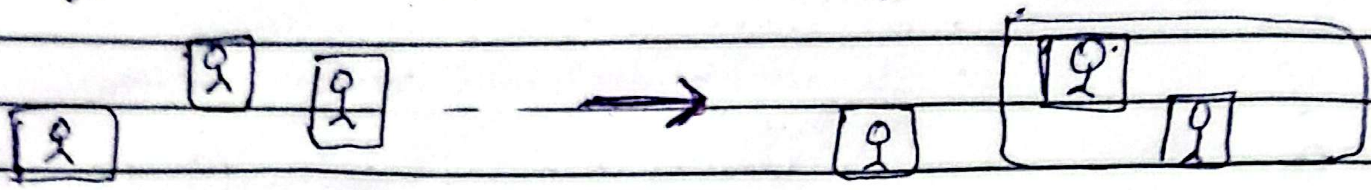
Hierarchical Clustering (Visualized as dendrogram)

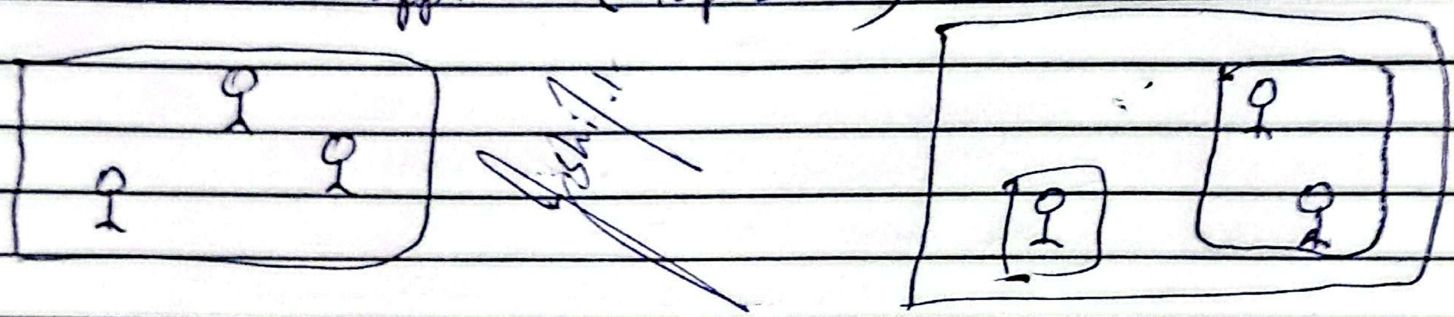
• — Approaches:

- Agglomerative approach (bottom-up)



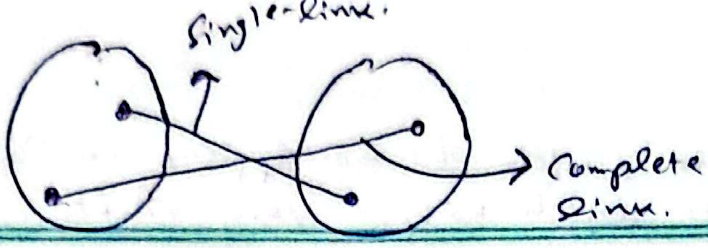
Phly sb ka 1 banao, phir see agg.

- Divisive approach (top-down)



Start with large cluster then 1 1

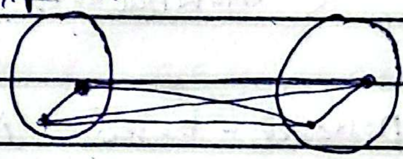
• — Agglomerative → • Make cluster for every data point
- Check similarity b/w every cluster. — (2)
  If m clusters, mXm matrix.
- Clusters having more similarity, merge them
- If left with more clusters start again from (2).
- Ways for calculating similarities.
  - Single-link Clustering (MIN)
  - Complete-link Clustering (MAX)
  - Group Average
  - Distance b/w centroids.



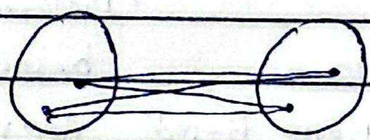
- Single link Similarity (MIN)
  → Represents similarity of most close or similar members by measuring Euclidian distance.
  → Problem is that it produces long chains.

- Complete-linkage Similarity (MAX)
  → Similarity of their most dissimilar members.
  → Problem is that it is sensitive to outliers.

- Group Average Clustering
  → Problem that computation is expensive.



- Centroid Clustering
  → Problem is Jb hum clusters ko merge krtay hain tou similarity dono clusters kay sath beh improve ho skti hay. Is wajah sey dendogram main overlaps bantay hain which makes analysis difficult.



- Divisive Clustering

- points which do not appear in Clusters
- points which behave very differently from normal.

• - Outlier Detection

• Automatic outlier detection.

Methods → • Statistical Approach
• Distance-Based Approach
• Deviation-Based Approach

a) Statistical Approach → • We assume that data has a model. e.g., normal distribution with 3 Sigma Rule.
• Jo ±3 std sy bhr woo outliers.

3 sigma rule (Data key zaida tar points mean key ass pass (±3 std) key andr.)

Drawbacks → • Tested mostly for 1 attribute
• Data distribution is unknown.

b) Distance - Based Approach → • Need multi-dimensional analysis without knowing data distribution.
• Different Algo for mining e.g., index-based, nested-loop, cell-based algo.

c) Deviation - Based Approach → • Identify outliers by examining main characteristics of objects in a group.
• Sequential exception technique → like human can distinguish.
• OLAP data cube technique → • uses data cubes to identify region of anomalies in large multidimensional data.

Learning
- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Association
- Semi-Supervised
- Meta learning
- Deep learning
  - Neural networks (CNN, RNN, LSTM)

• — Semi-Supervised Learning (SSL) Garbage In, Garbage out.

• When training data ka zyada data unlabeled ho aur thora sa labeled ho then we use SSL.

• These Algo learns from both ~~diff~~ labeled & unlabeled data.

• Labeled data ke labels main high certainty (confidence or yakeen) hoo.

• Self learning (use unlabeled data to improve itself)

• Procedure → • Train with labeled data

• Use this model and make predictions for unlabeled data.

L: Labeled

U: Unlabeled.

$y' = f(x)$ where $x \in U$

• Add these predicted labels in L

$L = L \cup (x, y')$

• Repeat.

• Problem → • Performance may degrade due to noisy instances.

• — Multi-View Learning (Ek observation ko 2 different independent features set say set kiya jata hy)

e.g A webpage can be described by its content or links which point to that page.

• Learning process zyda accurate hta hy.

• Takes advantage of every view so redundant (baar baar) view ko exploit (full advantage taken) kr kay performance ↑.

• 2 classifier work together to enlarge the training set L & increase performance. + called Co-training.

• Might cause overfitting.

- Imbalance Data. (Don't trust accuracy blindly).

- no Cancer = 998/1000

  Cancer = 2/1000

Model will be biased to no cancer so accuracy will be 99.8% - Dataset imbalance

Solution → • Collect more data
- Delete data from majority class
- Create synthetic data (artificially data points bencano to ~~enhance rare class~~).
- Adapt your learning algo.

- Random Oversampling → Minority class ke data points ko randomly duplicate kine so its quantity increases

  ⇓

  Overfitting + fixed boundaries

Random Under Sampling → Randomly delete data points from majority class.

⇓

loss of information

supported by weka.

SMOT (Synthetic Minority Over Sampling technique)
- In order to minimize risk of overfitting, we donot replicate minority instances but create new minority instances
- Operates in feature space.

Steps:-

1. Take difference b/w sample point & one of its nearest neighbours.
2. Multiply difference by random number b/w 0 & 1 and add it to feature vector.

- SMOTE generally works better than over sampling and under sampling.

- SMOTE filters has 3 parameters that needed to be specified.
    1) Class value of minority class
    2) no. of nearest neighbours
    3) %age of new minority instances to be created.

- Higher nearest neighbours, diversity ↑, generalization power ↑

- %age of instances to be created depends upon degree of class imbalance.
  Higher imbalance, %age ↑

- Best values for both step should be obtained through experimentation.