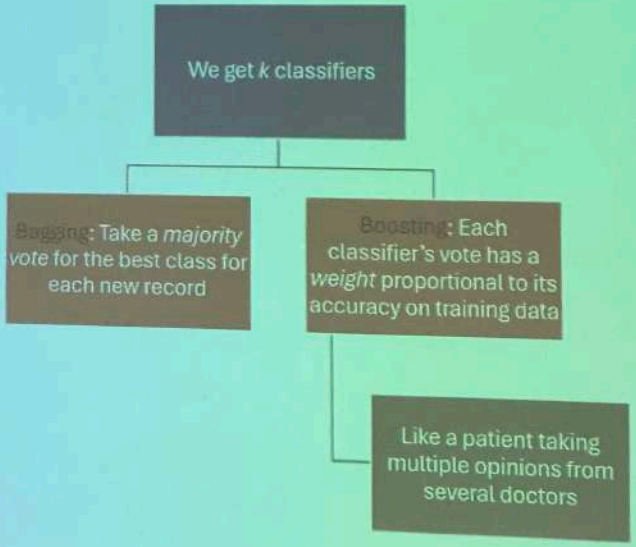
Combining Classifiers



Applications

- Targetting similar people or objects
  - Student tutorial groups
  - Hobby groups
  - Health support groups
  - Customer groups for marketing
  - Organizing e-mail
- Spatial clustering
  - Exam centres
  - Locations for a business chain
  - Planning a political strategy

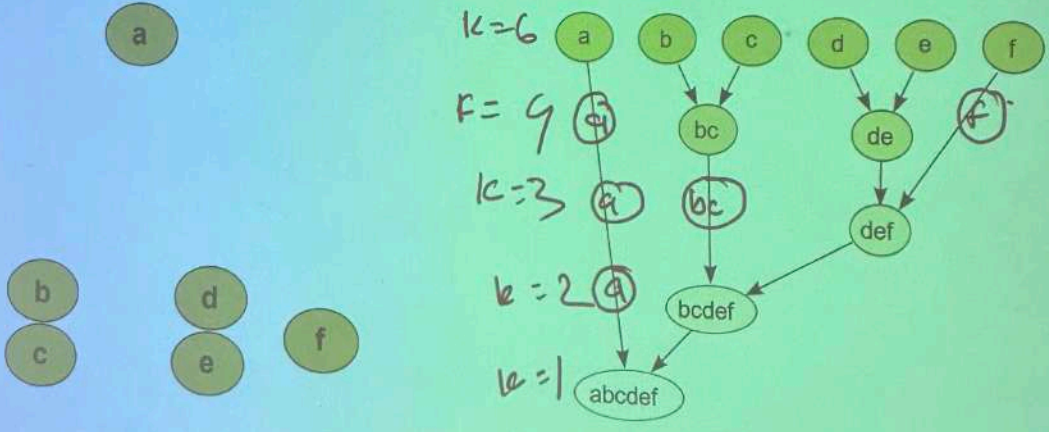
- Two Types of algorithms
  - Hard Clustering: Each point is assigned to some cluster
  - Soft Clustering: Each point is assigned probabilities of being assigned to each cluster

Hierarchical Clustering

- Agglomerative (e.g. AGNES):
- Start: Each point in separate cluster
- Merge 2 closest clusters
- Repeat until all records are in a single cluster
- Computationally Expensive
- Better at handling outliers

- Divisive (e.g. DIANA)
- Start: All points in 1 cluster
- Find most extreme points in each cluster.
- Regroup points based on closest extreme point
- Repeat until each record is in its own cluster
- Outliers may disrupt the

These are called Connectivity-based Clustering


a
b
c
d
e
f
k=6
F=9
k=3
k=2
k=1
bc
de
def
bcdef
abcdef

k-means

- Randomly select k points as centers
- Assign each point to the cluster closest to these centers
- Compute mean of all the points in the cluster
- Call these points as new centers of the cluster
- Continue still no further improvement

Step 1

Step 2

Step 3

Step 4

This technique falls under Centroid-based Clustering

DBSCAN

- Density-based spatial clustering of applications with noise (DBSCAN)
- By Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu in 1996... received Test of Time Award (ACM SIGKDD 2014)
- Key idea:
  - Closely connected points are marked into one cluster
  - Loosely connected are called noise
  - Non-parametric Clustering Algorithm

Optimization of DBSCAN

- $C = \{c_1, c_2, \ldots c_k\}$ be the clusters

- min k

st. $\mathrm{d}(\mathrm{p}, \mathrm{q}) \leq \epsilon \; \forall p, q \in c_j \; \forall c_j$

- A point p is a core point if at least minPts points with ϵ distance
- A point is directly reachable from a core point if it is within ϵ distance from a core point
  - Note, we talk about reachability only from core point

- A point q is reachable from p if there is a path $pp_1p_2 \dots p_nq$ such that p_{i+1} is directly reachable from p_i
  - what can we then say about p_i?

DBSCAN -- Algorithm

Find the points in the ε (eps) neighborhood of every point, and identify the core points with more than minPts neighbors.

Find the connected components of *core* points on the neighbor graph, ignoring all non-core points.

Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assign it to noise.

- minPts 4
- A and other red points are in the core
- B and C are not in the core
  - But reachable from A
  - All red points + B + C form a cluster
  - N is noise

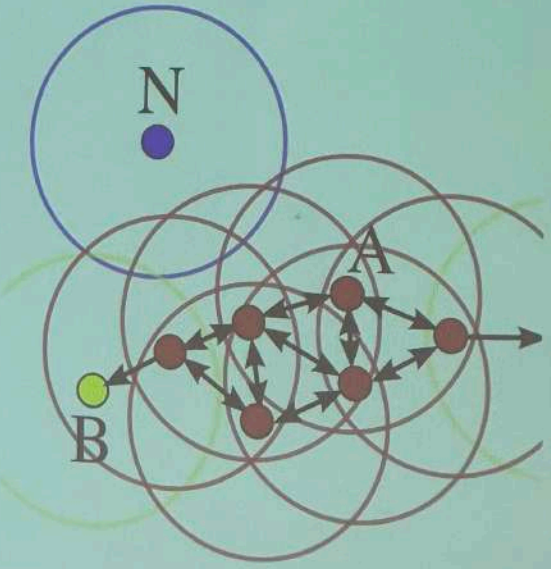


Image credits: wikipedia

Which k? Which Algorithm?

- Project on 2D/3D and determine if they are visually nicely separated
- Silhouette Score
  - S(i) = Based on average distance with points in the same cluster $a(i)$ vs smallest distance to the points not in the cluster $(b(i))$
  - $S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$
  - Average $S(i) \in [-1, 1]$, higher the better)

- Davies-Bouldin Index
  - Captures compactness of a cluster with its separation across clusters
  - $DB = \frac{1}{k}\sum_i \max_j \frac{\Delta X_i + \Delta X_j}{\delta(X_i, X_j)}$ Δ: *Intracluster distance* δ: *intercluster distance*