DATE : __/__/__
PAGE : _____

* Unsupervised learning *

↳ Clustering

Have input & no output

$$D_n = \{ (x_i)_{i=1}^{n} \mid x_i \in \mathbb{R}^d \}$$

Problem: given some newspaper articles tell me the output label that is sports or non sports.

Sol:

Article	Category

sol: it will be classification

problem: given some newspaper articles group all similar articles togather

Solution: Clustering / grouping

Types of clustering algorithm: 1) K-means, 2) DBSCAN

Application of clustering: Ecommerce: task is to group similar customers based on their purchase behavior

Customer behavior include:-

- how much money they spend
- kind of credit card
- kind of customers product they buy
- geographical area

K-means algorithm:- assign K - clusters $\{C_1, C_2, C_3 \cdots C_K\}$

for $K_3 = \{ C_1, C_2, C_3 \}$

Set of points in: C_1 are S_1

C_2 are S_2

C_3 are S_3

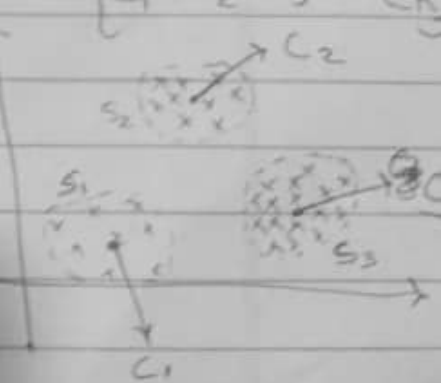
$S_1 \cup S_2 \cup S_3 = D_n$

$S_1 \cap S_2 \cap S_3 = \phi$

mean of $S_1 = C_1$

$S_2 = C_2$

$S_3 = C_3$



$$C_1 = \frac{\sum_{x \in S_1} x_j}{|| S_i ||}$$

DATE :__/__/__ ★
PAGE :_____

* Lloyd's Algorithm derives K-mean Clustering

Its steps are :-

1. initialization :- randomly pick 'K' different points from D_n an call them $C_1, C_2, C_3 \cdots C_K$
2. Assignment :- for each point x_i in D_n :-
   - Select the nearest centroid C_s
   - add x_i to set S_s
3. Recomputation :- recalculate C_s as follows

$$C_s = \frac{\sum_{x_i \in S_s} x_i}{||S_s||}$$

4. repeat step 2 & 3 until convergence (centroid ~~should~~ shouldn't change much).

from sklearn.cluster import KMeans