U-5
(ML)

* clustering

cluster → a no. of similar things that occur together.

→ technique in which data points are arranged in similar groups dynamically without any pre-assign^n of groups.

App^n → • Image Processing • Insurance
• Recommendation Engine

* properties of clustering algo.

• scalability
• ability to deal with diff data types
• Minimal to requirements for domain knowledge for determine input parameters
• Interpretability & usability

* Typical Requirements → clustering in data mining

↳ • Scalability.

All these + • Discovery of clusters with arbitrary shape.

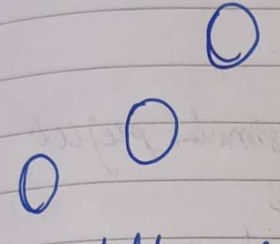
* Problems with clustering

↳ • not address all requir^n
↳ • dealing with large no. of dimensions & data items → problematic
↳ • effectiveness depends on distance
↳ • distance measure shouldn't define it.
↳ • result of clustering → interpreted in diff ways.

* Types of Clustering

a) well seperated clusters

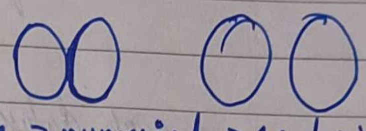
• a cluster of set of points such than any point in a clusters in closes to every other point in cluster



• threashold → used to specify all objects in clust or sufficiently close

b) prototype based clusters.

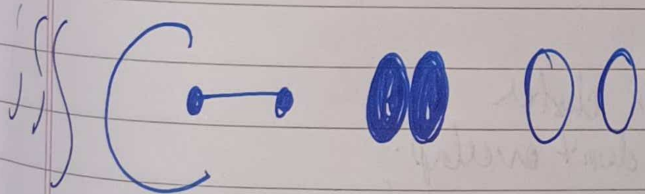
• set of object → is closer to prototype or central of cluster.



• data → numerical → centroid.
• data → categorical → prototype
• kmeans & kmedoids are example.
•

c) contiguity based clusters.

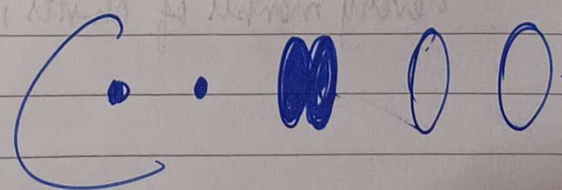
point in cluster is close to one or more point in cluster than any other



d) Density based.

• cluster → dense region of points seperated by low-density region from other regions of hight density

• Used when cluster are irregular or intertwined & when noise & outlier are present



K-Means -

- heuristic method.
- supervised learnig algo.
- solve clustering problems.
- group are unlabelled dataset in diff clusters.
- k → defines no. of predefined clusters needs to be created.
  k=2 ⇒ 2 clusters

- each dataset belongs only one group has similar project
- mimizes sum b/w data points & clusters
- distance calculation → Euclidean distance

Adv → efficient in compution.
→ easy to implement.

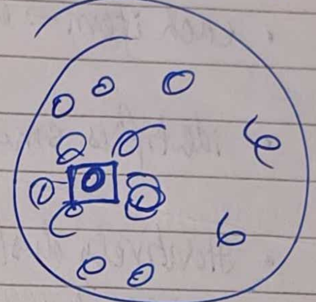
Disadv → only when mean is defined
→ need to specify k, no. of clusters, in advance
→ Trouble with noisy data & outliers
→ not suitable to discover cluster with ~~non~~ non complex shapes

Properties —

- always k clusters
- always atleast one item in cluster
- cluster are non - hierarchical & don't overlap.
- every member of cluster is closes to clusters.

k-medoids -

- each cluster represented by one of objects in cluster.
- Data points are chosen by medoids.
- classic partition clustering technique that groups data set of n objects into k groups
- known as priori
- less delicate to outlier
- convex shape not required.
- More robust to noises.



Hierarchical Clustering

- method of cluster analysis in which data points are arranged in hierarchy of clusters.

Adv →
- simple to implement
- easy & results in hierarchy (more info)
- doesn't need prespecify no. of clusters.

Disadv →
- large clusters breaks
- diff to handle diff sized clusters & shapes
- sensitive to noise & outliers
- can't be changed or deleted once done.

Hierachical Clustering

Types

Agglomerative Clustering	Divisive Clustering
• bottom up approach	• town down approach
• each item → own cluster	• all item → one cluster.
• identifies small cluster.	• identifies large cluster.
• Iteratively clustered all merged together	• Large clusters are successively divided.
• also know AGNES (Agglomerative Nesting)	• also known DIANA (divise analysis)

★ Dendogram

- Diagram representing tree of hierachy.
- The similar two objects are less is the height of link that joins them
- major info lost here

* DBSCAN

- Density Based spatial clustering of Application with Noise.
- groups together closely packed points.
- unsupervised learning.

ϵ → radius of neighborhood around point x.
minpts → required to form a density cluster

Adv → • don't need to specify no. of clusters
- flexible in shape & size of clusters
- Able to deal with noise & outliers
- Ability to identify uneven shape.
- easy → someone who knows dataset → to set parameters.

Disadv → • Input parameters → difficult to determine
- some situation very sensitive to input parameter.
- confused → border point belong to two cluster
- result → distance metric
- hard to guess correct parameters

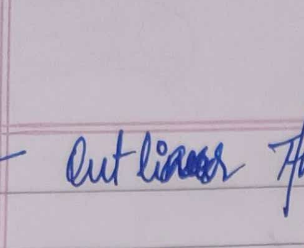
* Spectral clustering → • solves by creating clusters with arbitrary & non-linear shapes, no assumption → shape of clusters.

Adv → • no assumption → statistics of cluster
- Easy to implement
- Good clustering result.
- fast to sparse data set of several thousand elements

Disadv → • May be sensitive to choice of parameters
- computationally expensive for large dataset

* Outlier Analysis -

• statistical observation i.e. marked differently in value from others in sample.



→ Types -

1) Global
- a data obj is called global outliers.
- if it deviates from rest of data set

2) Contextual (Conditional)
- deviates significantly on context of object
- only with context → outlier
- generalization of local outliers

3) Collective Outliers -
- obj as a whole deviates significantly from entire dataset.
  it is collective.

• Challenges in Outlier detection

→ modeling normal objects & outliers
→ appln specific outlier detection
→ handling noise in " "
→ Understandability

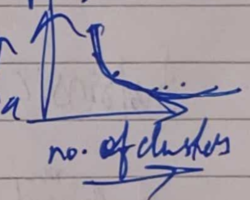
Local Outlier Factor (LOF)

- concept of local density
- locality → k nearest neighbors
- distance b/w k's used to determine density
- points that lower local density ⇒ their neighbors are outliers
- unsupervised anomaly detection - method

Elbow method -

- used to determine no. of clusters
- Inertia is sum of all distance of data point from centroid of clusters.

algo →
- Start with any value of k & perform k-mean algo.
- Determine total inertia
- Increase k by 1 & carry out step 1 & 2 until inertia is not significant.



- Measuring Clustering Quality

- Extrinsic → Ground truth available
  → reward behaviour

- cluster homogenity → purer cluster → better clustering
- cluster completeness → If 2 two obj → same category → same cluster
- Rag Bag → obj can't be merged with other obj.
- Small cluster preservation → splitting small category into smaller is more harmful than large → small.

Intrinsic Method –

- Ground truth not available.
- evaluate how well clusters separated → compactness
- Silhoutee coefficient → defines goodness of clustering technique.

values → -1 to 1

1 → cluster well part & distinguished
0 → cluster indifferent
-1 → clusters are assigned in wrong way.

Silhouette Score → $\frac{b-a}{\max(a,b)}$