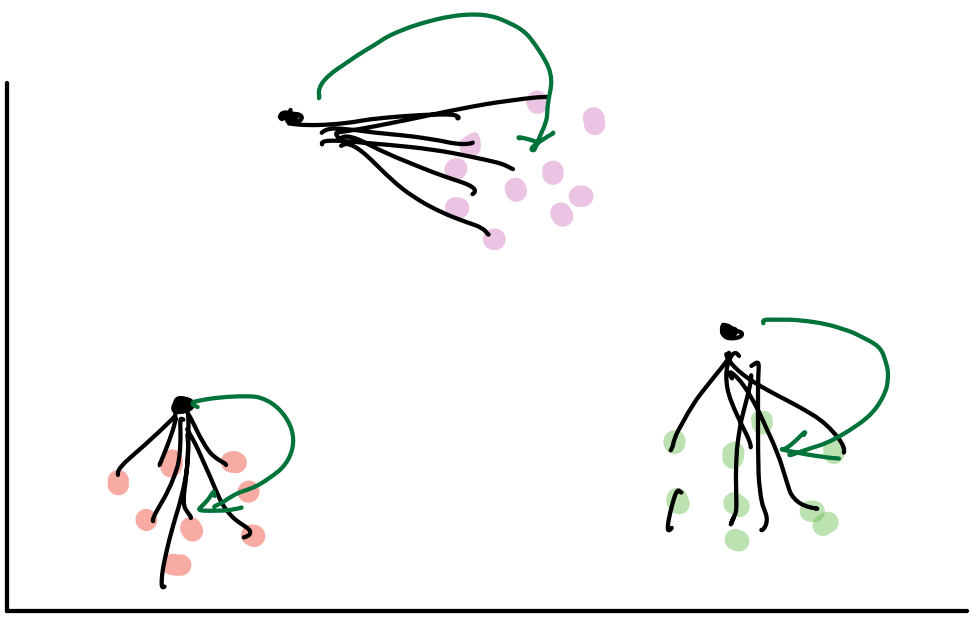
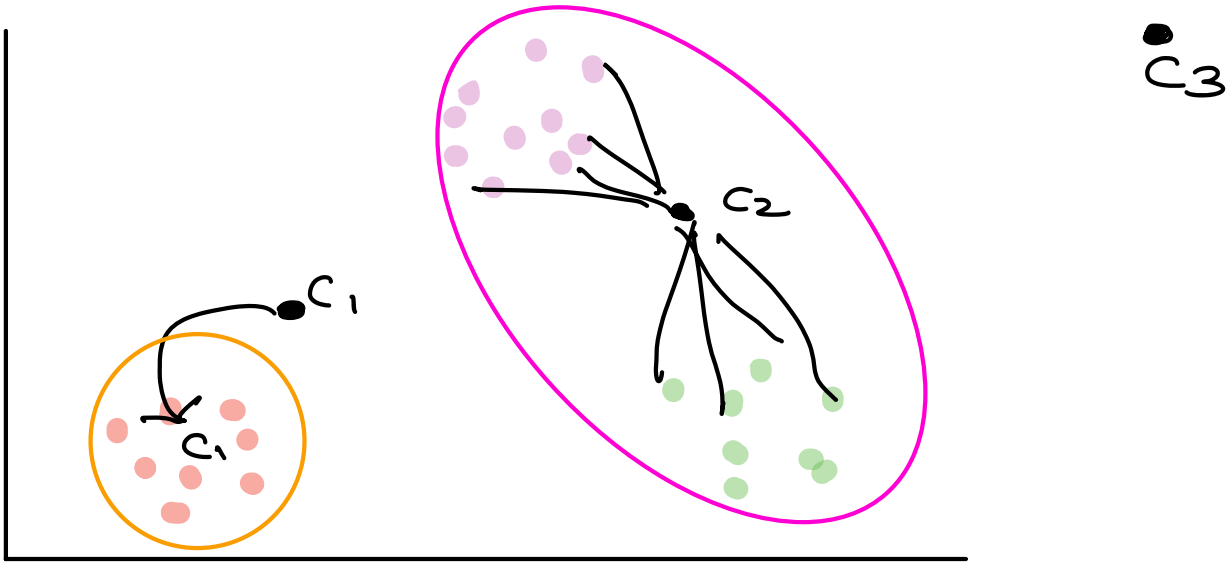
- Recap ✓
- WCSS and Elbow Method ✓
- Initialization Trap
- Kmeans ++
- Limitations of K-means
- K- median
- K- mediods

Limitations of K-means

Initialization Trap



Ideal Scenario



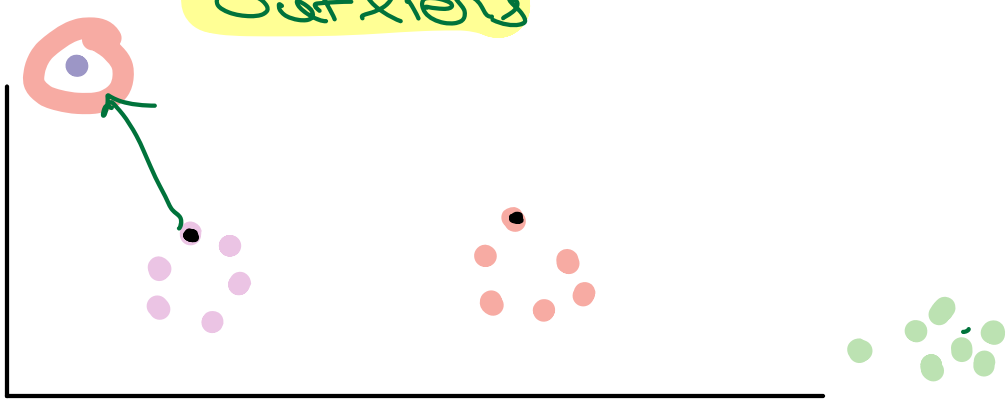
init – Trap

Kmean ++

Step -1 : Consider any one data point as first centroid randomly

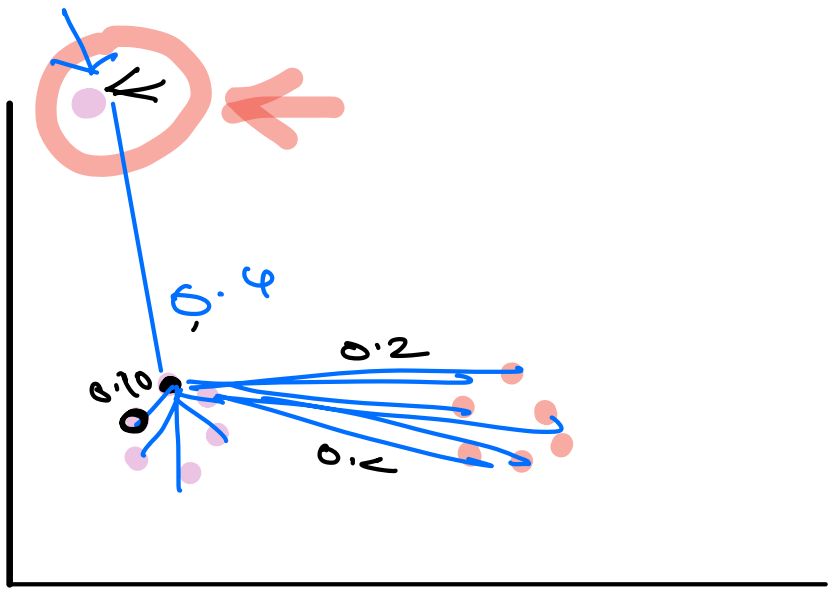
Step 2 : Pick the second centroid as far away as possible from first

Outliers



3) Handle outliers Beforehands

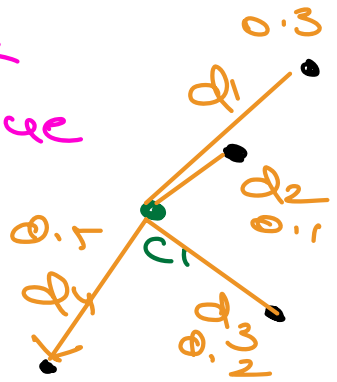
Pick centroids probabilitically



① Pick 1st Centroid Randomly

② Calculate $dist(C_1, x_i)$ and convert to probability

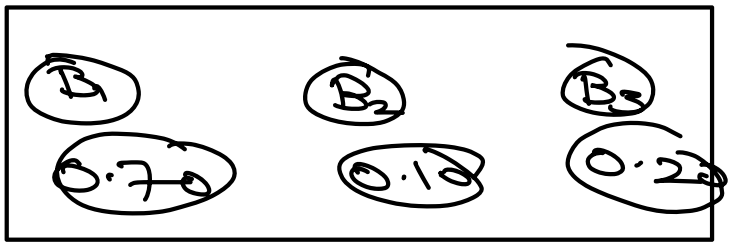
③ Based on P.D. pick the next Centroid Value



$P_1 \rightarrow \frac{d_1}{D}$ $P_2 \rightarrow \frac{d_2}{D}$ ----

$d_1 + d_2 + d_3 + d_4 = D$

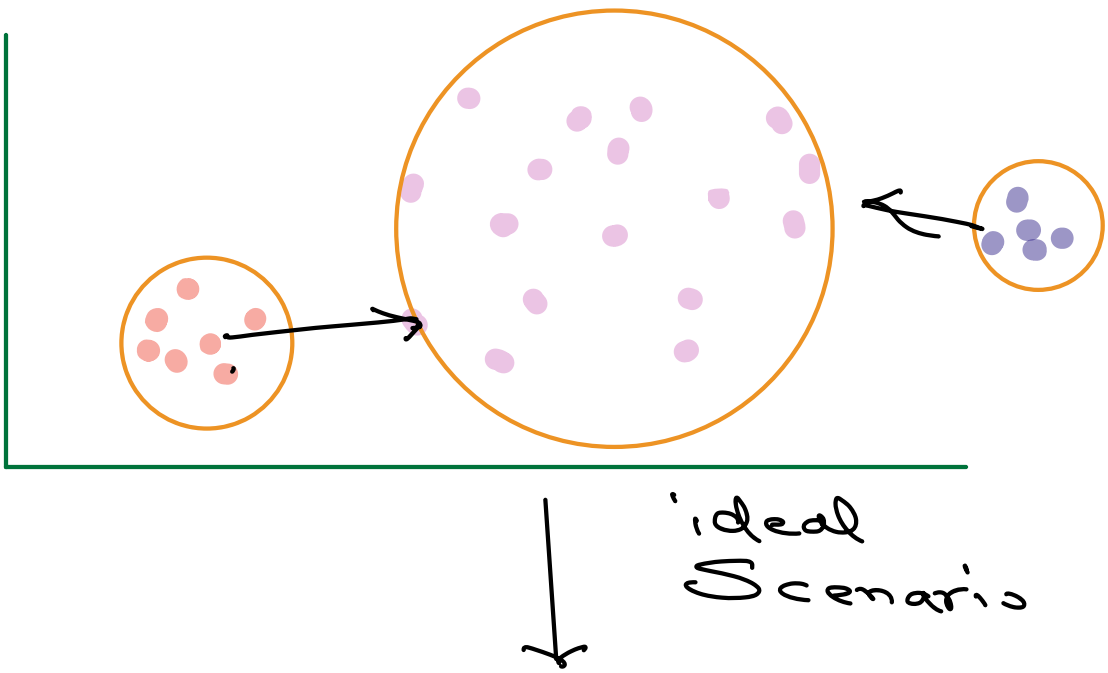
By 70%

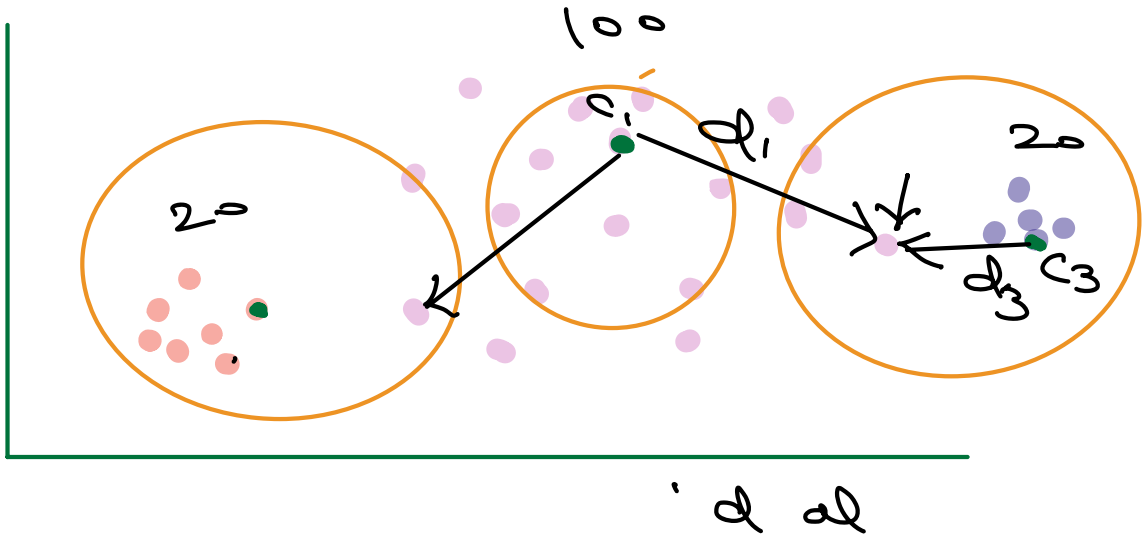


* The chances of a point getting picked will be proportional to its distance from previous Centroids

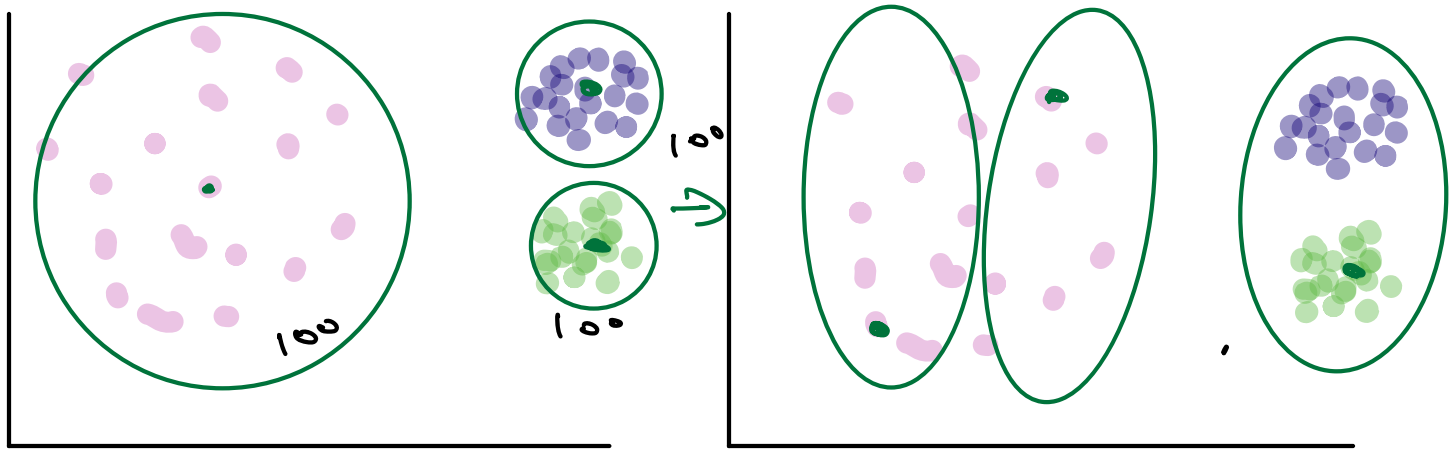
Limitations of kmeans / kmeans++

1. Different sized Clusters

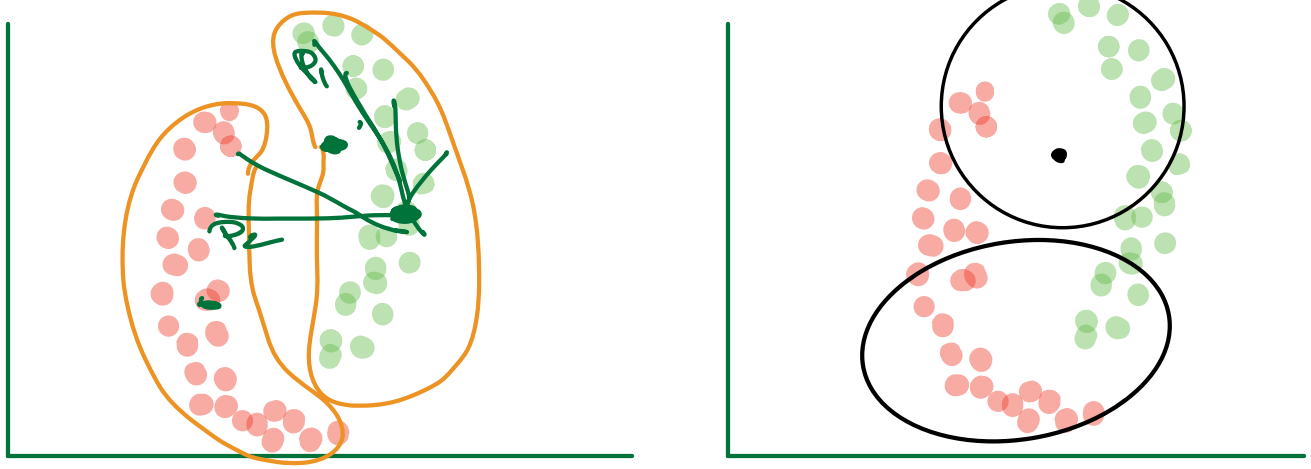




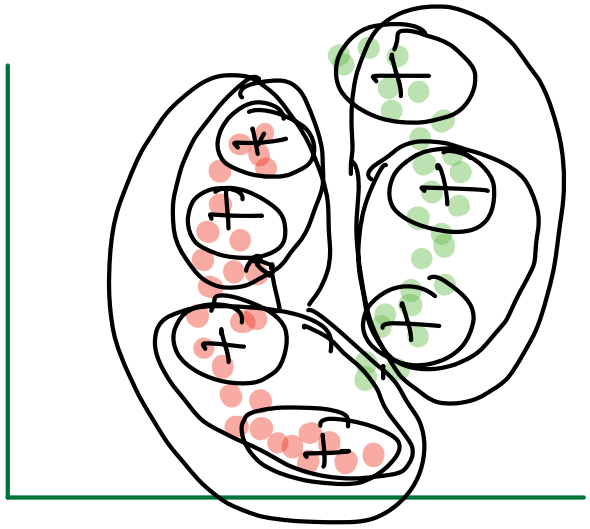
② Density of Clusters is Different



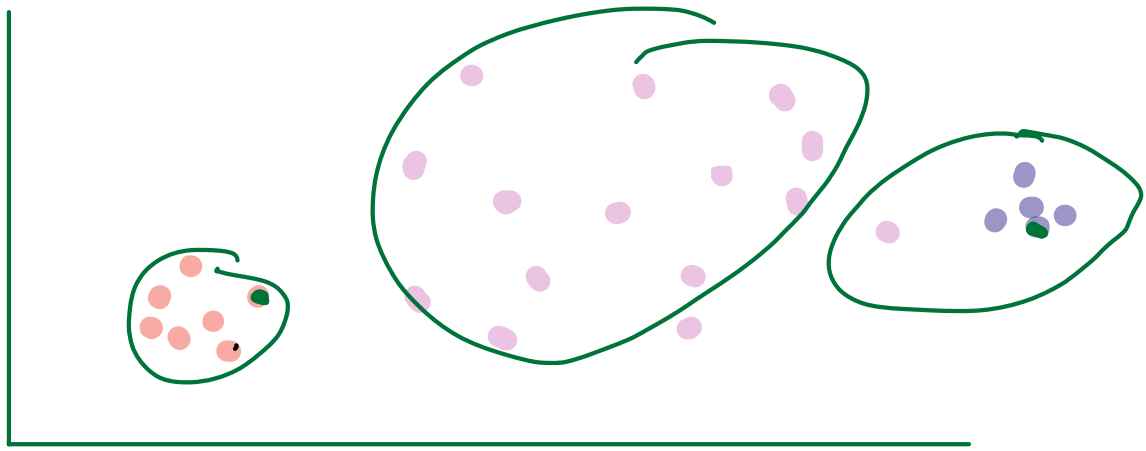
③ Clusters are Not Globular Shaped



Resolution



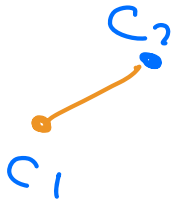
2



d d

K-median

1. Initialization
2. Assignment
3. Update Step



In K-means, we were taking mean of data points assigned points

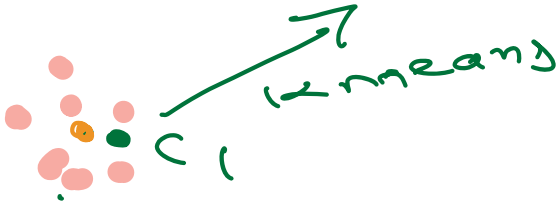
* In K-median, we use median of distance to update centroid.

Sorted Distances

d_1 d_2 d_3 d_4 - ---- d_N

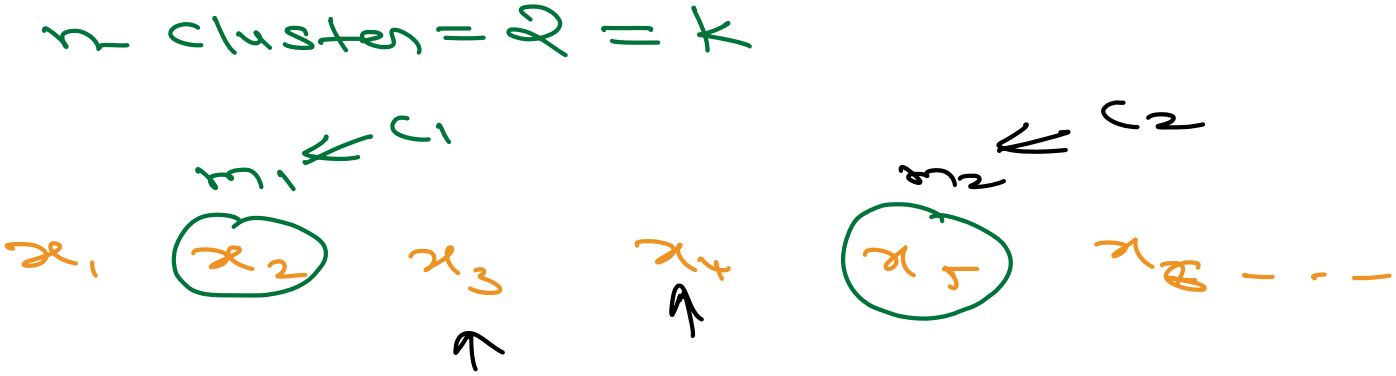
Which scenario can this help in?

Outliers



Q What if you want the centroid to be one of existing point always

K-medoids



①

$$\text{Loss} = \sum_{i=1}^{k} \sum_{x_i \in c_i} \text{distance}(x_i, \ldots$$

② Iterate over all Data Find the value of m_i such that loss is minimum