K-means

Assume K components

Each has a mean ("centroid") μ^k

Minimize:

$$\sum_{k=1}^{K} |C_k| \sum_{x \in C_k} d(x, \mu^k)^2$$

$$d(x, \mu^k)^2 = \left(\left[\sum_j (x_j - \mu_j^k)^2 \right]^{\frac{1}{2}} \right)^2$$

(Euclidean)

$$= \sum_j (x_j - \mu_j^k)^2$$

Suppose we were given $\{\mu^1 \dots \mu^K\}$. How would we assign $x_1 \dots x_N$?

$$C(x_i) \leftarrow \arg\min_z d(x_i, \mu^z)$$

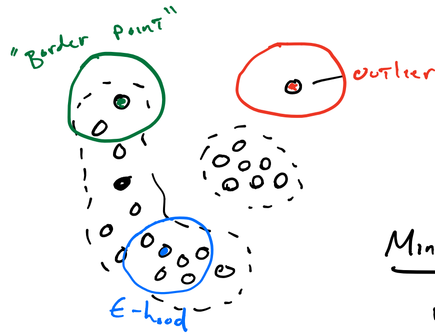
Conversely, if we are given cluster assignments C we can easily derive cluster means

$$\mu_z = \frac{1}{|C_z|} \sum_{x \in C_z} x$$

K-means just repeats these two steps until cluster assignments stop changing.

DBSCAN

Clusters $\stackrel{def}{=}$ "Densely" connected points



Clusters are defined by:

$\underline{\epsilon}$: Max radius of a neighborhood

$\underline{\text{Min Pts}}$: Minimum number of points in the ϵ-hood of a point.

Consider a point q

$$\epsilon\text{-hood}(q) \stackrel{def}{=} \{ p \in D \mid dist(p,q) \leq \epsilon \}$$