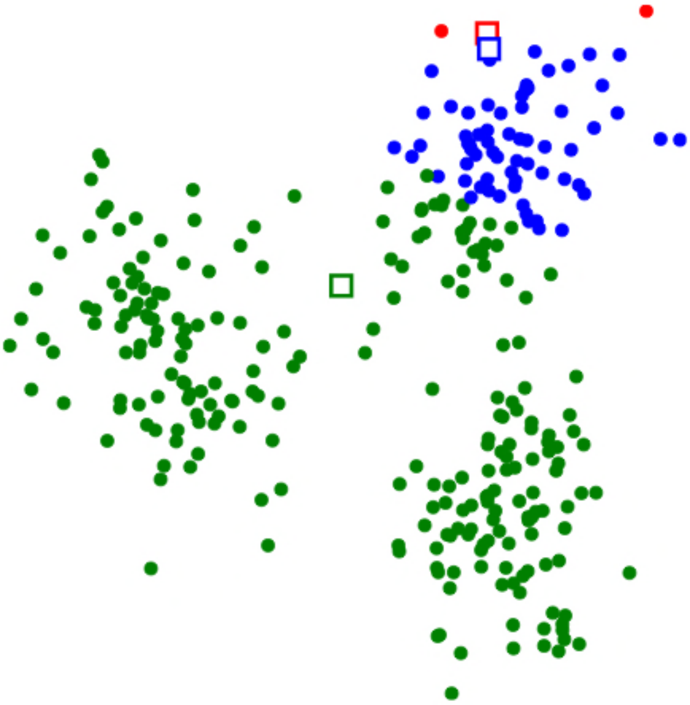
Initial

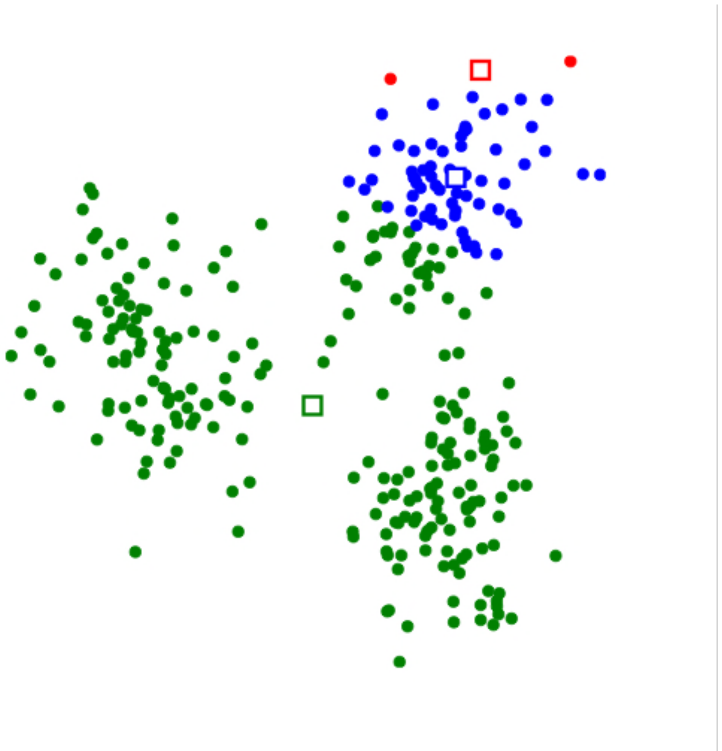


Itr: 1

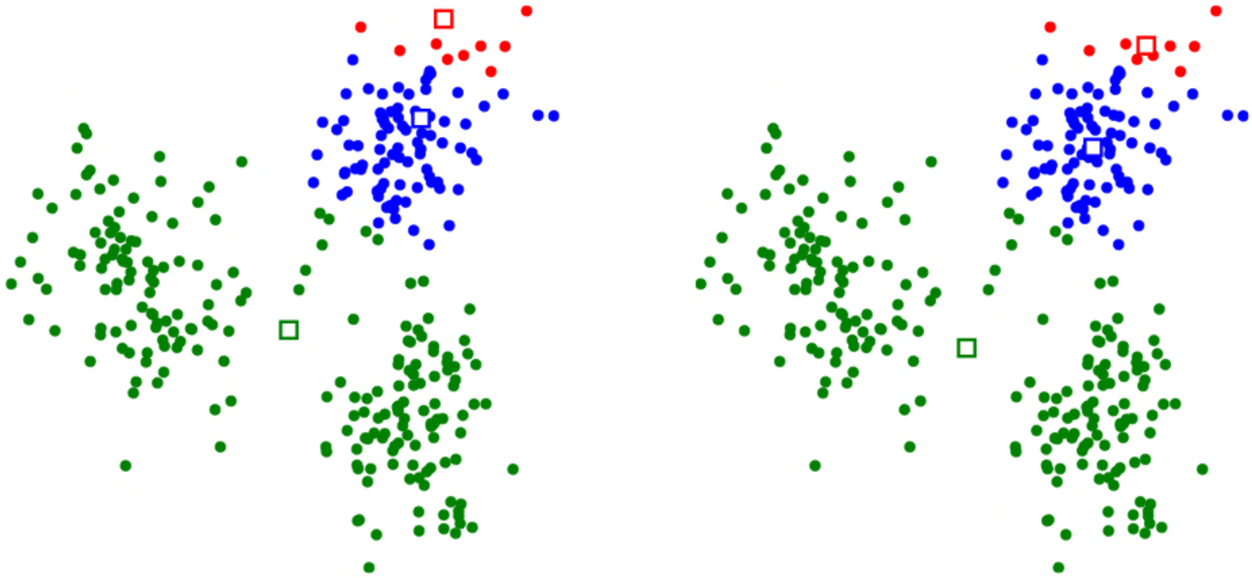


ItR : 1

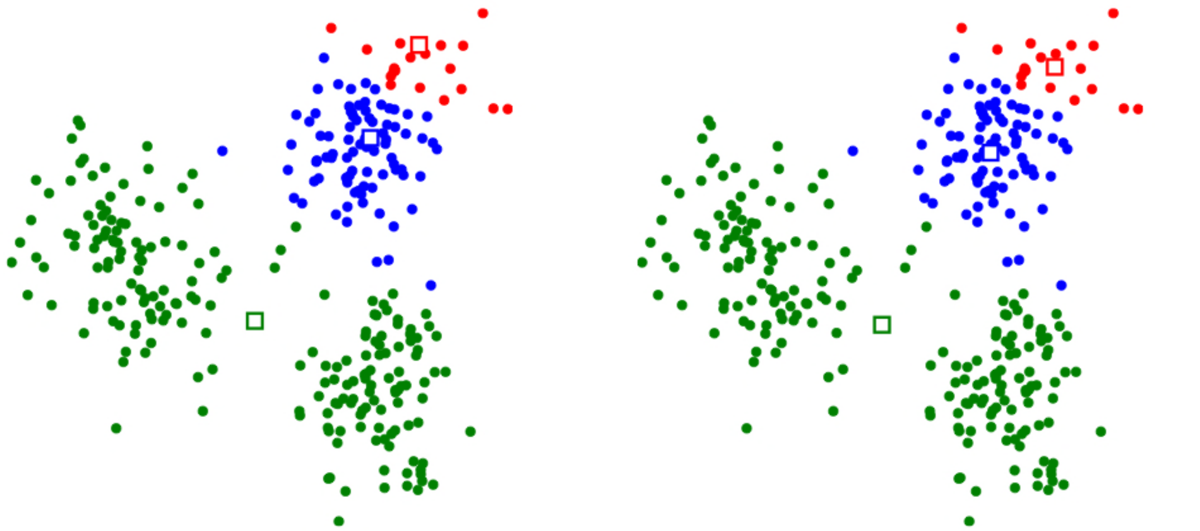




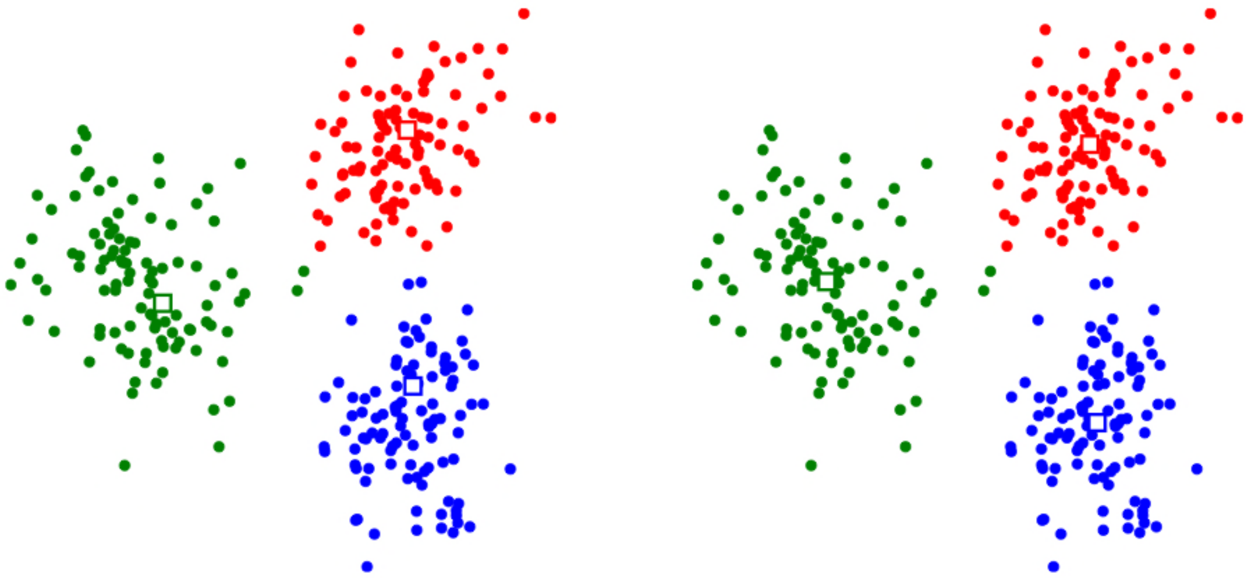
ITR : 2



ITR 3 :



ITR : 11



stop : u_k didn't change that much.

$$r_{nk} \in \{0,1\}, \quad r_{nk} = \begin{cases} 1 & x_n \text{ belong } z_k \\ 0. & \text{otherwise.} \end{cases}$$

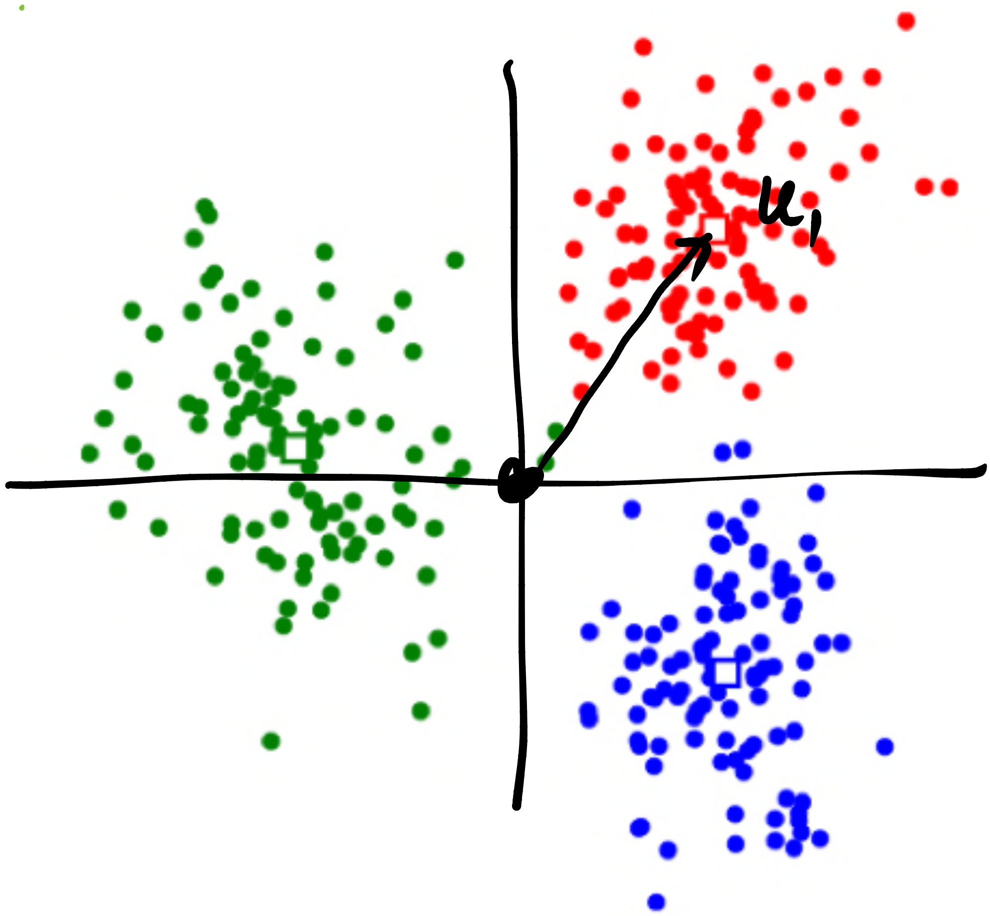
k-mean w/o regulization:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2$$

k-mean with regulization

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2 + \sum_{k=1}^{K} \lambda \underline{\|u_k\|_2^2}$$

e.g.



k-mean with regulization

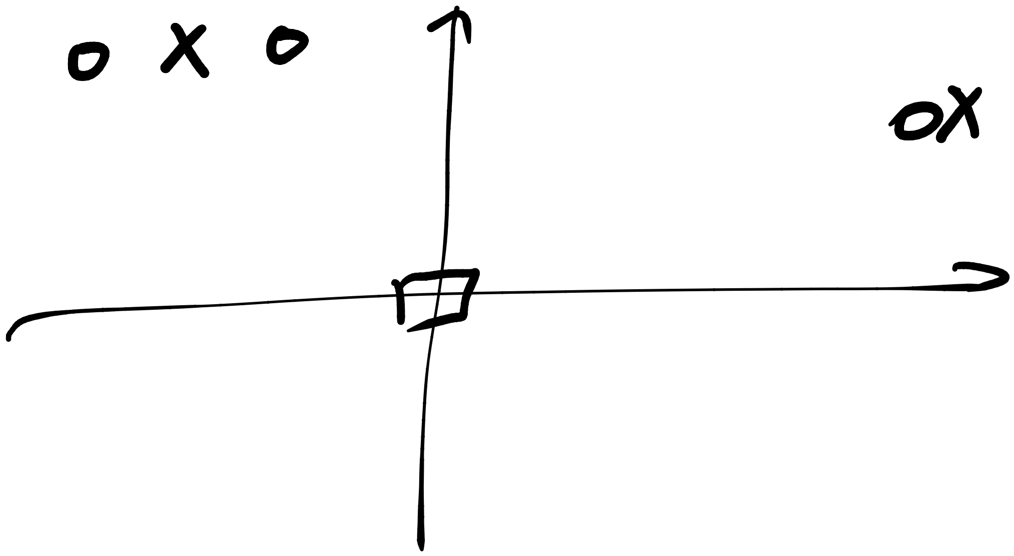
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2 + \sum_{k=1}^{K} \lambda \|\mu_k\|_2^2$$

There are K teaching assistants who must collect the answer scripts once the time is up. The TAs need to figure out good locations to position themselves so that the students can walk to the nearest TA and submit their answers. Once the TAs have all the answer sheets, they must return to the front desk located at (0,0) while handling the returned answer sheets carefully. To reduce the possibility of mishaps related to handling of the papers, write down an objective which can be used to minimize the total distance that both students and TAs need to walk to bring the papers to the front desk. Assume thateveryone can walk by taking the shortest path between two points.

$K=2$.

O: students

X: TA



Data set:

$$D = \{1, 2, 3, 3, 3, 4, 5, 6, 7, 8, 10\}$$

$|D| = 11 \qquad x_i \in D.$

$$\arg\min_{\mu_2} \sum_{i=1}^{N} (x_i - \mu_2)^2 \leftarrow \text{mean}$$

$$\arg\min_{\mu_1} \sum_{i=1}^{N} |x_i - \mu_1| \leftarrow \text{median}$$

$$\arg\min_{\mu_0} \sum_{i=1}^{N} (x_i - \mu)^0 = Y \qquad x^0 = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & x = 0 \end{cases}$$

mode

mode, mean, median.

$\mu_0 \notin D \quad , Y = N,$

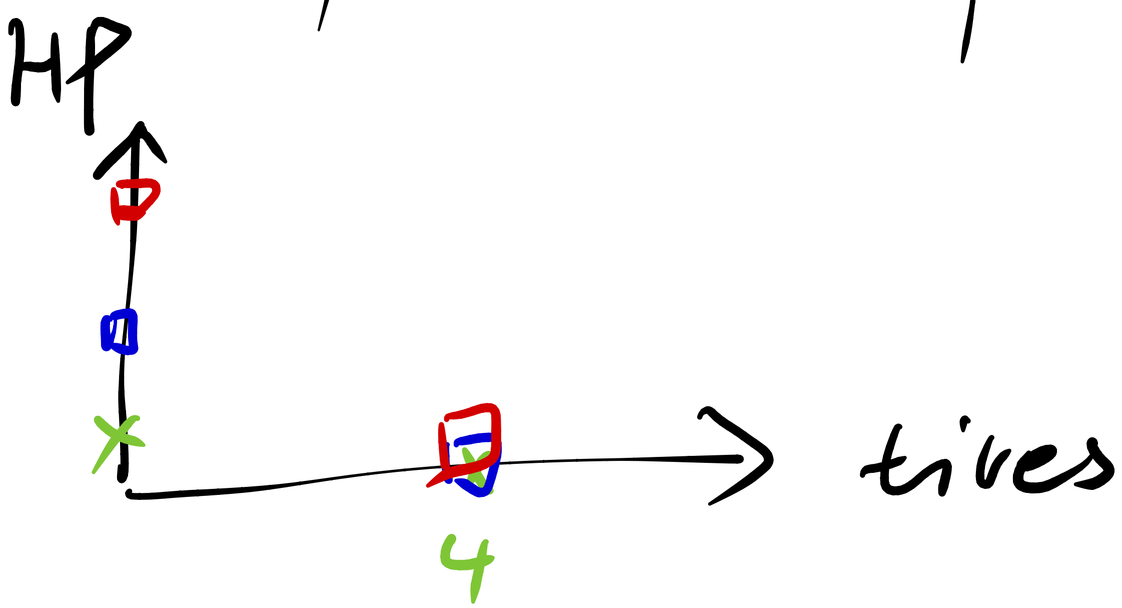
$\mu_0 \in D \qquad Y = N - |\{x_i | x_i = \mu\}|$

PCA:

why PCA?

e.g.

	# of tires	HP	Y: price
car$_1$	4	280	20k
car$_2$	4	80	1k
car$_3$	4	350	100k



$$\max_{u_1} \quad u_1^T S u_1$$

$$\text{s.t.} \quad \|u_1\|_2^2 = 1$$

$S = \frac{1}{N}(X^T X)$

$\downarrow$

SVD

$$L(u_1, \lambda) = u_1^T S u_1 + \lambda(\|u_1\|_2^2 - 1)$$

$$2Su_1 - 2\lambda u_1 = 0$$

$$Su_1 = \lambda_1 u_1$$

$$X = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$$

$$\max_{u_2} \quad u_2^T S u_2$$

$$\text{s.t.} \quad \|u_2\|_2^2 = 1$$

$$u_1^T u_2 = 0$$

$$\max_{u_2} u_2^T S u_2 \quad ①$$

$$\text{s.t. } \|u_2\|_2^2 = 1$$

$$u_1^T u_2 = 0$$

$$L(u_2, \lambda_2, \eta_2) = \underline{u}_{m+1}^T \underline{\underline{S}}\, \underline{u}_{m+1} + \lambda_2(1 - \underline{u}_2^T u_2)$$

$$+ \eta_2 u_2^T u_1$$

$$\frac{\nabla L(u_2, \lambda_2, \eta_2)}{\partial u_2} = 0$$

$$\Rightarrow 0 = 2\underline{\underline{S}}\,\underline{u}_2 - 2\lambda_2 \underline{u}_2 + \eta_2 \underline{u}_1$$

$$\Rightarrow S\underline{u}_2 = \lambda_2 \underline{u}_2$$

$\lambda_2 = u_2^T S u_2$ → ① what we try to maximize.

but $\underline{u}_2 \neq \underline{u}_1$, $\therefore \lambda_2 \neq \lambda_1$.

$\therefore \lambda_2$ is the second largest eigenvalue.