Tour Version

Data Mining

Tour 5

Clustering

Clustering Pattern

Clustering

What is a Cluster Pattern?

- Partitioning a Set of Data into Subsets or Clusters

وهو إني أقسم البيانات لمجموعات كل مجموعة لها خصائص مشتركة يعني مثلاً الحيوانات لثديات و طيور وهكذا ،، ،، ،،

- objects in a Cluster are Similar ،، yet Dissimilar To objects in other Clusters

يعني إن ال objects اللي في نفس ال Cluster متشابهين والعكس لو Cluster تاني

يعني مثلاً الطائر لازم كلهم بيبيضوا وعندهم ريش لو جبت كائن تاني مش كده بيبقى مش طائر

Goal : discovery of Previously unknown groups within The data

- Implicit Classes

الهدف هو إكتشاف أصناف جديدة من البيانات ما كانش واخدين بالنا منها.

(أصناف مضمونة - مخفية)

remove noise عشان ن Pre-Processing منكم يستخدم من

وكمان Outlier Analysis

Requirements for Cluster Analysis

- Scalability =>
  معظم ال algorithms بت Handle Small Datasets
  فلو عندك Big Data حاول تعمل Sampling

- Handling Different attribute Types
  معظم ال Algorithms بتعالج بس ال Numerical Data

- Discovering Clusters with arbitrary Shape
  معظم ال Algorithms بطلع أشكال دائرية

- Domain Knowledge & input Parameters
  محتاج تفهم انت عايز ايه من عملية ال Clustering
  وإيه هي ال Parameters الي هتخش ال Algorithm

- Handling Noisy Data

حالياً ان Data فيها Noise بتأثر في شكل ال output بتاع ال Cluster (Sensetive)

- Incremental Clustring

موضوع الترتيب بيفرق ف عملية ال clustring ومحتاج تعملها كذا مرة عشان تعرف هل النتيجة مستقرة ولا لا

- Handling High Dimensionality

معظم ال Algorithm بتعالج كمية قليلة من ATTributes

- Constraint - Based Clustering

معظم ال Algorithms ماتقدرش تحدد لها ConstrainT إلا حاجة بسيطة زي K-Mean اللي بتقدر بس تتحكم في عدد ال Clusters

- Interpretability & usability

هل ال Algorithm طلع نتائج مفهومة ونقدر نستخدمها.

Comparing Cluster Analysis methods

ال Algorithms ممكن نقسمها على حسب

1. Partitioning Criteria

- flat — كل ~~Cluster~~ object بينتمي Cluster واحد
- Hiererical — ممكن ال object بينتمي ل different layers من Clusters

2- Separation of Clusters

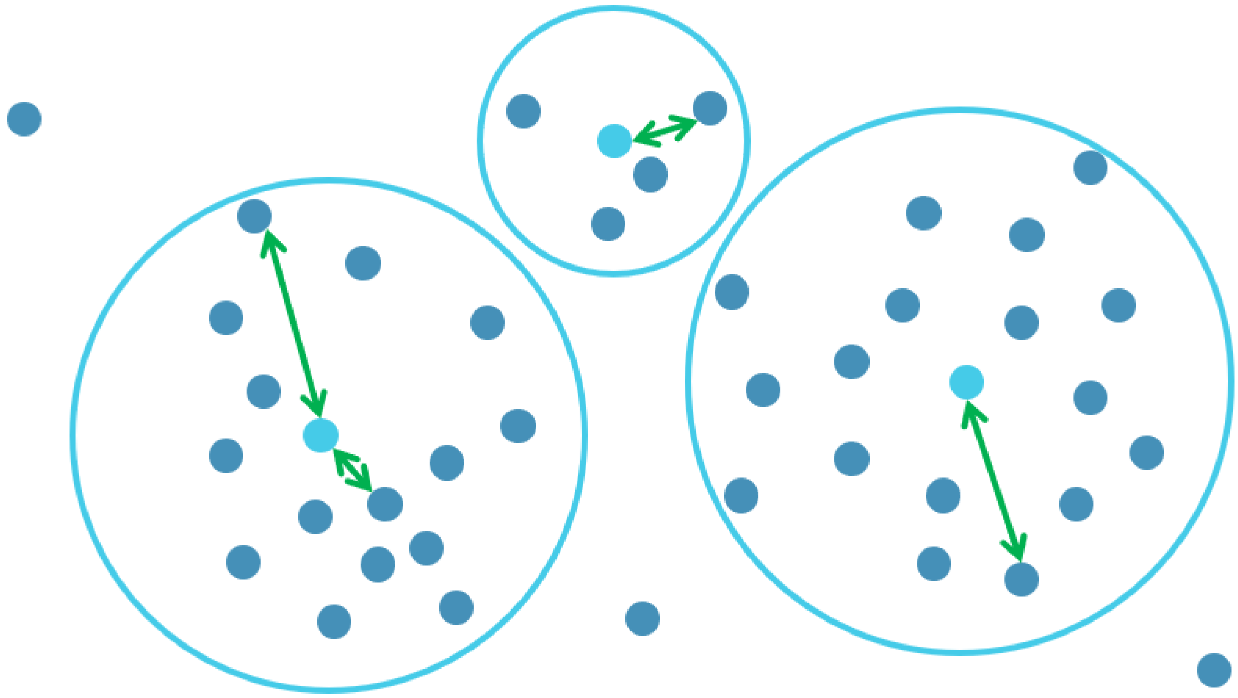
هل كل object بينتمي ل Cluster واحد mutualy execlusive

ان بينتمي لأكتر من Cluster (overlapping)

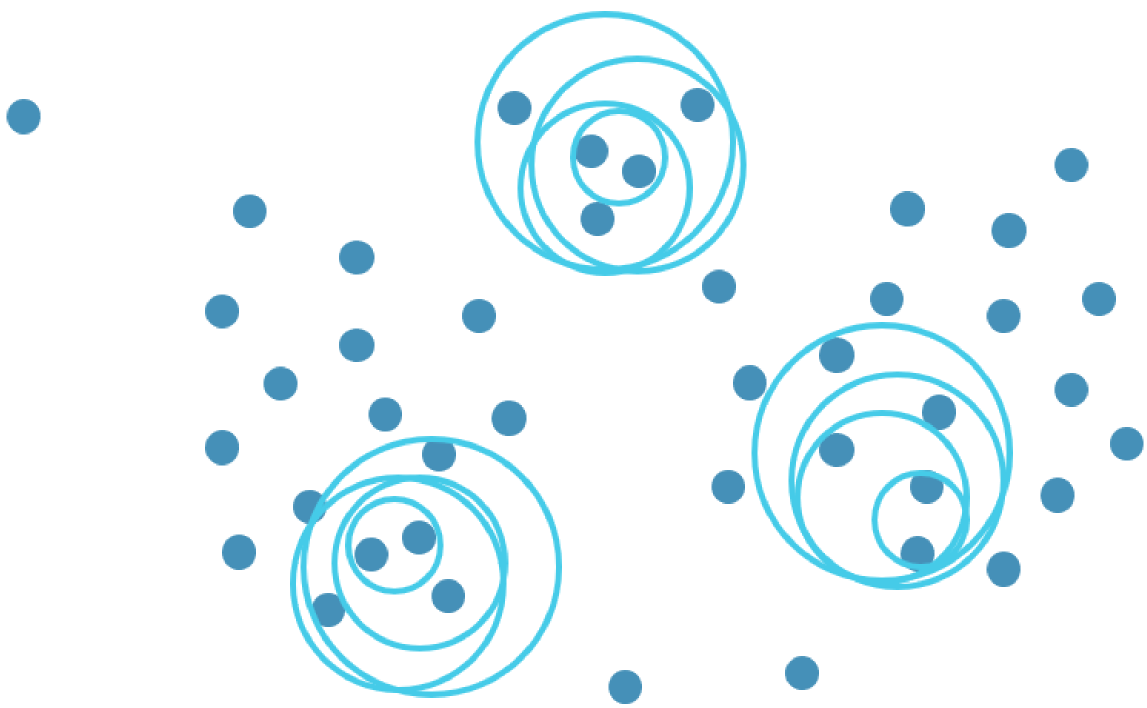
3. Similarity measure

هل ال Algorithm بيستخدم ال distance ولا denisity ← Threshold

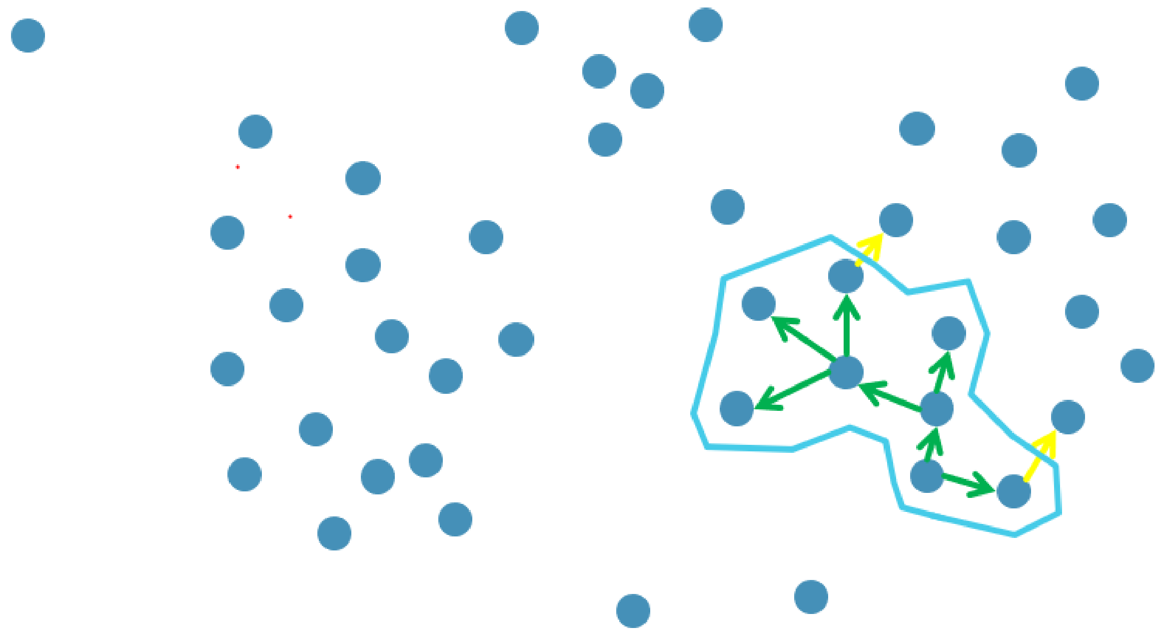
Partitioning methods	— Find *mutually exclusive* clusters of *spherical shape* — *Distance-based* — May *use mean or medoid* to represent cluster center — Effective for *small- to medium-size data sets*



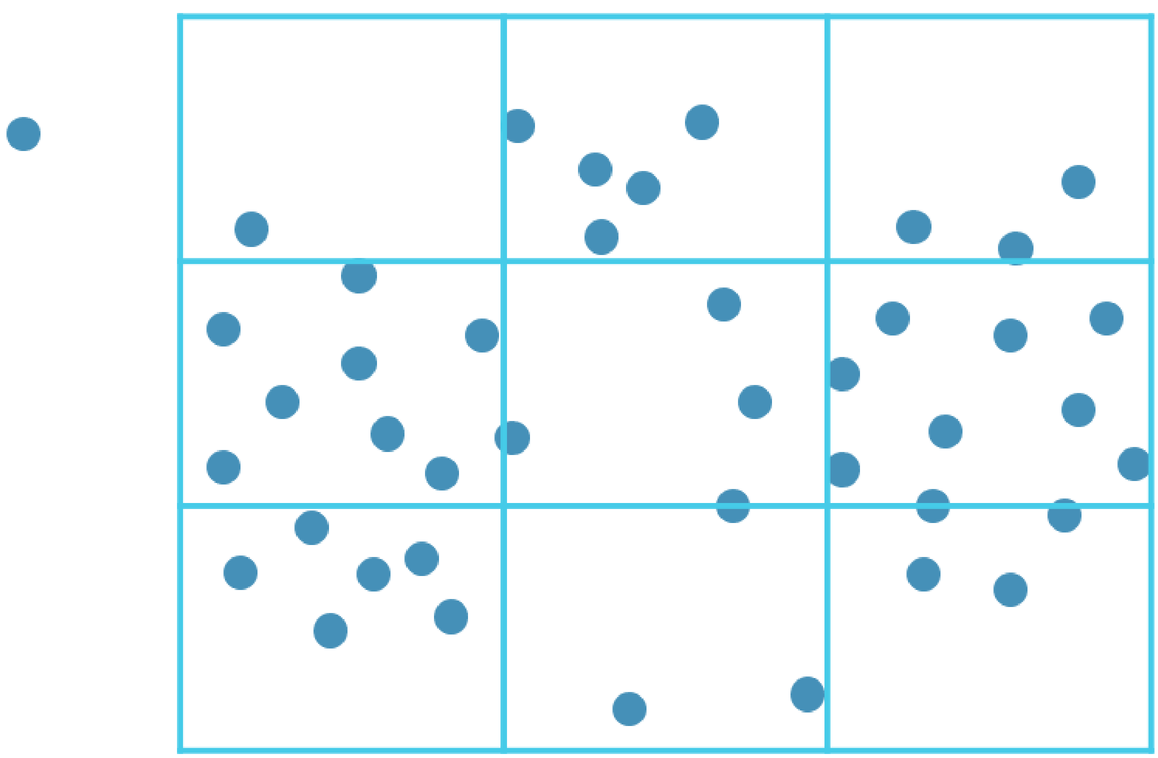
Hierarchical methods	— Clustering is *hierarchy* involving multiple levels — Cannot correct *erroneous merges/splits* — May consider object "*linkages*"



Density-based methods	— Can find *arbitrarily shaped clusters* — Clusters are *dense regions* separated by *low-density regions* — Each point must have a *minimum number of points within its "neighborhood"* — May *filter out outliers*



Grid-based methods	— Use a multi-resolution *grid data structure* — *Fast processing time*



Partitioning Methods {K-Means , K-Medoid}

K-MEANS

بيقسم البيانات لـ عدد K من ال Clusters

كل Cluster اللي بيعبر عنه Centroid وهي نقطة المنتصف للـ Cluster

ال object الأقرب لل Centroid اللي حددتهم بيبقى تبع لأقرب Centroid في ال Cluster

Quality of Clusters

حاول ان يكون الرقم ده أقل ما يمكن

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} dist(x, c_i)^2$$

حاجات خدها في اعتباراتك

- اختار عدد ال K كويس وايه هم ال Centroids
- احسب ال dissimilarity و ال means كويس
- لما تلاقي إن أحجام ال Clusters مختلفة جداً أو مش واحدة shape منتظم بيبقى حصل failure

	A1	A2
x1	2	10
x2	2	5
x3	8	4
x4	5	8
x5	7	5
x6	6	4
x7	1	2
x8	4	9

Cluster the eight points in table using k-means. Assume that k = 3 and that initially the points are assigned to clusters as follows: C1 = {x1, x2, x3}, C2 = {x4, x5, x6}, C3 = {x7, x8}.

• Apply the k-means algorithm until convergence (i.e., until the clusters do not change), using the Manhattan distance.

(Hint: The Manhattan distance is: $d(i, j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \ldots + |x_{in}-x_{jn}|$.) Make sure you clearly identify the final clustering and show your steps.

Algorithm: ***k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

K. MEANS

هنا كنا في المسألة بانه عمل أول iteration

iter 1

وعدد ال $K = 3$

$C_1 = \{X_1, X_2, X_3\} \Rightarrow \overline{C_1} = (4, 6\frac{1}{3})$

$C_2 = \{X_4, X_5, X_6\} \Rightarrow \overline{C_2} = (6, 5\frac{2}{3})$

$C_3 = \{X_7, X_8\} \Rightarrow \overline{C_3} = (2\frac{1}{2}, 5\frac{1}{2})$

iter 2	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	New mean
C1	$5\frac{2}{3}$	$3\frac{1}{3}$	$6\frac{1}{3}$	$2\frac{2}{3}$	$4\frac{1}{3}$	$4\frac{1}{3}$	$7\frac{1}{3}$	$2\frac{2}{3}$	$(4\frac{1}{2}, 8\frac{1}{2})$
C2	$8\frac{1}{3}$	$4\frac{2}{3}$	$3\frac{2}{3}$	$3\frac{1}{3}$	$1\frac{2}{3}$	$1\frac{2}{3}$	$1\frac{2}{3}$	$5\frac{1}{3}$	$(7, 4\frac{1}{3})$
C3	5	1	7	5	5	5	5	5	$(1\frac{2}{3}, 5\frac{2}{3})$

manhatin distance $(X_1, \overline{C_1}) = |4-2| + |10 - 6\frac{1}{3}| = 5\frac{2}{3}$

" " $(X_1, \overline{C_2}) = |6-2| + |10 - 5\frac{2}{3}| = 8\frac{1}{3}$

" " $(X_1, \overline{C_3}) = \ldots\ldots = 5$

$C(X_1) = \min(5\frac{2}{3}, 8\frac{1}{3}, 5) = 5$

وهكذا لحتى ما تعبي كل الجدول الي فوق

iter 2 Clusters

$C_1 = \{X_4, X_8\} \Rightarrow \overline{C_1} = \frac{(5+4, 8+9)}{2} = (4\frac{1}{2}, 8\frac{1}{2})$

$C_2 = \{X_3, X_5, X_6\}$

$C_3 = \{X_1, X_2, X_7\}$

- و هتفضل شغال كده في iterations لغيت اما تلاقى ال clusters شكلها ثبتت و ده الحل

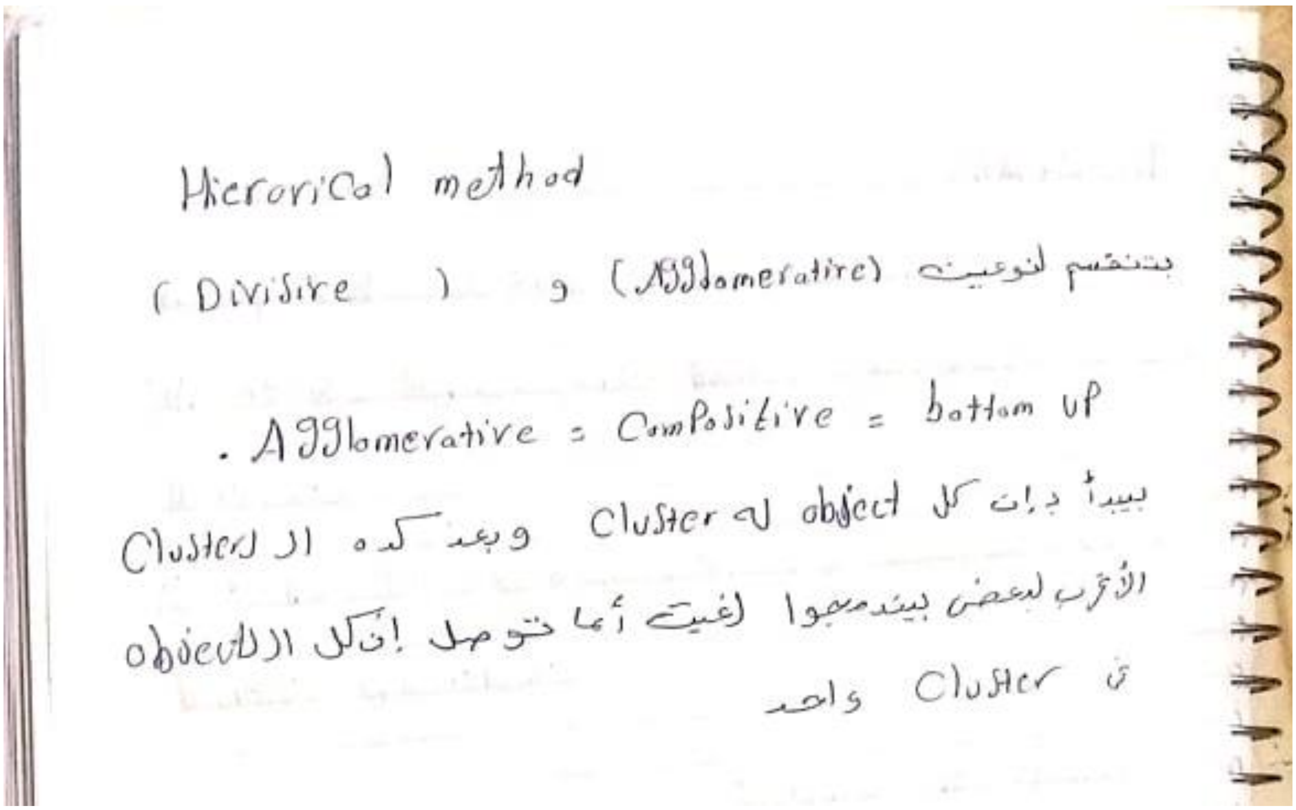
$C1 = \{x1,x4,x8\} = \{(2,10), (5,8), (4,9)\}$ Mean of $C1 = (2\frac{2}{3}, 9)$

$C2 = \{x3,x5,x6\} = \{(8,4), (7,5), (6,4)\}$ Mean of $C2 = (7, 4\frac{1}{3})$

$C3 = \{x2,x7\} = \{(2,5), (1,2)\}$ Mean of $C3 = (1\frac{1}{2}, 3\frac{1}{2})$

Hierarchical Methods {Agglomerative vs Devisive}

وبنطلع فيها ال clusters على هيئة شجرة متتابعة و فيه منها نوعين من ال methods

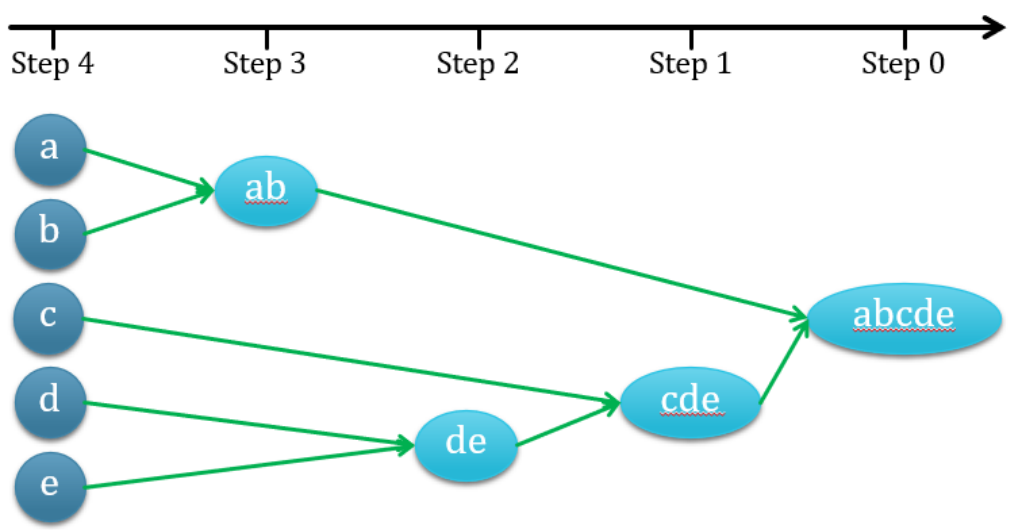


Hierarical method

بتنقسم لنوعين (Agglomerative) و (Divisive)

. Agglomerative = Compositive = bottom up

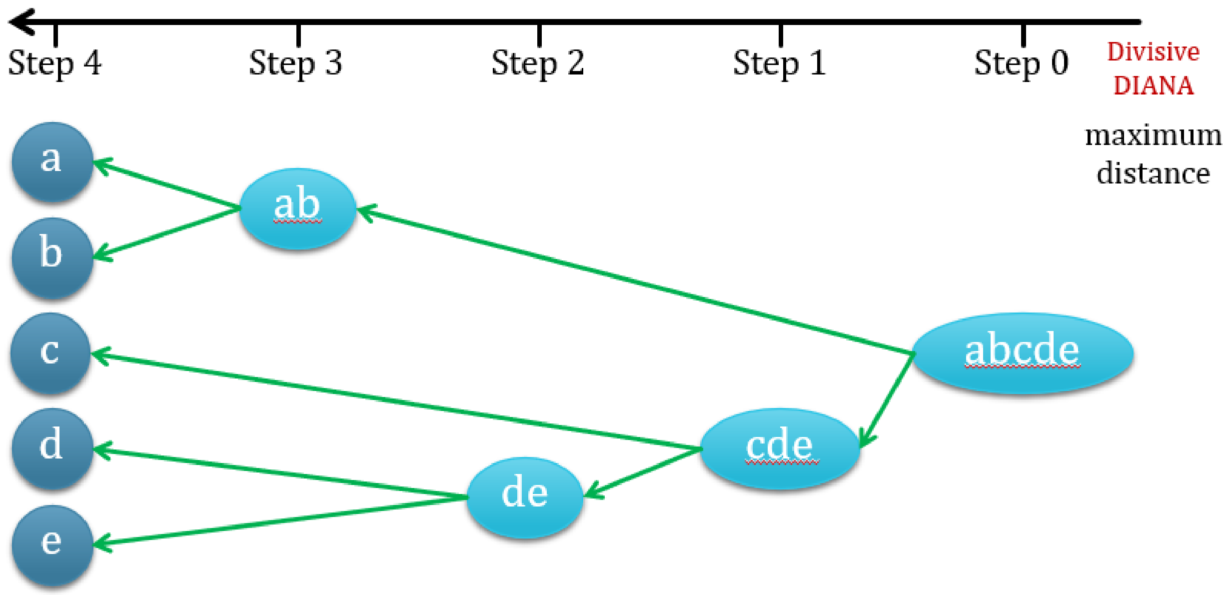
بيبدأ بذات كل object له cluster وبعد كده ال clusters الأقرب لبعض بيندمجوا لغيت أما توصل إن كل ال objects في cluster واحد



Divisive = Top-down = Split decomposition

ودي عكس اللي فاتت بيبدأ إن كل ال object في cluster

وبعد كده بينقسموا لغاية ما يكون لكل object ⇐ Cluster



How to divide a cluster is a challenge! Heuristic approaches may be used

Density Based Methods {DBSCAN}

DBSCAN

- Density - Based Spatial Clustring of Applcation with Noise

. find Core objects [with dence Neighborhoods].

- محور [DB method] هو Core object كان Partitioning

هو ال Centroid -

- بيحاول إنه يلاقي Core obj يكون له Neighbors متحققين كثافة معينة

Radius : ε

ε-neighborhoods : all obj inside the Circle

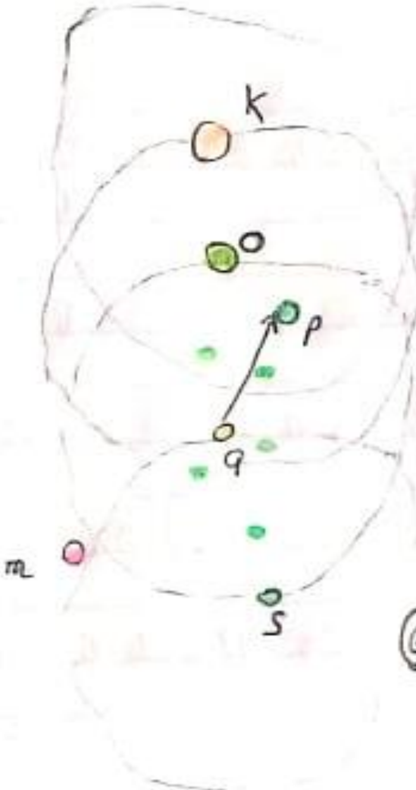
Min-Pts → Threshold for Neighborhood

Density = 6 # of ε-neigh

Core obj = o

هو أي obj محقق كثافة في ال region اللي تبعه أكبر من أو بيساوي ال min-Pts

Core object : obj whose ε-neighborhood Contains at least minPts objects



Given $\epsilon = 4$ and minPts $= 5$

(S) مش Core لأنه محقق (3) Neighborhood

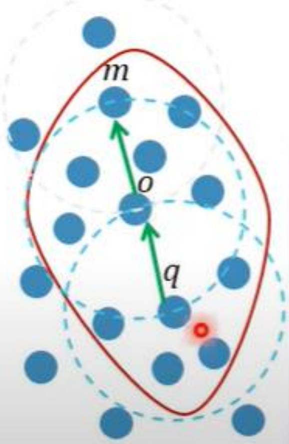
(q) هو Core لأنه محقق أكثر من $\epsilon = 4$

(P) يعتبر Density Reachable لأنه من (ε.q) لـ q

(m) مش Density Reachable لـ q لأنه مش (ε n q)

(q and k) هم density connected لو فيه عنصر مشترك زي object بينتمي لل cluster2 و Core objects ∈ q, k

Given $\epsilon = 4$ and $MinPts = 5$



an object ***p*** is directly **density-reachable** from another object ***q*** if and only if ***q*** is a core object and ***p*** is in the -neighborhood of ***q***

objects ***q*** & ***m*** are **density-connected** if there is an object ***o*** such that ***q*** & ***m*** are both density-reachable from ***o***

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ϵ-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ϵ-neighborhood of p;
(8)          for each point p′ in N
(9)              if p′ is unvisited
(10)                 mark p′ as visited;
(11)                 if the ϵ-neighborhood of p′ has at least MinPts points,
                     add those points to N;
(12)             if p′ is not yet a member of any cluster, add p′ to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

EVALUATION OF CLUSTERING ASSESSING CLUSTERING TENDENCY

هنعمل Cluster Tendency عشان نعرف ال Quality لل Clustring

ونكتشف إن ال Dataset مالهاش شكل عشوائي

عن طريق Hopkins Statistics

← بنختار أول حاجة Sample عددهم n من ال Points

ولكل نقطة هنشوف مين الأقرب Neighbors لها ونحسب

ال distance ما بينها وما بين كل ال neighbors

$$distance\ X_i = \min_{v \in D} \{dist(P_i, v)\}$$

نختار أقل مسافة طلعت لكل ال Samples

ونختار مرة Sample عدد هو n ونعمل نفس اللي حصل فوق Y_i

ونأخذ النتيجتين ونحسب

Hopkins Statistic

$$H = \frac{\sum Y_i}{\sum X_i + \sum Y_i}$$

المفروض إنه هي تساوي نحو ⇐ 0.5

$$H \ncong 0.5$$

عشان ال method تكون إشتغلت صح

Supervised Testing

Extrinsic method

بنقارن ما بين النتائج وأرض الواقع

(Ground Truth)

بيتطلب إنه نكون عارفين فعلياً ال classes

لبعض ال objects

ونقارن ما بينهم ونشوف النسبة

$Q(C, C_g)$

محتاجين نعرف من خلاله

- Cluster Homogeneity

هل كل object له class label واحد بس

على أساس إن cluster بيمثل class label

- Cluster Completeness

بيشوف هل ال object له class label اللي ال cluster بتاعه

بيعبر عن ال class label

- Rag Bag

ال objects إللي مالهاش cluster بنعبيها فيها

- Small Cluster Preservation

زي مثلاً ال divisive Hierarchy

لازم أحافظ على توازن أحجام

ال clusters

زى ال BCubed

بنحسب ال Precision وال recall لكل object

Precision =>
كم عدد ال object الى فى cluster واحد بينتموا لنفس category

Recall =>
كم عدد ال object الى ليهم نفس ال class label ليهم نفس ال clusters

Extrinsic methods → compare clustering against **ground truth** (*supervision*)

- Assign a score $Q(C, C_g)$ to capture:
  - **Cluster homogeneity** → the purer the better – clusters represent separate class labels
  - **Cluster completeness** → an object with a class label belongs to the cluster representing that class label
  - **Rag bag** → objects that can't be merged into clusters belong to a *rag bag* – penalize a *misc. object* when put in a *pure cluster* more than in a rag bag
  - **Small cluster preservation** → splitting a small category is *more harmful* than splitting a large category
- Ex. **BCubed** *precision* and *recall* of every object in dataset:
  - Precision → how many objects in the same cluster ∈ the same category as the object
  - Recall → how many objects of the same category are assigned to the same cluster

Intrinistic methods

ودي بتختبر أد إيه ال Clusters اللي طلعت مفصولة كويس

مثال عليها Silhouette Coefficient

بتقدر تفرق ما بين

① ← مقدار الفرق في المسافة (متوسط) ما بين object وكل ال objects اللي معاه في نفس ال Cluster

• كل ما كان المتوسط أقل طبعاً نتيجة أصح

② - مقدار أقل متوسط لفرق المسافة ما بين ال object وباقي ال Clusters دي مش ال Cluster اللي بينتمي له object

• كل ما كان كبير طبعاً هيبقى أفضل

لو المقدار التاني - الأول طلع موجب يبقى ال clustering كويس

لو العكس يبقى ال Clustering غلط

•Compute **the silhouette coefficient** for object **x1.**
What is the meaning of the computed value?

C1= {x1,x4,x8}={(2,10), (5,8), (4,9)} Mean of C1= $(2\frac{2}{3},9)$
C2= {x3,x5,x6}={(8,4), (7,5), (6,4)} Mean of C2 = $(7, 4\frac{1}{3})$
C3= {x2,x7}={(2,5), (1,2)} Mean of C3 = $(1\frac{1}{2}, 3\frac{1}{2})$

- $a(o) = \frac{\sum_{o' \in c_i} dis(O,O')}{|c_i|-1} = \frac{5+3}{2} = 4$
- $b(o) = \min\left\{\frac{\sum_{o' \in c_j} dis(O,O')}{|c_j|}\right\} = \min\left\{\frac{12+10+10}{3}, \frac{5+9}{2}\right\} = 7$
- $S(o) = \frac{b(o)-a(o)}{\max\{a(o),b(o)\}} = \frac{7-4}{7}$ ➔ +ve
- This mean the cluster containing o is compact and o is far from other cluster