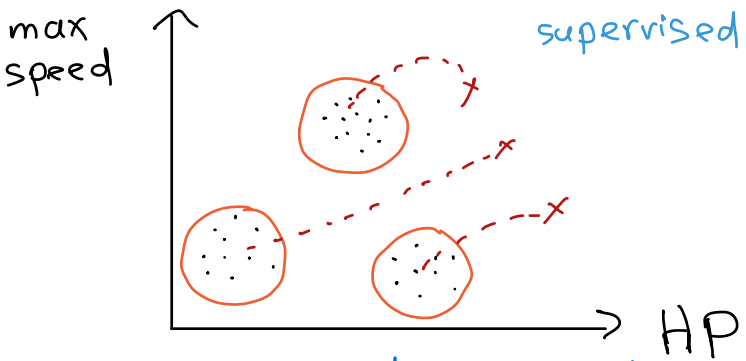
Clustering : Intelligence is the capability of grouping similar objects.

Clustering groups "unlabled" data into clusters of similar inputs.

Unsupervised <x>

vs

supervised <x, d>



Are clustered well seperated ?

Are clustered linearly seperable ?

difficulties :- overlaps, complicated shape

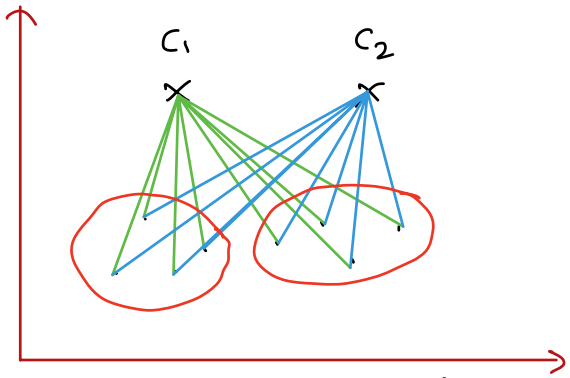
Clustering algorithm → need to know # of clusters
→ don't needed

K-means algorithm = find the centroid (prototypes) (means) of K clusters

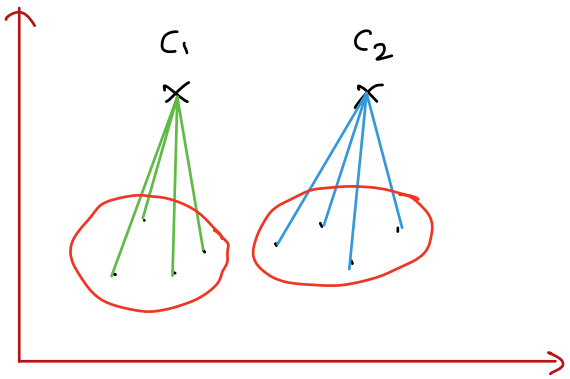
① Randomly placed K centroids

② Assign each data points to its closest cluster K.

③ Update the centroids

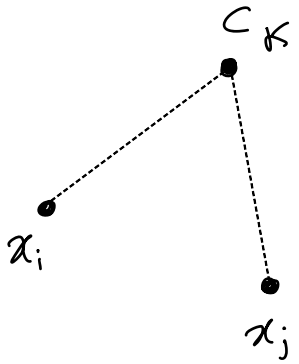


distances for class assignment



distance calculation for updates (error)

Similarity grouping happening via distance measurement.



$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \ldots}$$

$$= \| x_i - x_j \|_2$$

$i \longrightarrow$ measurement

$1, 2, 3, 4 \ldots \longrightarrow$ features

Objective (error)

$$E = \sum_{K=1}^{K} \sum_{x=C_K} \| x - m_K \|_K$$

m_K ↑ centroid of K th cluster

Minimize the sum of squared errors to its prototype in each cluster.

Update ⟶ centroids are the average of all $x \in c_k$ for $k \in \{1, 2, 3, \ldots K\}$

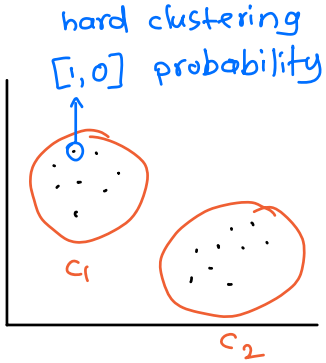
Stopping ⟶ ① After some iterations

② When centroids dont change anymore

③ When few/no data points change cluster

Problems of K means :

1. Needs K
2. Outlier sensitive
3. Hard clustering

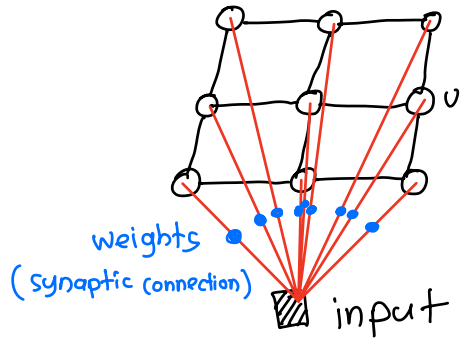


Clustering is unsupervised learning

Use processing units (neurons) to place centroids on an adjustable map: Self organizing maps

Hypothesis : The model self organizes based on learning rules and interactions.

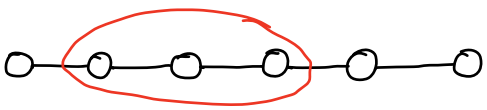
Processing units maintain proximity (geographical closeness) relationships as they grow.



The input is connected with each unit (neuron) of a lattice (map)

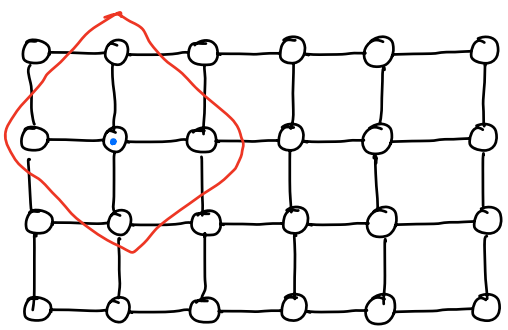
Concept of neighbourhood :

1D

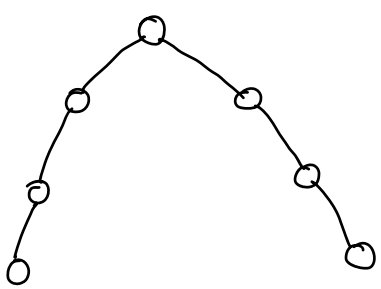


possible neighbourhoods

2D



Goal - find weight values s.t adjacent units have similar values.



neurons grow together
(piecewise continuity)

→ Inputs are assigned similar to units that are similar to them.

→ Each unit become center of a cluster.

SOM = constrained K means

→ Given input vector x find the ith unit with closest weight vector by competition.

$W_i^T x$ will be maximum.

→ For each unit j inthe neighbourhood N(i) of the winning neuron i, we update the weights of j (W_j)

→ Weights outside of N(i) are not updated.

→ SOM has 3 stages, ① Competion

② Collaboration

③ update weights

Competion)

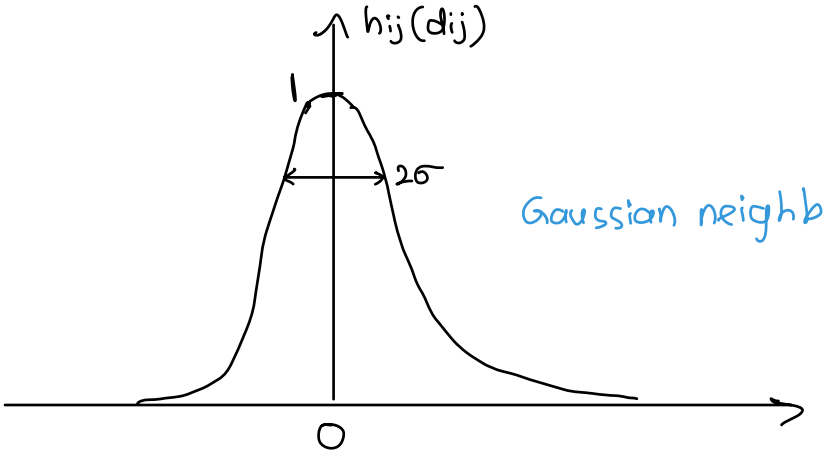
Find the most similar unit

$$i(x) = \arg\max_j \| x - w_j \|_2 \qquad j = 1, 2, 3 \ldots n \; ; \; m \neq \text{units}$$

Collaboration)

Use the latteral distance d_{ij} between the winner unit i and unit j

$$h_{i,j}(d_{ij}) = \exp\left(\frac{-d_{ij}^2}{2\sigma^2}\right)$$ Gaussian neighbourhood



Gaussian neighbourhood

$$\sigma(n) = \sigma_0 \exp\left(\frac{-n}{T}\right)$$

n = number of iterations

T = constant