1. K-mean Clustering.

We have a date set $\{x_1, \cdots, x_N\}$, we want to find out K clusters. We will suppose that K is given.

Our setting:

(1) Each point x_i will be assigned to exact one cluster

(2) The distance is measured by $\|\cdot\|_2$

(3) Cluster is represented by it's sample mean

We will use an indicator r_{nk} to show whether x_n is assigned to cluster k. We are trying to minimize an overall distance which is called distortion measure.

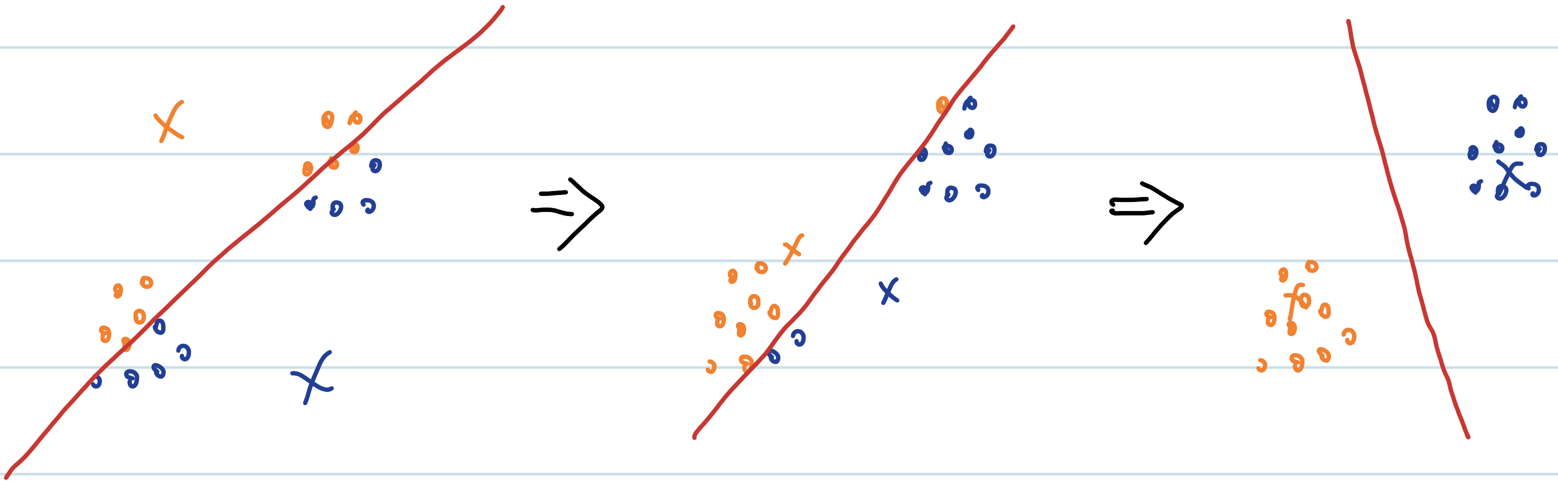
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \underset{\substack{\uparrow \\ \text{center of cluster } k}}{\mu_k} \|_2^2$$

if μ_k is given, r_{nk} is easy to be determined.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_k \|x_n - \mu_k\|^2 \\ 0 & \text{o.w} \end{cases}$$

Now, let's consider to optimize μ_k. when r_{nk} is fixed.

$$\frac{\partial}{\partial \mu_k} J = 2 \sum_{n=1}^{N} r_{nk} (x_n - \mu_k) = 0 \Rightarrow \mu_k = \underbrace{\frac{\sum r_{nk} x_n}{\sum r_{nk}}}_{\text{Cluster's sample mean.}}$$



We deliberately choose bad initial points. but we can still converge very well. K-mean is often used to initial the parameters in a Gaussian mixture model befering applying EM.

K-mean method can also be used to do image segmentation.