When to use cluster analysis

* Data Summarization
  + Data reduction
  + Cluster centers, shapes, and statistic
* Customer segmentation
  + Collaborative filtering
* Social Network analysis
  + Find similar groups of friends
* Precursor to other analysis
  + Use as a preprocessing step for classification and outlier detection
  + Use it for sampling and data reduction

Attribute selection

* Measure attribute “worthiness”
  + Use entropy
* Entropy
  + Measure lack of order or predictability
  + In statistics and information theory
    - Has a value of 1 for uniform distribution (predictable)
    - Knowing the value has a lot of information (high surprise)
    - Has a value of 0 for a constant signal (fully predictable)
    - Knowing the value has zero information (low surprise)
  + Algorithm
    - Start with all attribute and compute distance entropy
    - Greedily eliminate attributes that reduce the entropy the most
    - Stop when entropy no longer reduces or even increase

Hierarchical Clustering

* Two options for building the dendrogram on the left
  + Top down (divisive)
  + Bottom up (agglomerative)

Merge Criteria

* Single (best-case) linkage
  + Distance = minimum distance between all mi mj pairs of objects
  + Joins the closest pair
* Complete (worst-case) linkage
  + Distance = maximum distance between all mi mj pairs of objects
  + Joins the pair furthest apart
* Group-average linkage
  + Distance = average distance between all object pairs in the groups
* Other methods
  + Closest centroid, variance-minimization, ward’s method

Comparison

* Centroid-based methods tends to merge large cluster
* Single linkage method can merge chains of closely related points to discover cluster of arbitrary shape
  + But also merge two unrelated clusters, when the chaining is caused by noisy points between two clusters
* Complete linkage method tends to create spherical clusters with similar diameter
  + Will break up the larger odd-shape clusters into smaller spheres
  + Give too much importance to data points at the noisy fringes of a cluster
* The group average, variance, and Ward’s methods are more robust to noise due to use of multiple linkages in the distance computation
* Hierarchical methods are sensitive to a small number of mistakes made during the merging process
  + Can be due to noise
  + No way to undo these mistakes

DBSCAN

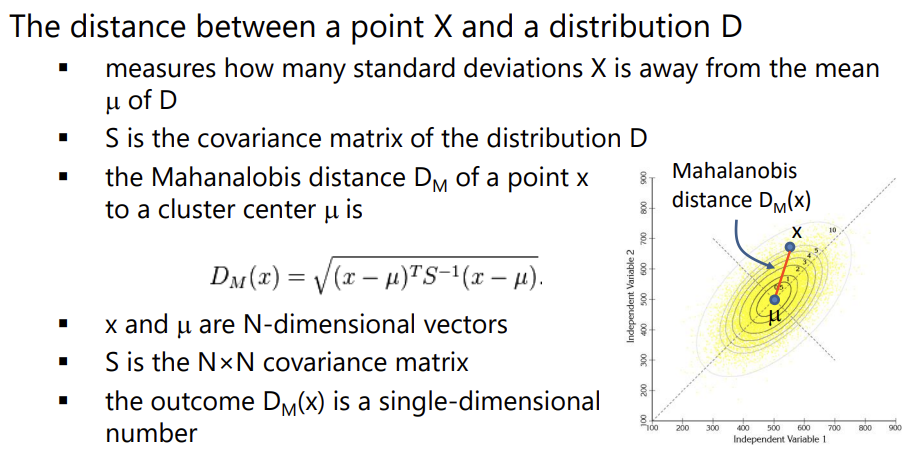
Text

Description automatically generated

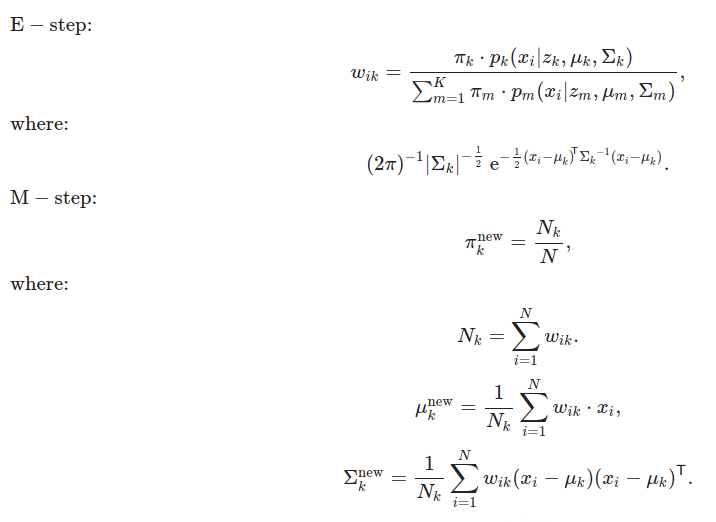
Diagram, schematic

Description automatically generated

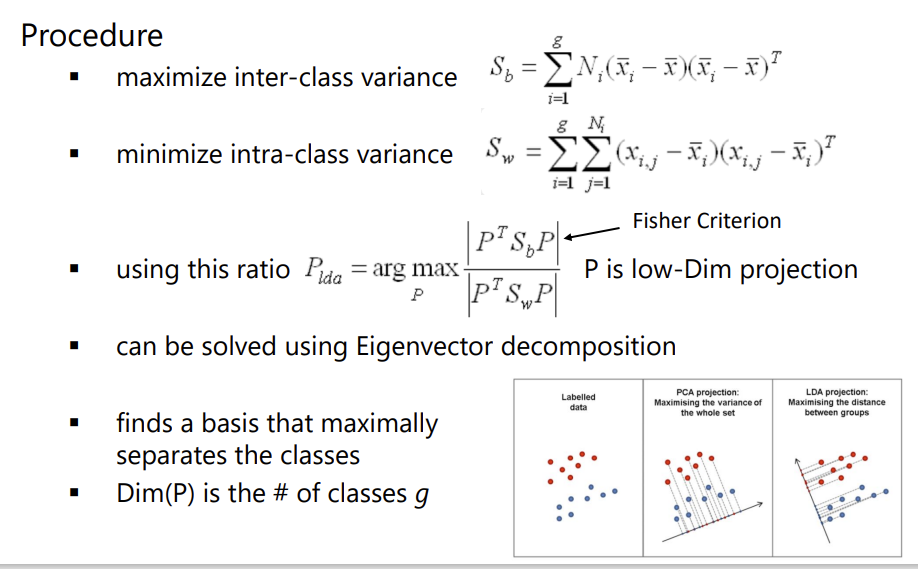
Mahala Nobis Distance



Probabilistic Clustering

* Expectation-Maximization
* 

Linear Discriminate Analysis (LDA)

* Require class labels
  + Having class labels enables better segmentation
* 

T-SNE

* T-distributed stochastic neighbor embedding
* Text, letter

  Description automatically generated
* Measures how (relatively) close Xj is from Xi, considering a Gaussian distribution around Xi with a given variance
  + This variance is different for every point
  + T is chosen such that points in dense areas are given a smaller variance than points in sparse areas
* Text, letter

  Description automatically generated
* Disadvantages
  + Does not preserve global data structure
    - Only within cluster distances are meaningful
    - Between cluster similarities are not guaranteed

Reduction VIA Neural Network

* Train a Variational Autoencoder (VAE)
  + Optimize the output reconstruction loss of the input
  + Optimize the latent distribution to be standard normal
* Advantages of interpolation in Latent space
  + Latent space allows easy interpolation
  + Move between samples in latent space and reconstruct novel instances by the decoder
  + Not easily possible using other non-linear layouts like MDS, T-SNE