Lec - 24

- K-means :-

- Convergence :-

1) Proving congergence ⟷ Find a quantity that continously decreases. (or) [diff b/w G's at each iter]

2) ~~decreases~~ Show no. of G's is finite & G's decreasing / Σ center, point

$G_k = \Sigma_i (x_i - c_k)^2$: goodness measure

[But tbh, G decreases, goodness increases)

1) For the first step [assignment]

Any x_i changes from c_p to c_o

then $\text{argmin} [c_p - x_i]^2 < \text{argmin} [c_o - x_i]^2$

So each x_i decreases, if reassigned

⇒ Overall G. decreases.

2) mean is the point where error is min

$$\frac{\partial G_k}{\partial \mu_k} = \Sigma (x_{ik} - \mu_k)^2 \Rightarrow \Sigma c_k = \frac{\Sigma x_i}{n}$$

- Effect of PCA:

If not much information loss; → K-means will work pne

→ Highly influenced by outliers

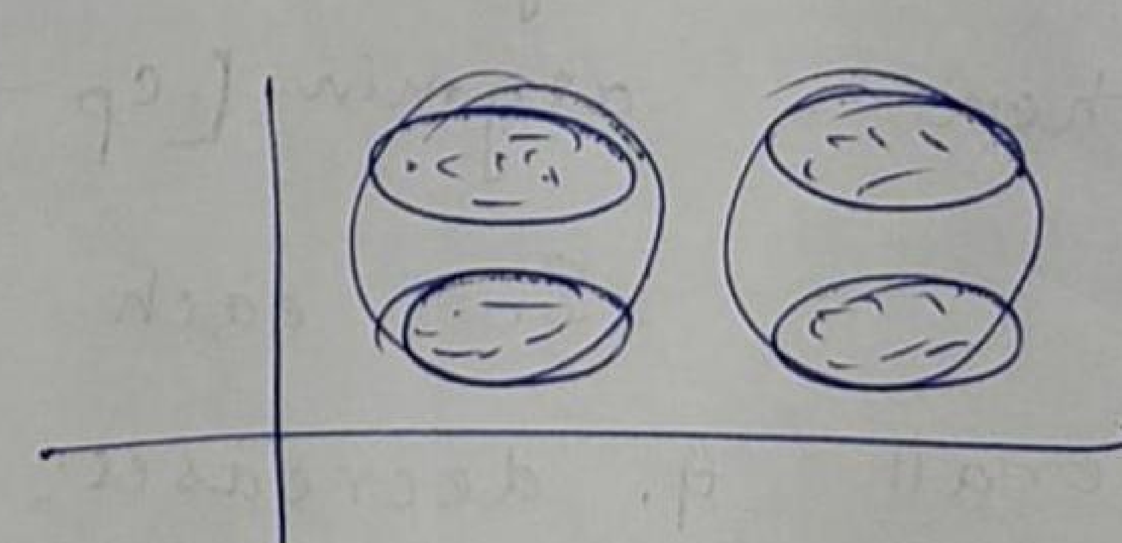
↳ Avoid by some threshold (distance b/w clusters)

• Variant of k-means
↳ One point is changed → μ changes.

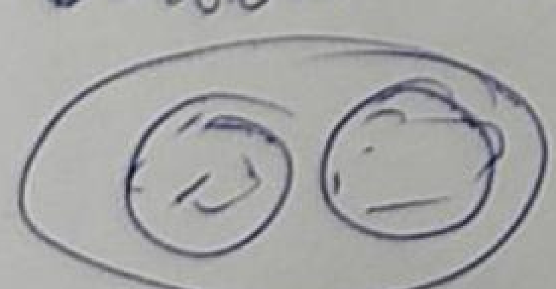
• Scaling affects, ~~PCA~~ k-means.
• PCA → won't affect much if enough PComps

• k vs G_k – tradeoff
• cosine distances are used instead of Euclidean distances

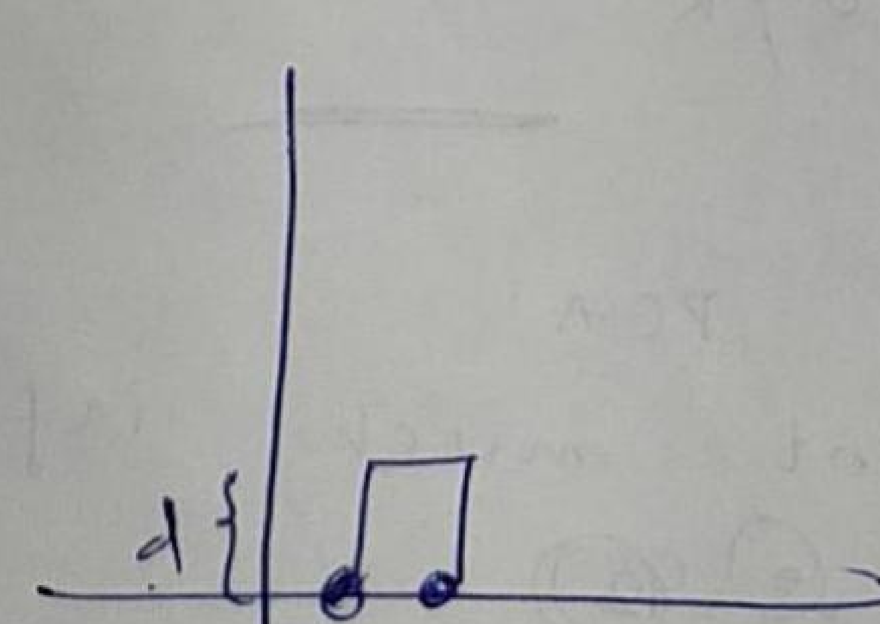
Guassian Mixture Model.
↳ far more slower convergency because of partial assignments, when compared to k-means
• Subclusters in real world
• Heirarchical clustering



• Top-down
↓
Apply k-means
• Bottom up – // Start with 1 cluster → 1 point

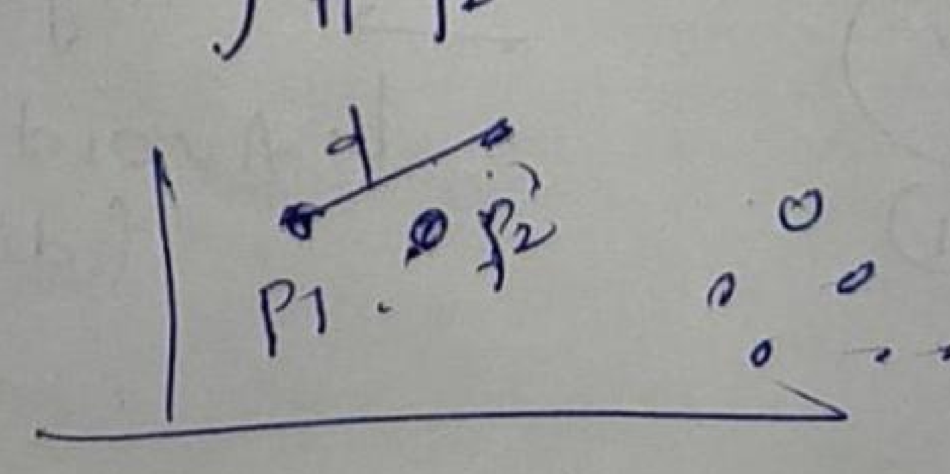


• Many points → find cluster distance

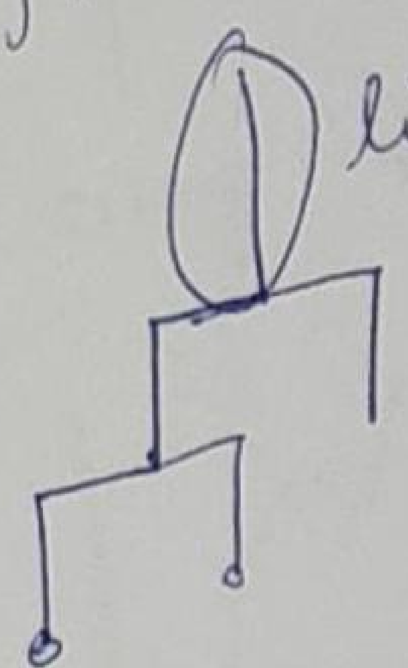


• p_1
• p_2
• d

dist (p_1, d) + dist (p_2, d) / 2



- Minimum ~~of~~ spanning tree.
- ~~[scribbled out]~~ O(   )
- After heirarchy is generated. Cut the tree
- life time



If limetime is long → Valid Cluster

How long → user defined

Distance
- Single link / (MST based) → min
  [$O(N^2 \log N)$ at ~~best~~] max
- Multiple link
- Avg link
- Centroid based

↳ As dendogram distance ↑, a single link is formed

↳ All ~~...~~ Mult. link points are to be clustered before further clusters

---

Single link prefers

Centroid updation ·cp np ·cq nq → $C_{new} = \frac{n_p c_p + n_q c_q}{p+q}$

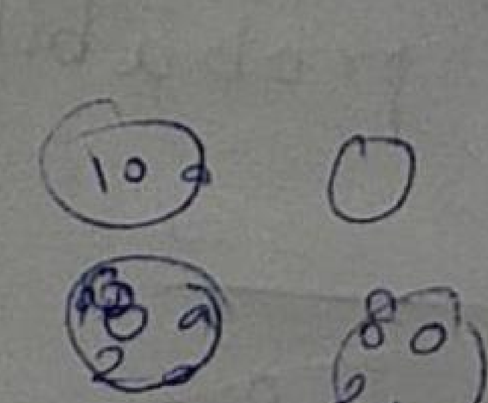
- Standard of the clustering algo
  - Cluster level-purity metric →
    ↳ maximally present are assumed to be correct
  - Biased; n clusters →

→ Over all purity:

↳ Rand index:

	A	B
	20	24
	20	78
	C	D

$$\text{Index} = \frac{A + D}{A+B+C+D}$$



• Clustering algorithms
↳ Semi- supervised learning
↓
Partial labels
↳ So: a, b ∈ One cluster
e, f definitely don't belong, together
Neednot be labels for everypair.

• Penality → Hard → ∞ penalty
↳ → intermediate penalty

Some samples, you assume labels and proceed
$a,b \in C \Rightarrow a, b \in$ apple
$a,e \notin C \Rightarrow a \Rightarrow$ apple
$e \Rightarrow$ orange

Stanford → PGM (Probabilistic Graphical Model – Daphne (Kohler)