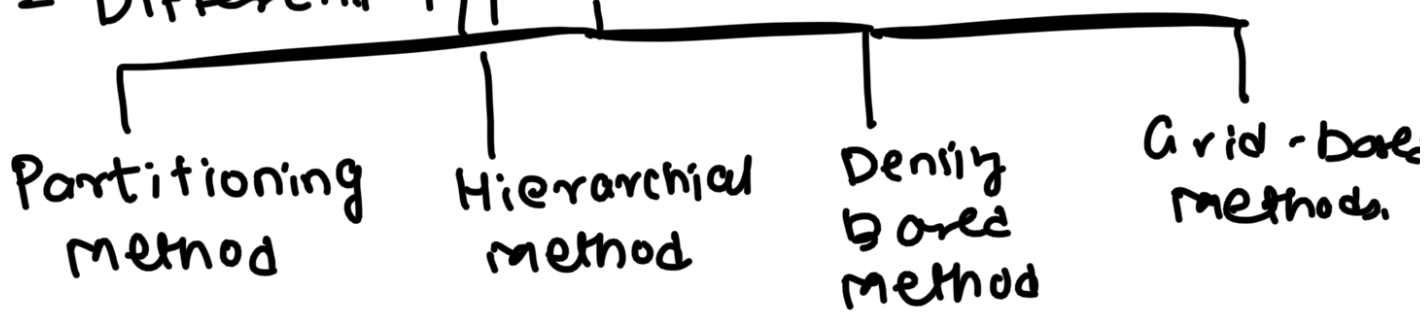
Unit 5

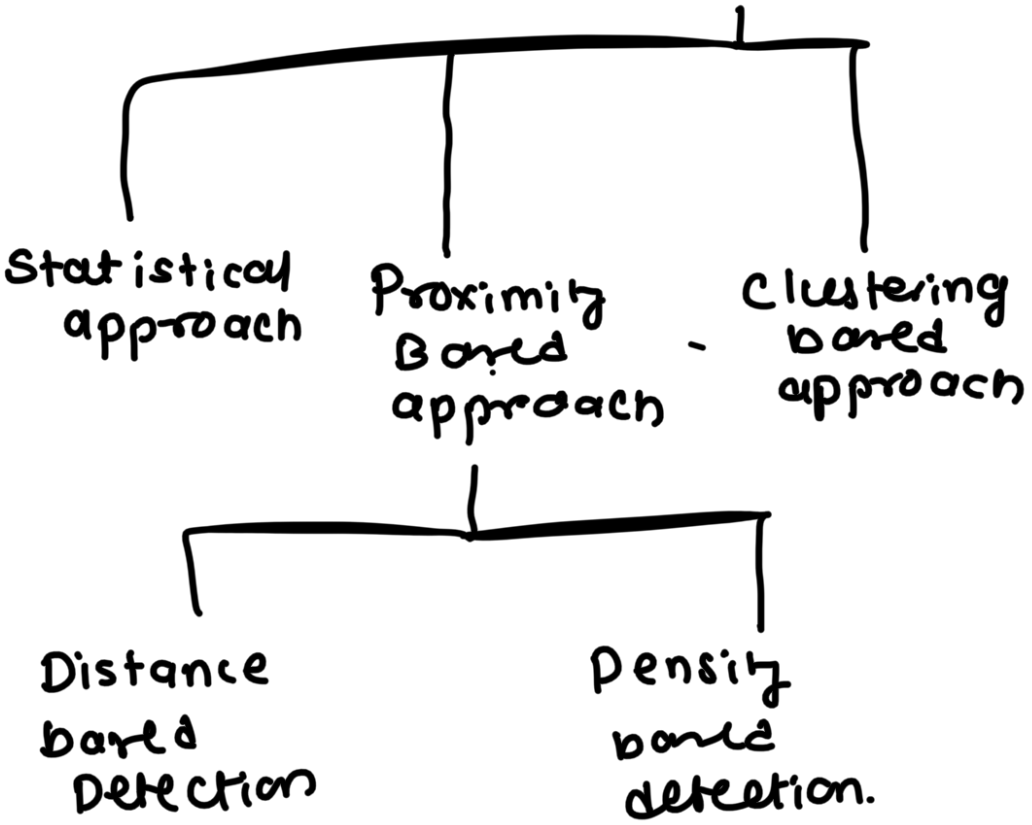
- Introduction to cluster analysis.

- Different types of clustering methods:



- Partitioning Methods
  - Global optimal
    - k-means
    - k-medoid (PAM)
  - Heuristic
- Clustering High Dim. Data
  - Need
  - Types
    - Subspace clustering
    - Dimension ality redu ction
- Clustering with constraints.
  - Constraints on instances
  - Constraints on clusters
  - Constraint on similarity measure ment.
- Outlier Analysis
- Applications
- Types of outliers
  - Global outlie
  - Contextual
  - Collective outlier
- Need/challenges of Outlie Detection.

- Outliers Detection Methods



- Outlier Detection in High Dimensional Data.

Unit 5

- Clustering :
  - unsupervised learning
  - process of partitioning set of data objects into subsets.
  - Cluster has similar data.
  - Set of clusters resulting from cluster analysis is called clustering.

- Importance :
  - widely used in many application like business intelligence, image pattern recognition, web search, biology.
  - Used in outliers detection.
  - Business intelligence → identify customer, group them
  - web search → organize the results into groups.
  - helps segregate population

- Types :

- Types:

1. Partitioning methods — 1. Distance-based.
   2. Use mean or medoid to represent cluster center
   3. Effective for small and medium sized data.

2. Hierarchical Methods — 1. Decomposed hierarchically
   2. Incorporate other tech. like micro clustering.

3. Density-based methods — 1. Clusters formed are dense regions
   2. Can find arbitrarily shaped cluster

4. Grid-based methods — 1. Use multidimensional grid
   2. Fast processing Time

- **Partitioning Method:** Partition datasets into k clusters.

- **Clustering High Dimensional Data:**

  more than 10 attribute → High Dimensional data

  Need: Example of supermarket → customer purchase data is high dimensionality → Traditional distance methods can't be used (Euclidean or manhatten distance).

1. **Subspace clustering approach:**
   - Subspace — subset of attributes in full space.
   - ... existing subspaces of given high

- searching cluster in [illegible]
  dimensional data.

2. Dimensionality reduction approach
   - Tries to construct lower dimensional space and search for cluster

* Clustering with Constraints:

1. Constraints on instances:
   - specify how pair of instances should be grouped.
   - must link constraint
   - cannot link constraint

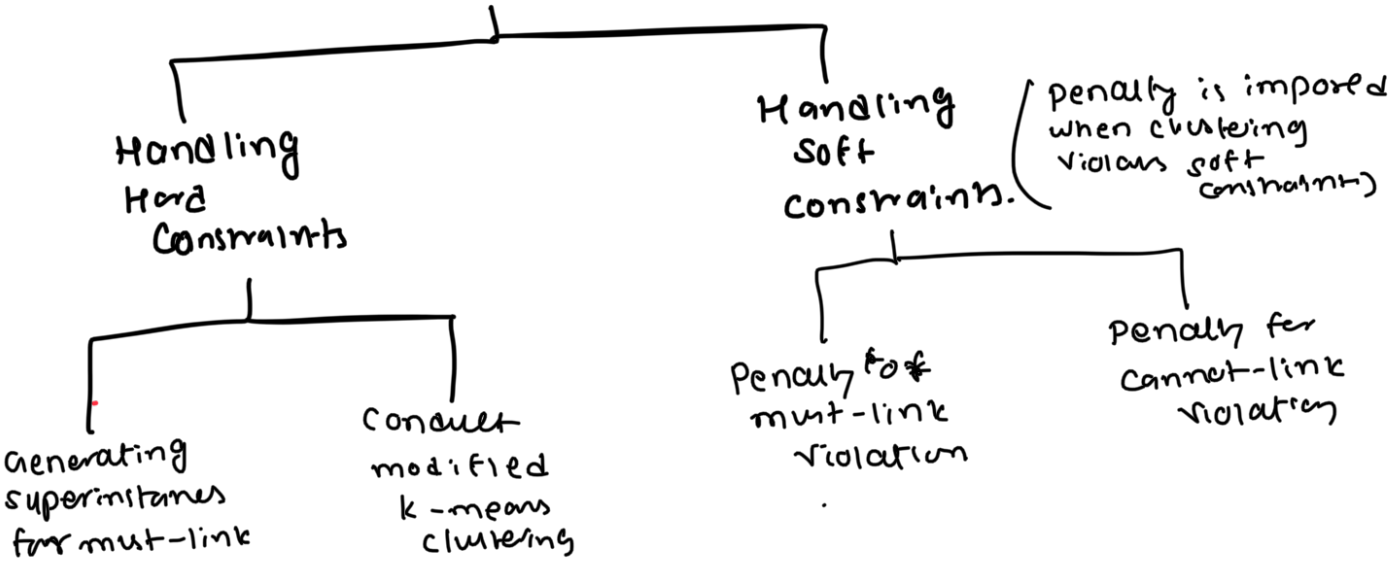
must link — - constraint on a and b, then a and b in same cluster
- Transitive in nature
  must link(a,b) and must-link(b,c) then must-link(a,c).

cannot link — If cannot-link (a,b) then a and b have diff. clusters.

2. Constraints on clusters: specifies requirement on cluster, using attributes

3. Constraints on similarity measure: specifies similarity measure
   e.g. Euclidean distance must be respected.

Methods for Clustering with Constraints:



1. Generating super instances for must-link:

1. Transitive closure is computed on must-link constraints. [illegible] subsets of objects, all objects

2. This closure gives multiple ... in subset must be assigned in one cluster.
3. Replace all these objects by mean
4. Weight of superinstance, is no. of objects in represen
5. Must-link constraint is satisfied.

2. Conduct modified k-means clustering:

- we modify the center assignment process of k-means to nearest feasible center assignment.
- when objects are assigned to center in sequence, make sure the assignments do not violate cannot-link.
- Therefore, respects all cannot link.

3. Penalty of must-link violation: The must-link constraint(a,b) are present in diff cluster then the dist(a,b) is added to objective fun. as penalty.

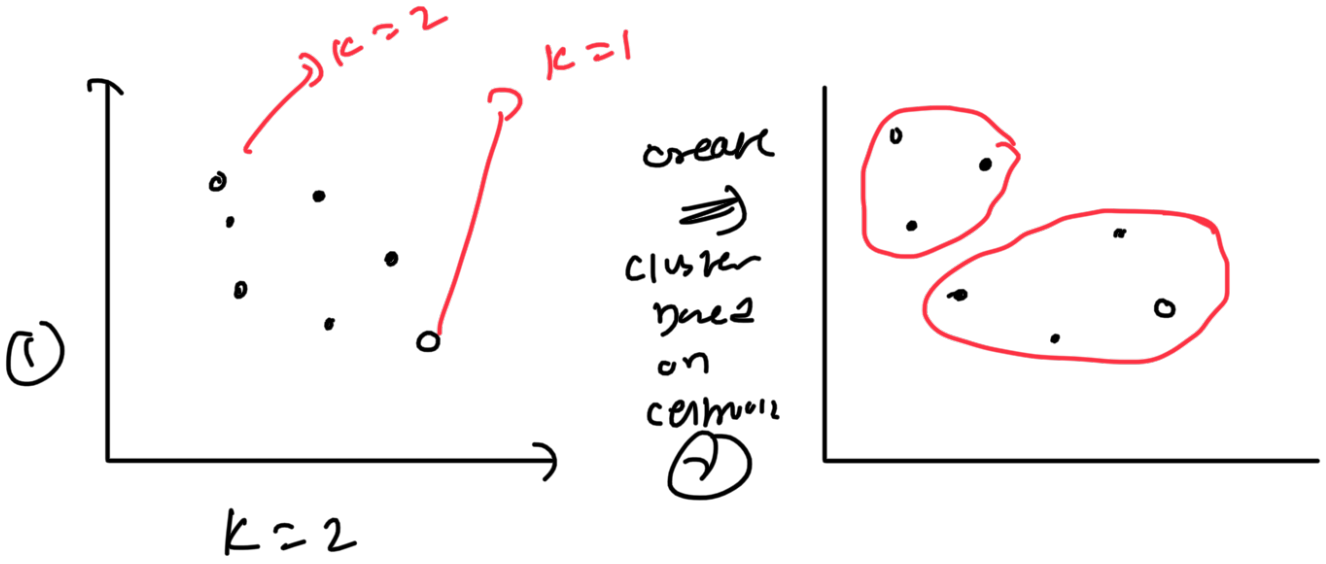
4. Penalty of cannot-link violation: The cannot-link constraint(a,b) are present in same cluster then dist(a,b) is added to objective func. as penalty.

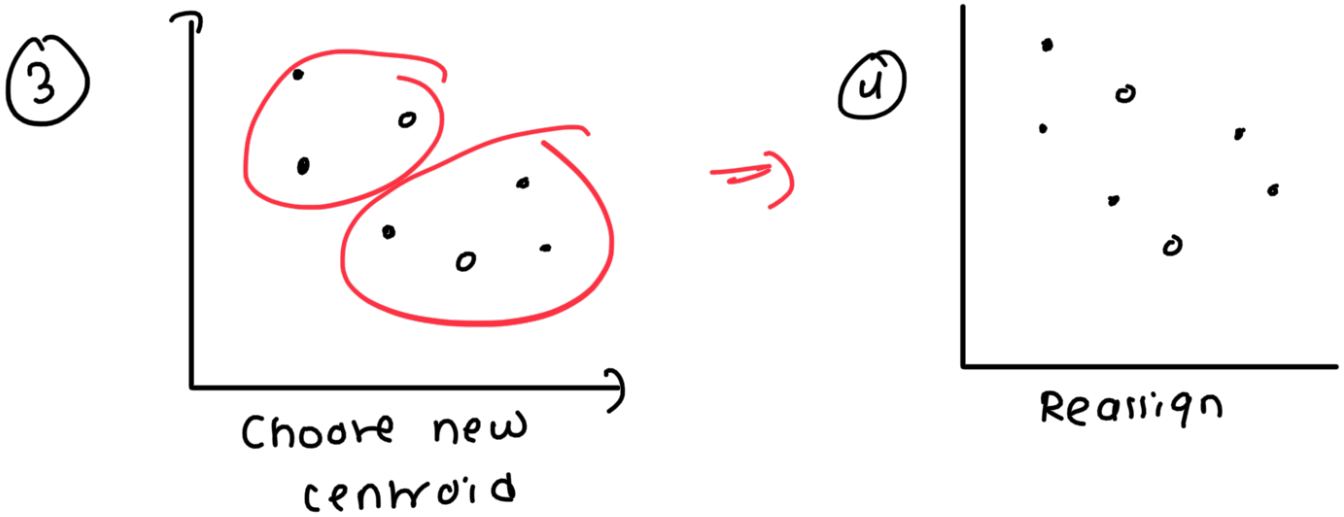
* K-mean:

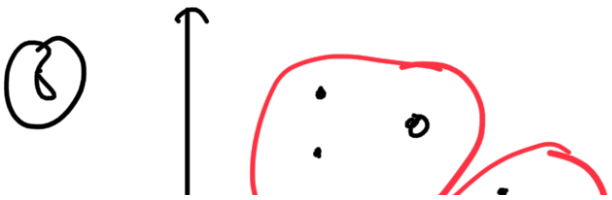
- unsupervised learning algo.
- In k-means, the data object are classified based on their attributes into k number of cluster.
- Algo:
  1. Define k centroids for k clusters which are generally fare away.
  2. Group elements into clusters, which are nearer to centroid of cluster
  3. Now calculate new centroid for each

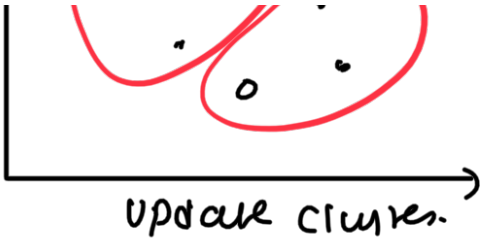
cluster.

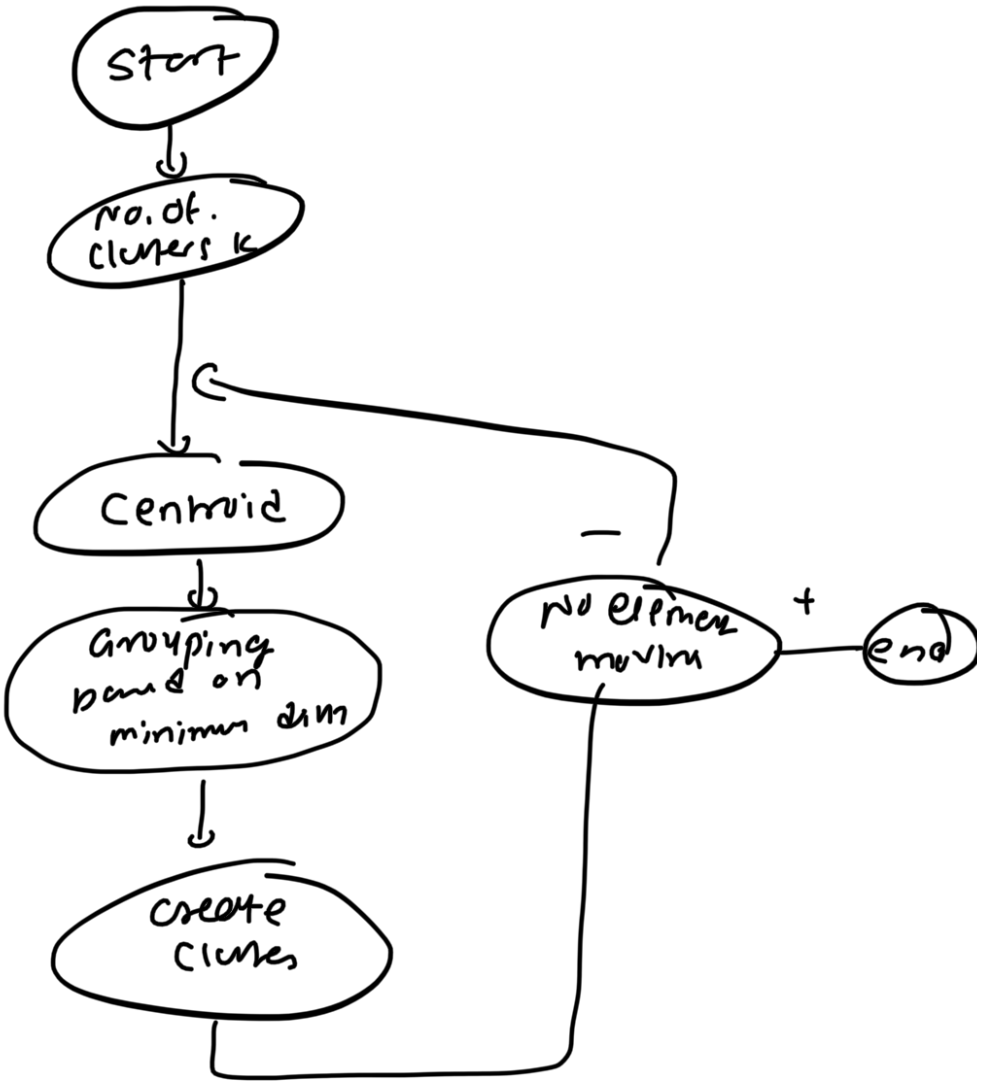
4. Again group elements based on new centroid.
5. In every step, the centroid changes the elements more from one to another cluster.
6. Do same process till no element is moving from one to another.











K-medoids:

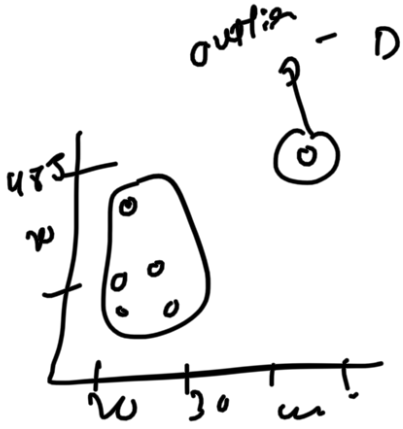
- Take most centrally located object in a cluster.
- PAM. (
- Handles outlier well.
- Ordering of input does not impact results.

Each cluster ⇒ medoids.

Algo:
- select k points as medoids.
- Assign all points to closed medoids
- see for better medoids.
- Repeat until no medoids change.
- medoids are changed if overall cost is important.

Advantage:
1. effective on small data sets.
2. PAM robust than k-mean

* Outlier Analysis:
- observation or subset of observation diff from remaining set of data
- unusually large or small value.
- rise due to instrumental error, Human error.



- Detection is imp:
1. Human error outlier needs to be deleted or corrected.
2. outlier does not mean bad data, statistical analysis can be used to analyze.

- Applications:
1. Fraud detection.

2. Loan processing
3. Intrusion detection
4. Unexpected entries in database

- Types:

1. Global outlier : - simplest outlier to detect
   - when data obj diff from rest of given data.

2. Contextual outlie :-
   - Time series data.
   - data obj anomalous within its context.
   - winter season temp rises to 35°.

3. Collective outlier :-
   - A subset of data obj outlying to entire data set.
   - individual data obj is not outlier, but their collection

:- Human electrocardiogram

- Detection methods.

1) supervised method :
   - Training data → normal class
                   → outlier class
   - Data compared and find out which class it belongs to
   - limitations is accurately labelled training data

2) Semi-Supervised method :

2) Semi-supervised -

- only one type of class
- If data does not fit the one model, assumed to belong to another class.

3) Unsupervised - method.

- uses notion of outliers and then makes use of same to detect outliers.

- Three methods
  - Statistical Based
  - Proximity Based
  - Clustering Based

1. Statistical Based.

Z-score

↓

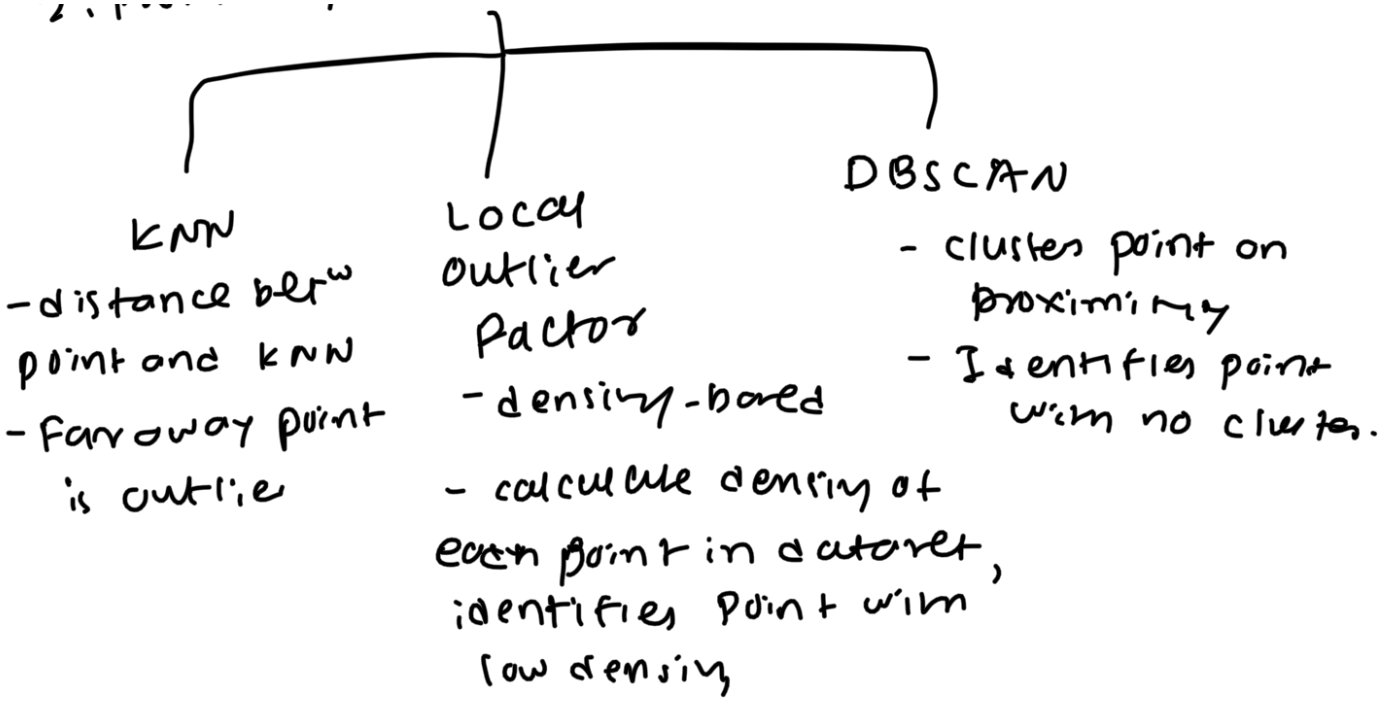
calculate
- Z-score of each point
- Z-score: no. of standard deviations that a point is from mean

Interquartile range

↓

- divide data into quantiles
- identifies points that are more than a certain number of IQR,

- Proximity Based:



3. Clustering Based

* Outlier Detection in High Dimensional Data:

- Dimensionality increase, noise increases.
- Conventional methods cannot be used.

1) Data sparsity. — high-D often sparse
nearly dominated by noise.

2) Data subspace — - Adaptive to subspaces signifying the data
- Capturing local behaviour of data