Klustring

Agenda

- Klustring - K-means
- Hierarkisk klustring

Kod

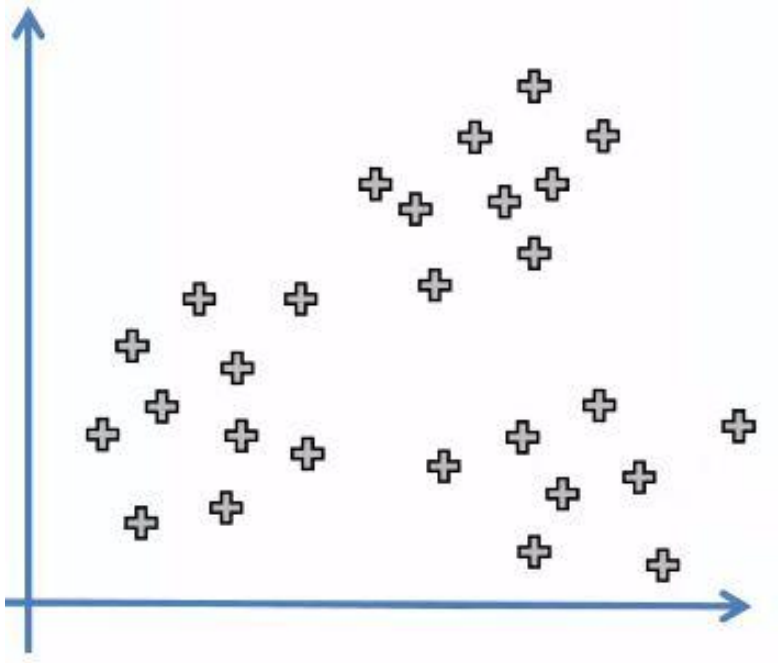
- Klustring

Klustring - Grupper

Till skillnad från *supervised* ML - inga “svar”

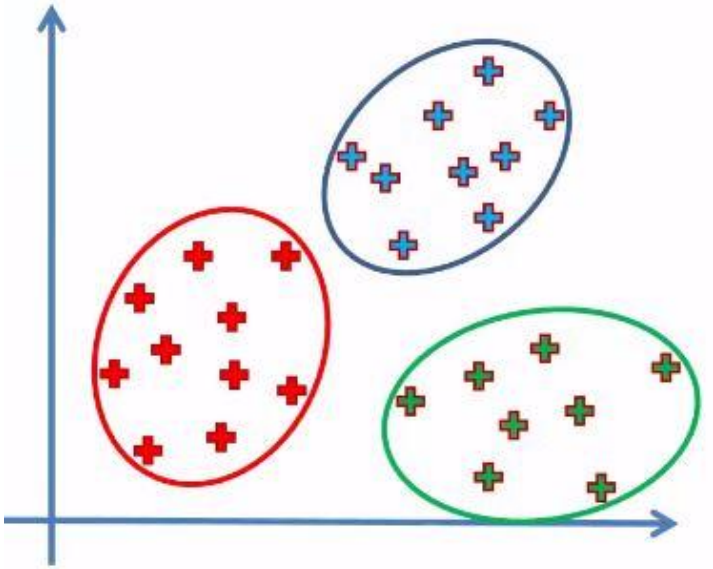
Hur grupperar man

How to cluster



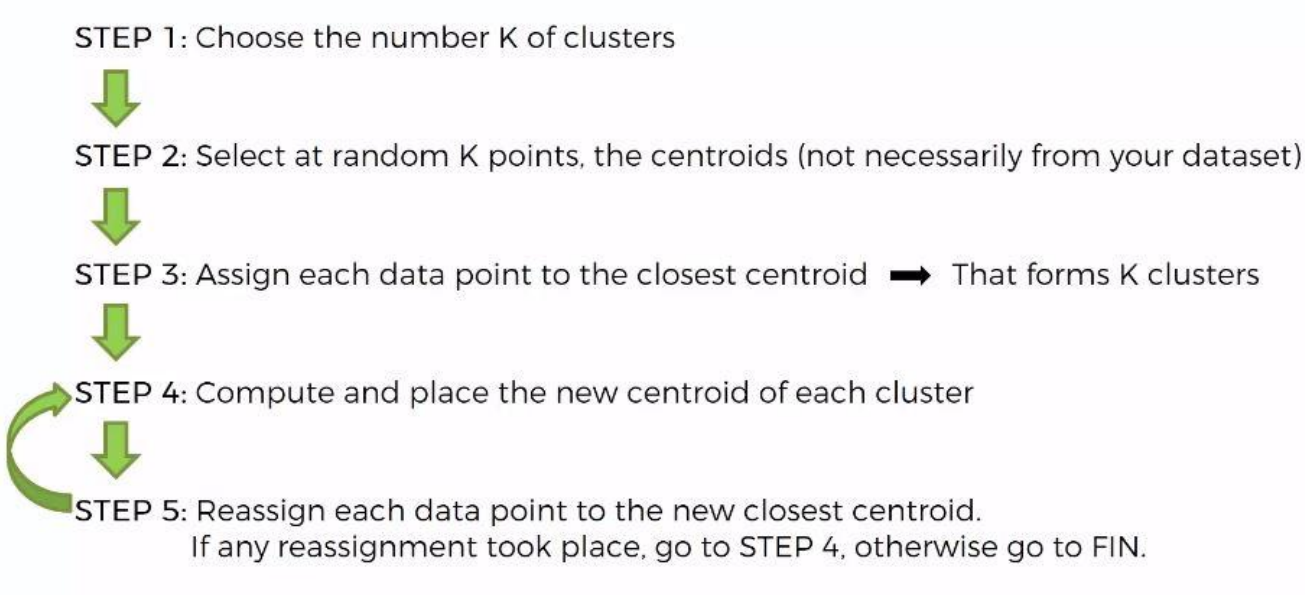
Klustring -

Intuitivt: ”Klumpar” som hör ihop



Klustring - Algoritm

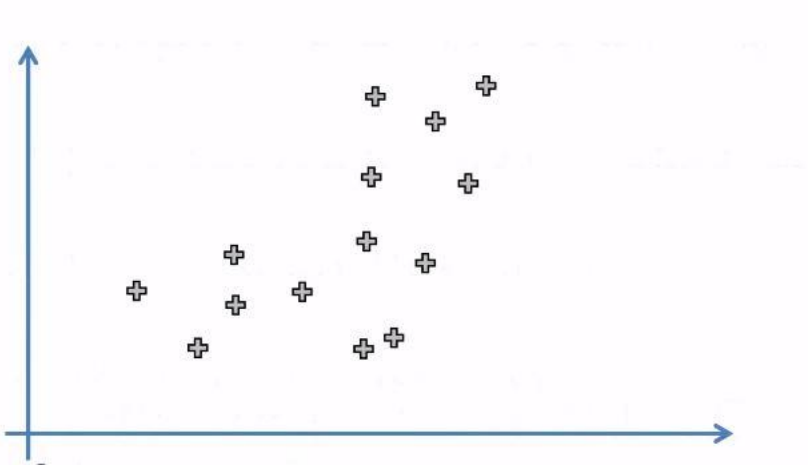
Princip för K-Means algoritmen



Upprepade steg för att utföra den s.k. kustringen / clustering

Klustring - Algoritm Exempel

Exempel: Dela in detta dataset i två kluster enligt K-means metoden

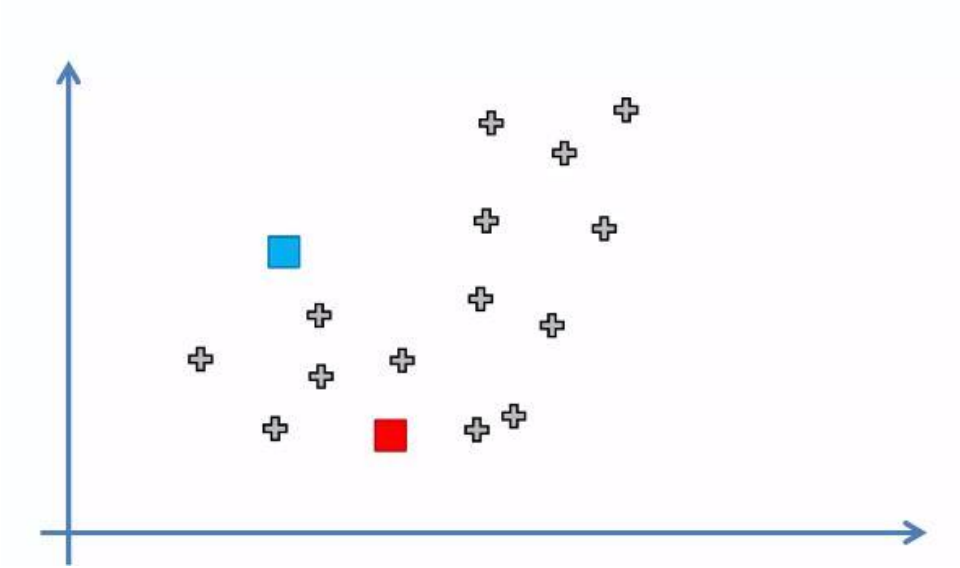


Klustring - Algoritm steg 1&2

Steg 1: Välj antal kluster: 2

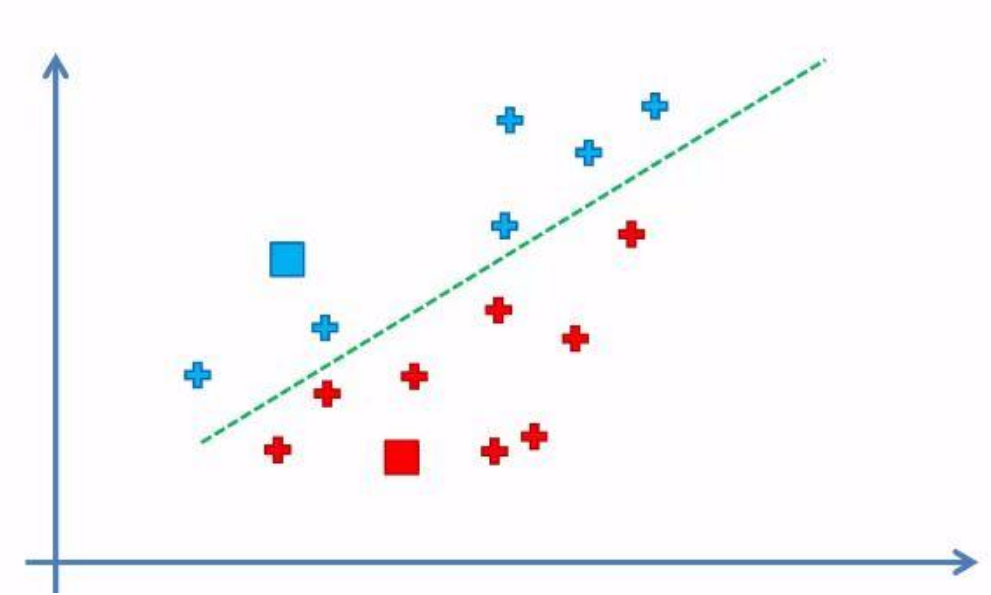
Steg 2: Placera ut 2 (antal kluster) punkter slumpmässigt.

Dessa kallas centroider.



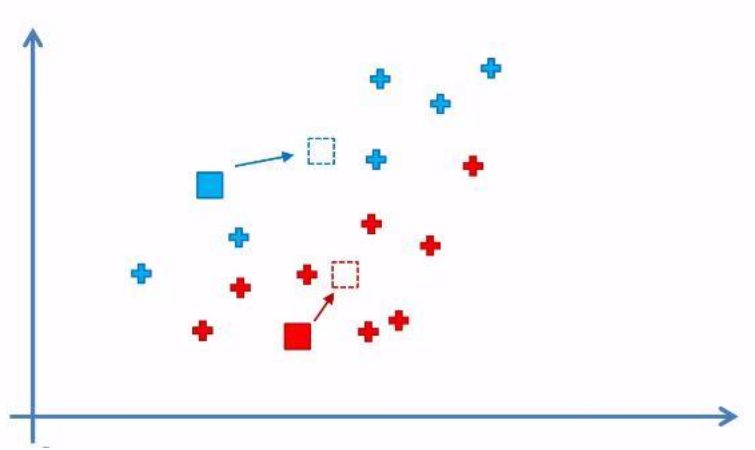
Klustring - Algoritm steg 3

Steg 3: Tilldela nu varje datapunkt den klass som centroiden har som ligger närmast



Klustring - Algoritm steg 4

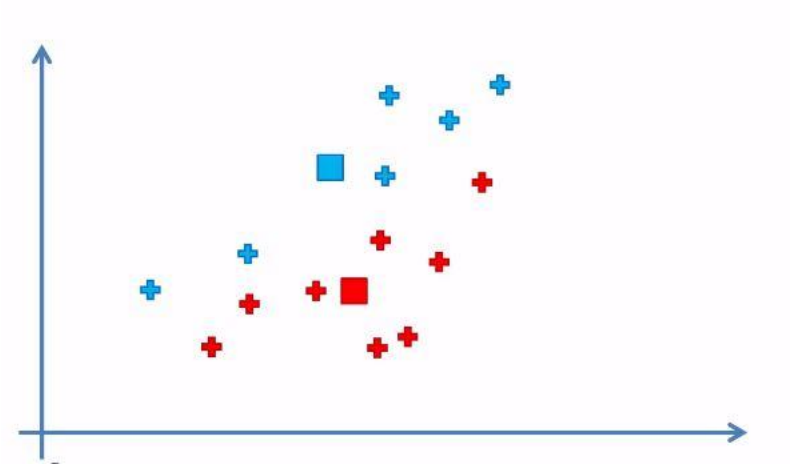
Steg 4: Flytta nu centroiden för varje kluster till respektive tyngdpunkt.



Klustring - Steg 3

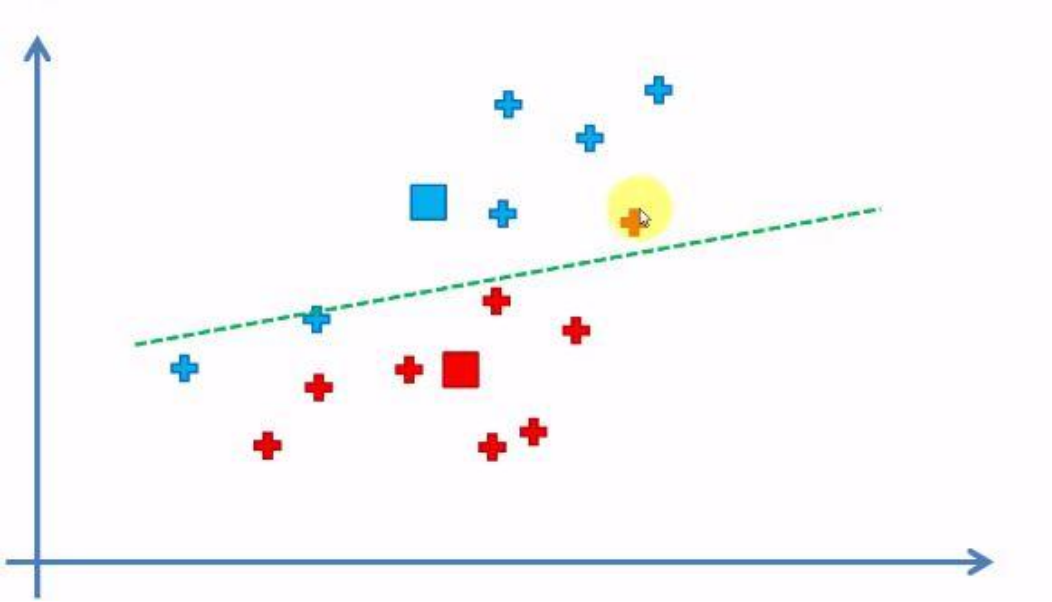
Upprepa steg 3.

När centroiderna är flyttade, förberedd för att åter tilldela varje punkt en klass.



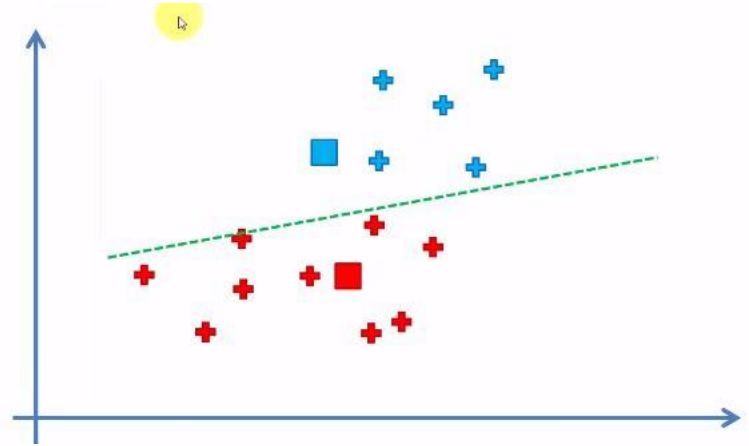
Klustring - Steg 3

Det kan nu hända att vissa punkter kommer närmare en annan centroid än innan



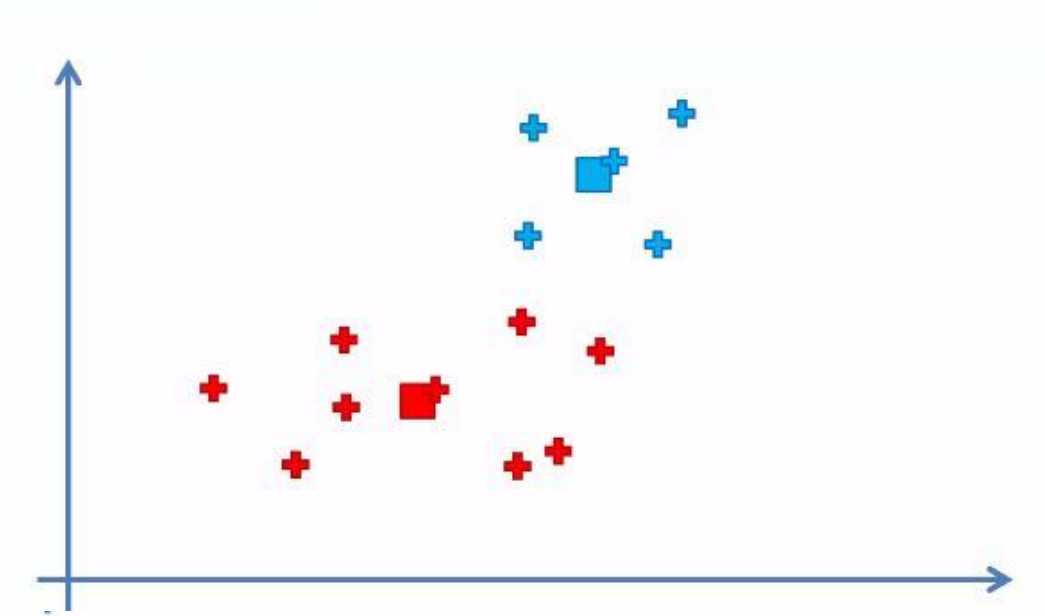
Klustring - Algoritm steg 5

Steg 5: Om någon datapunkt kommer närmare en annan centroid, ändra dess klass och gå tillbaka till steg 4



Klustring - Algoritm steg 4

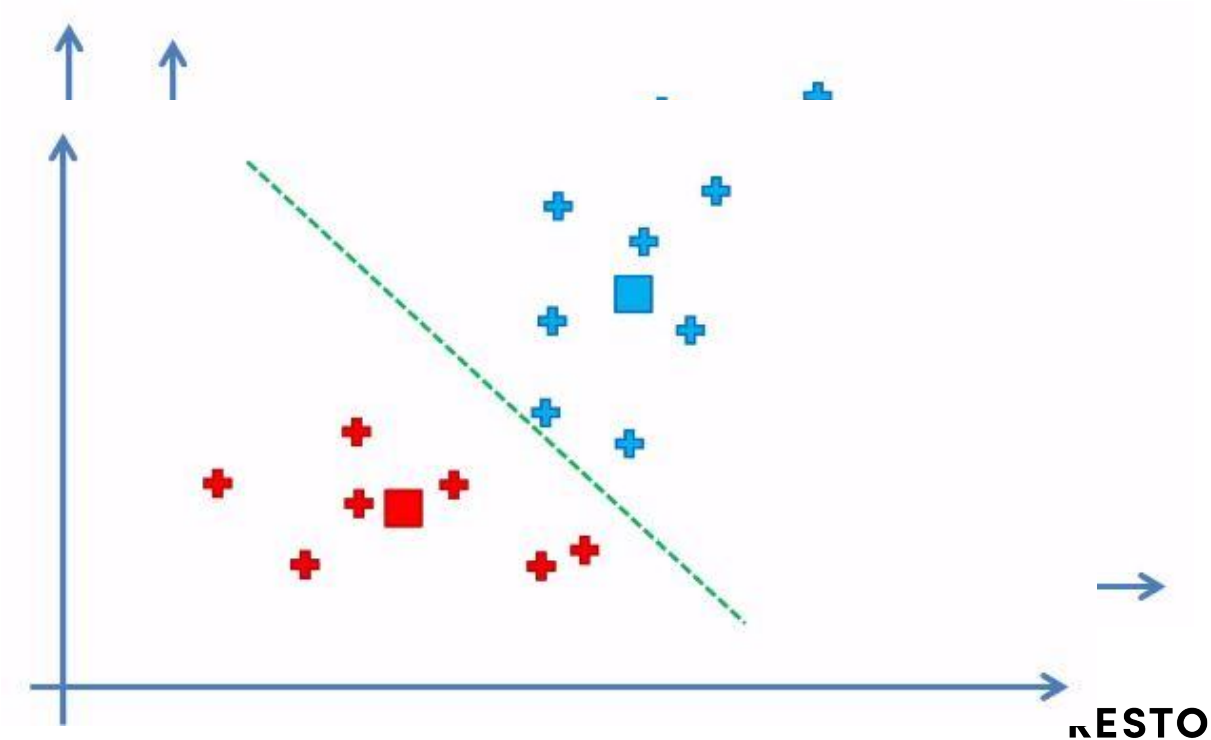
Repetera steg 4



Klustring - Algoritm steg 5 & 4

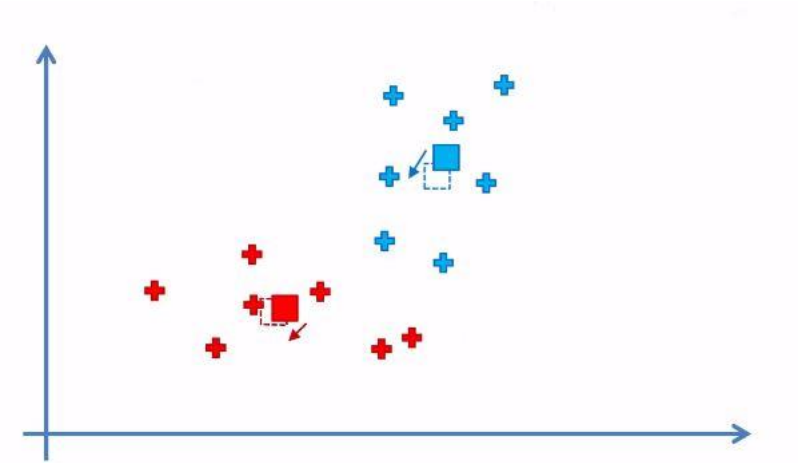
Repetera steg 5

Repetera steg 4



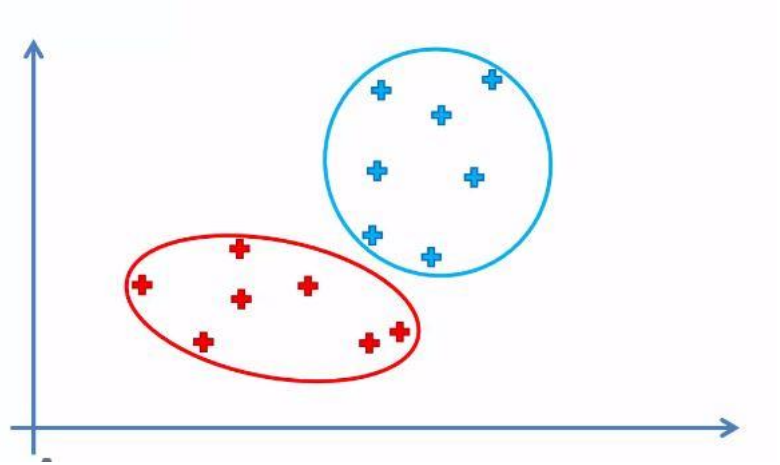
Klustring - Algoritm

När ingen punkt längre ändrar klass mellan steg 4 och steg 5 är modellen färdig!



Klustring - Klar

Slutgiltig klustring

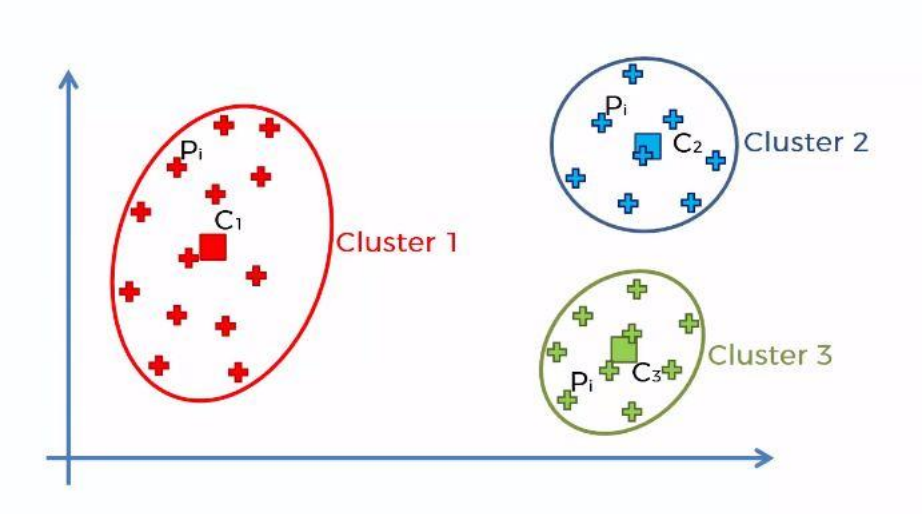


Klustring - Val av K

Hur väljer man bäst antal kluster K?

Klustring - Val av K

Vi behöver ett mått som beskriver klustringens effektivitet!



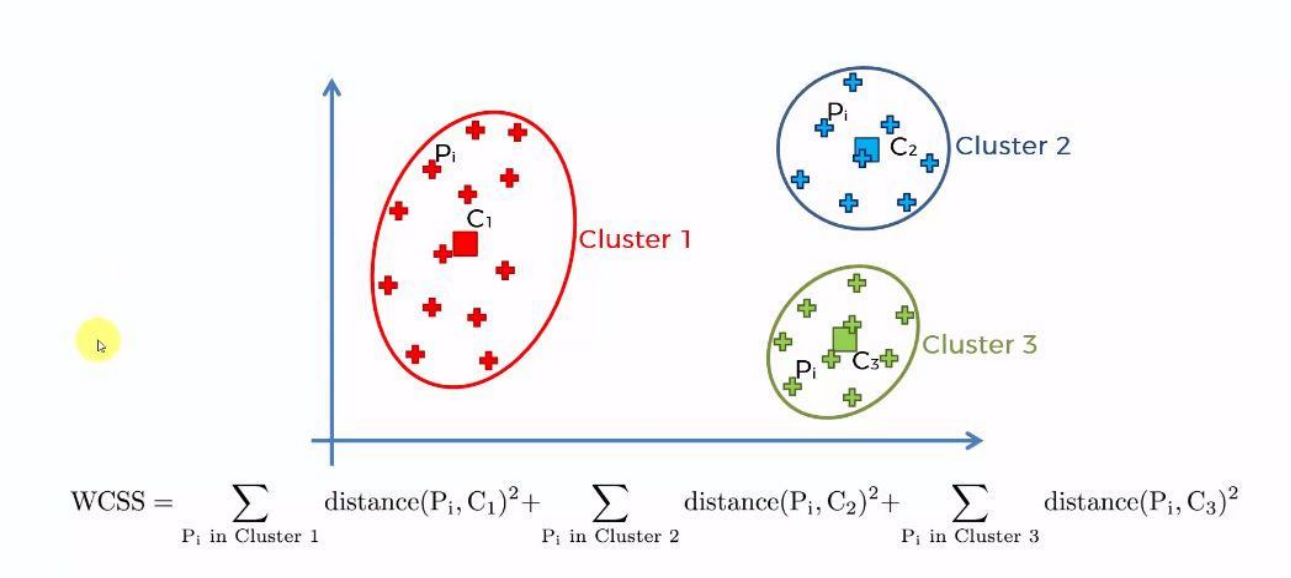
Klustring - Val av K

WCSS - Within Cluster Sum of Squares

$$\mathrm{WCSS} = \sum_{P_i \text{ in Cluster } 1} \mathrm{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster } 2} \mathrm{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster } 3} \mathrm{distance}(P_i, C_3)^2$$

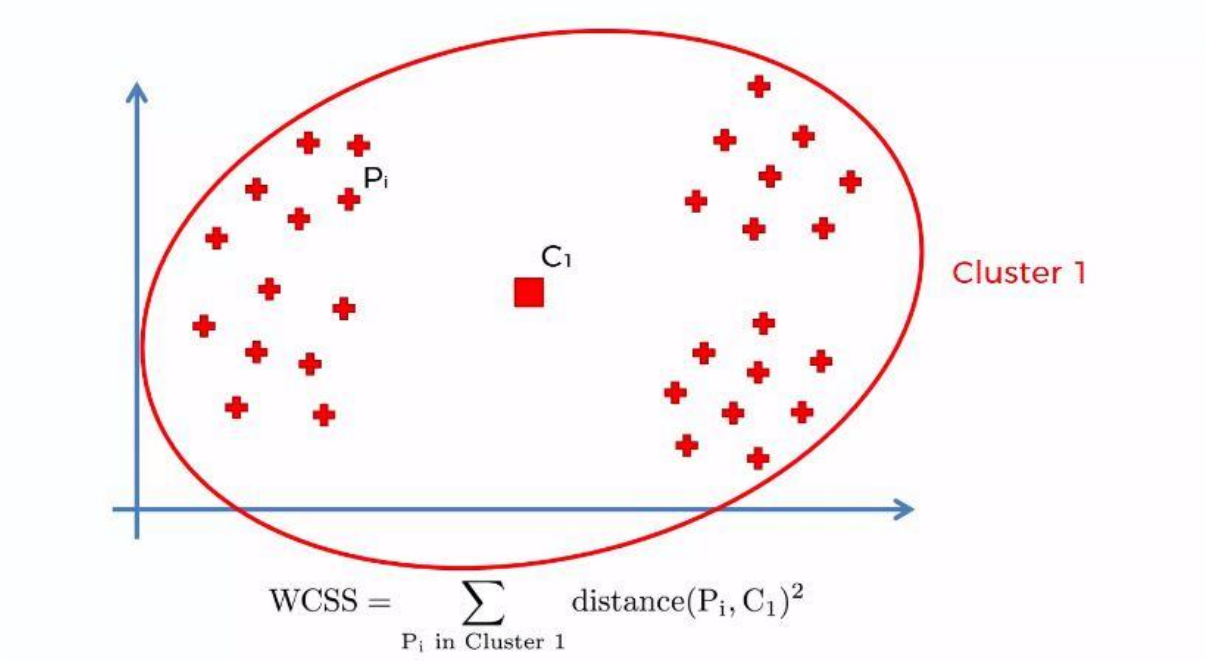
Klustring - Val av K

Räkna ihop de kvadrerade avstånden från varje centroid till datapunkterna i klustret.



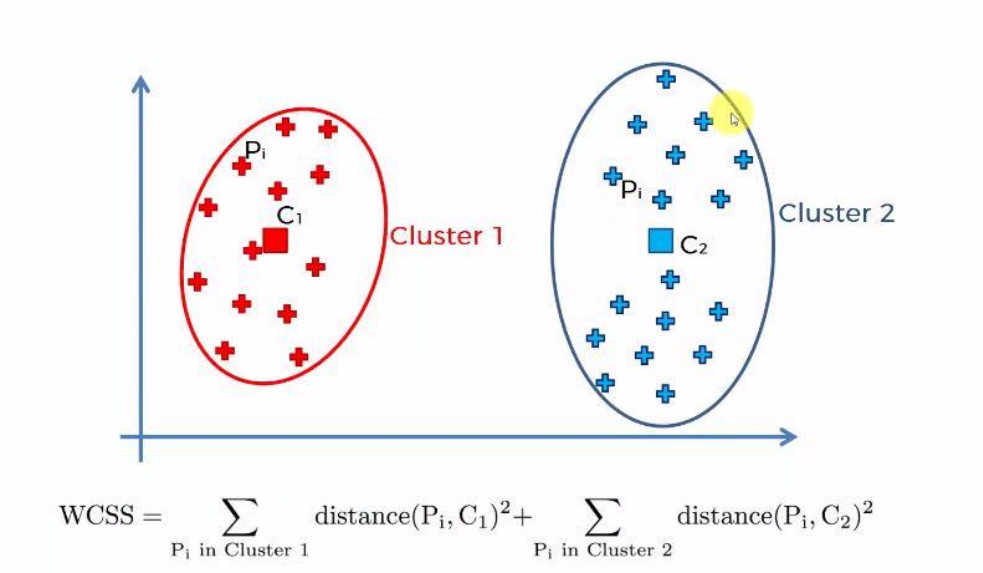
Klustring - Val av K

Om vi bara har ett kluster?



Klustring - Val av K

Om vi har två kluster

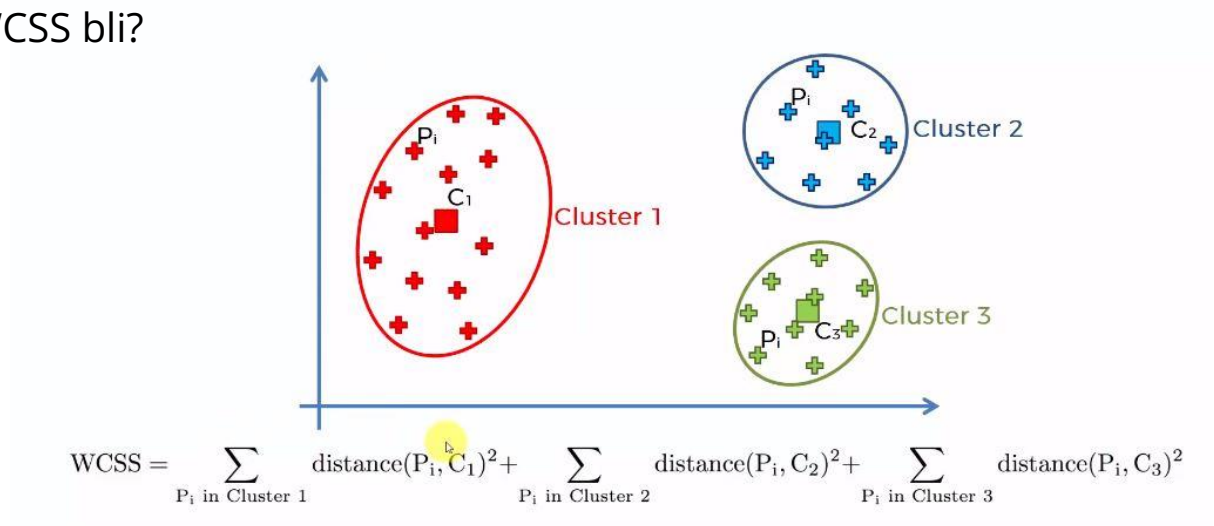


Klustring - Val av K

Om vi har tre kluster?

Vad händer med WCSS när K växer?

Hur litet kan WCSS bli?

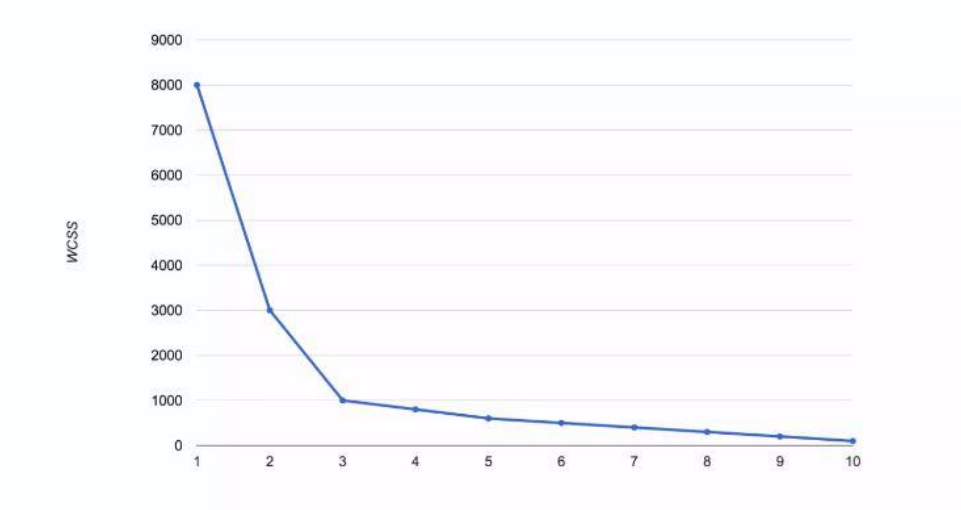


RESTORY...

Klustring - Val av K

Armbågsmetoden/ Elbow method

Hur länge ’tjänar’ man på att lägga till ett kluster



Hierarkisk klustring

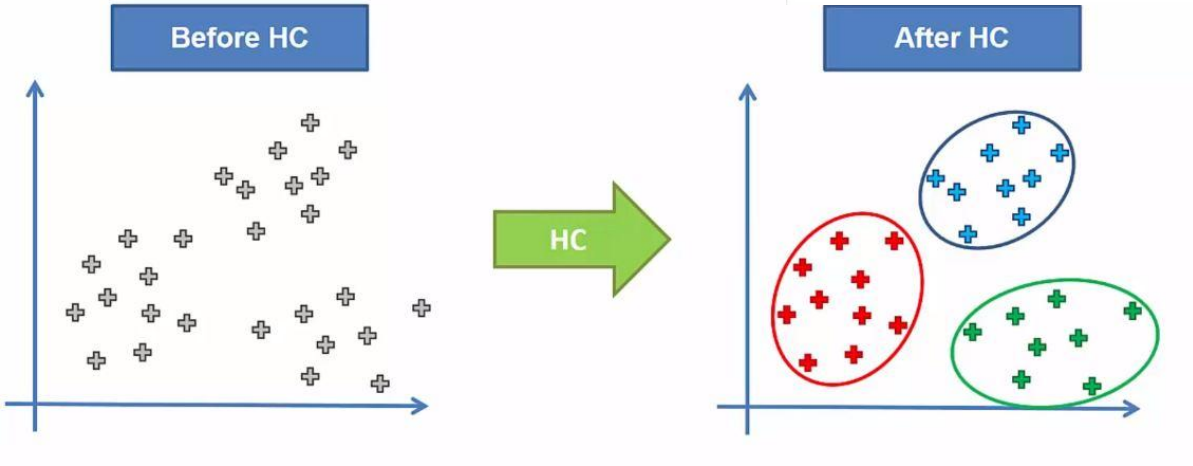
Hierarkisk Klustring - Agglomerativa metoden

Definitions of *agglomerative*

1. adjective clustered together but not coherent

synonyms: agglomerate, agglomerated, clustered

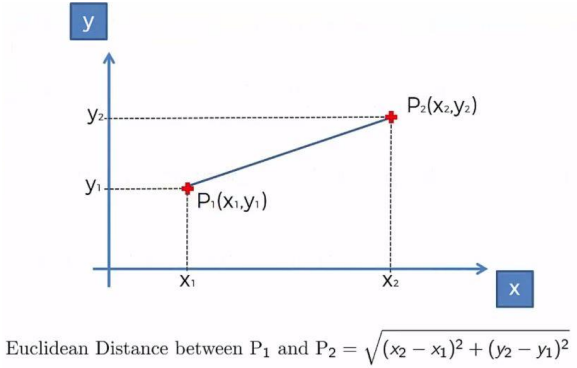
collective
forming a whole or aggregate

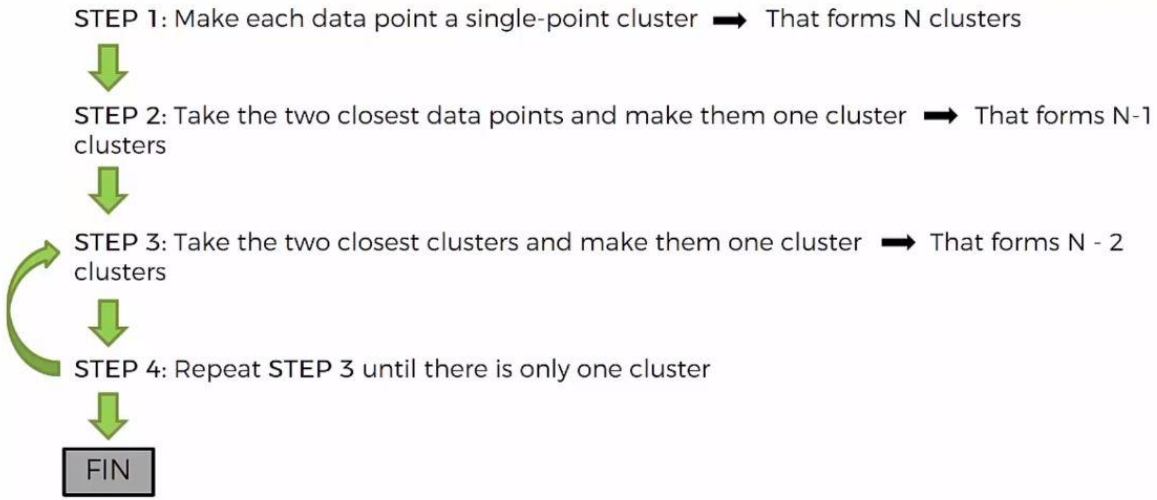


Hierarkisk Klustring -

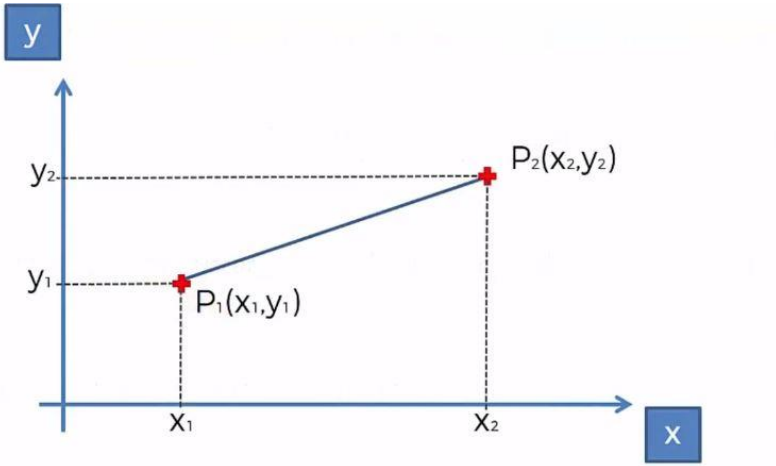
Börja med att betrakta varje enskild datapunkt som ett enskilt kluster.

Klumpa efter hand ihop kluster baserade på minsta avstånd.
Euklidiskt avstånd.



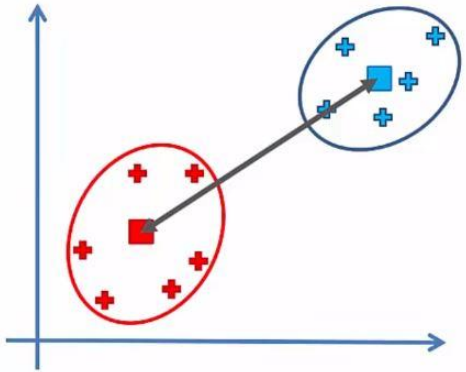


Hierarkisk Klustring - Avstånd



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

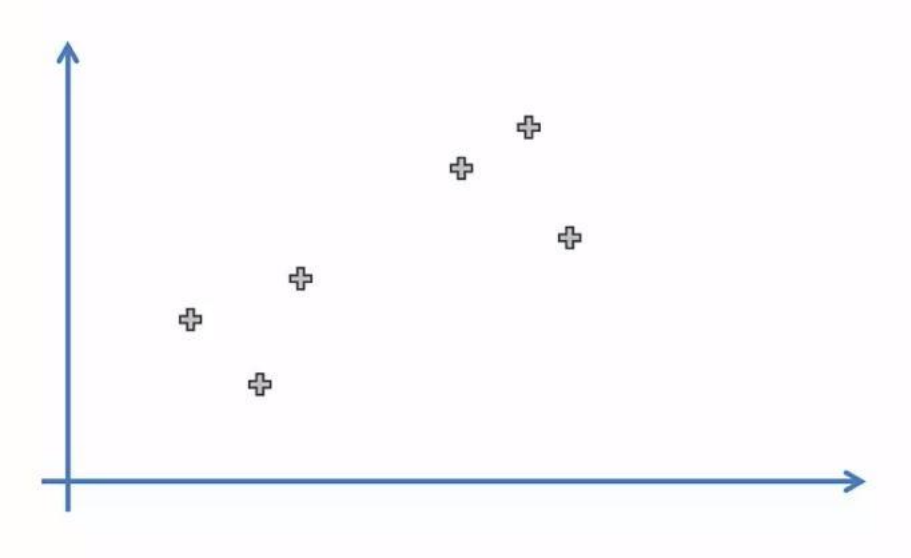
Hierarkisk Klustring - Avstånd



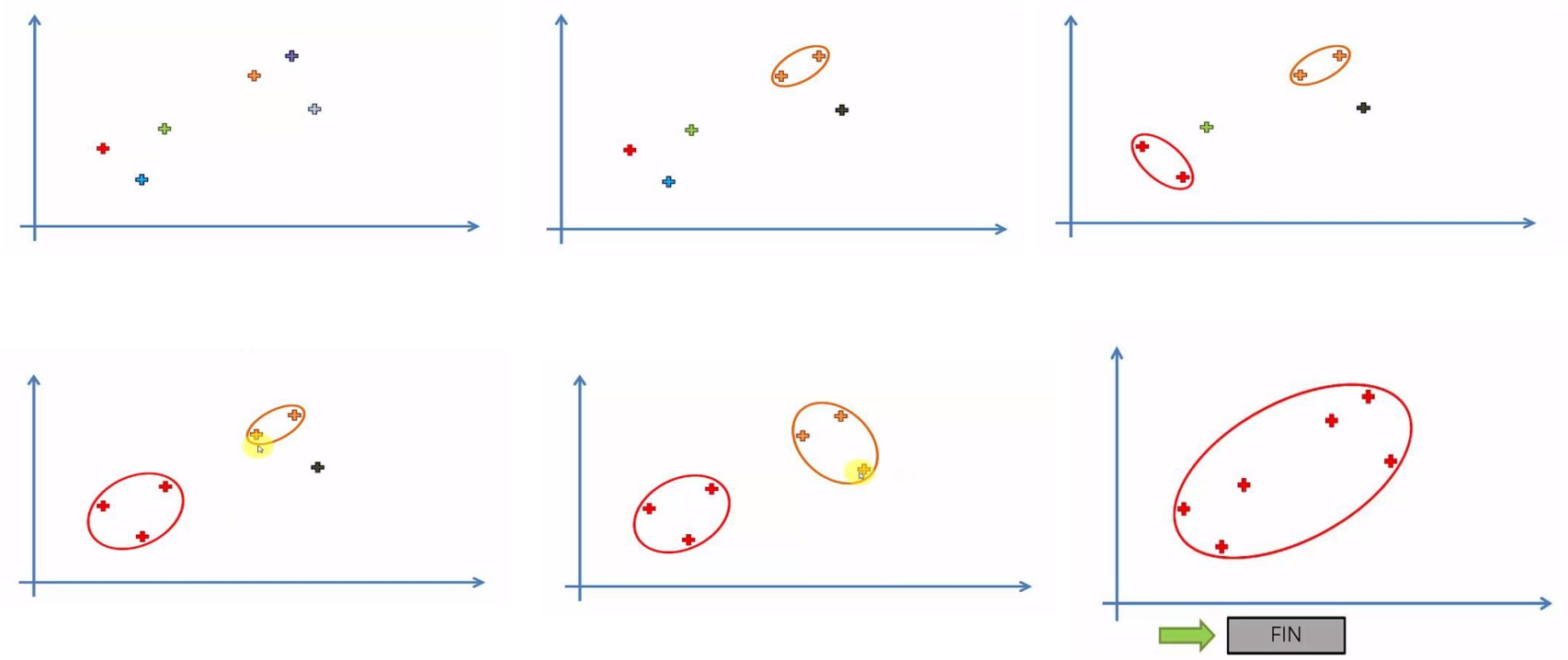
Distance Between Two Clusters:

- Option 1: Closest Points
- Option 2: Furthest Points
- Option 3: Average Distance
- Option 4: Distance Between Centroids

Hierarkisk Klustring - Exempel

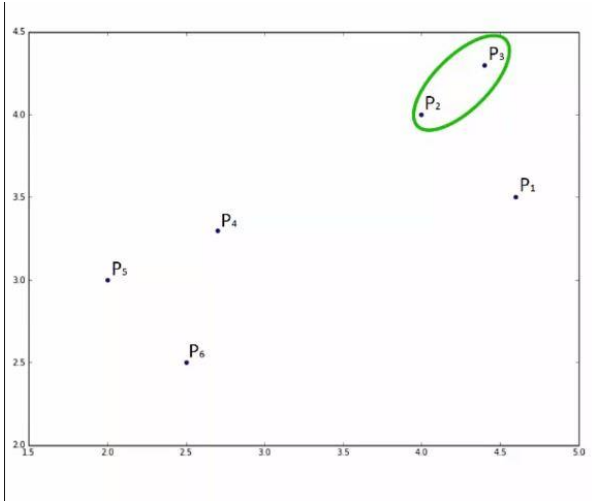


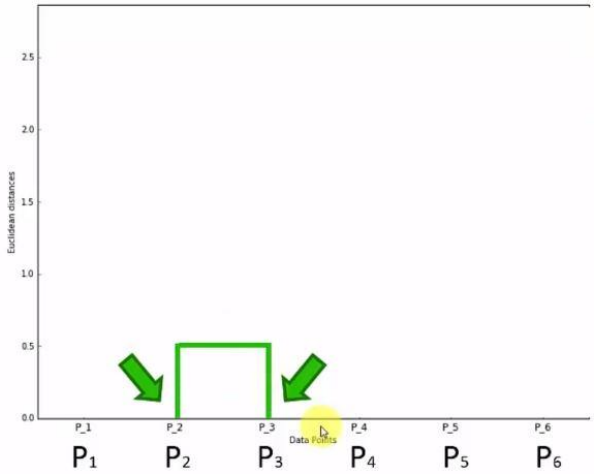
Hierarkisk Klustring - Exempel



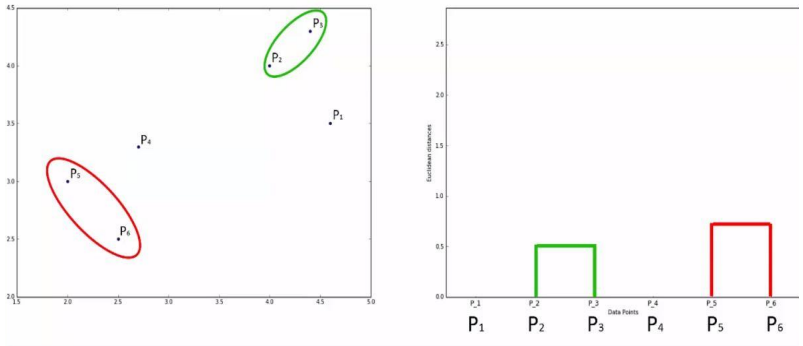
Hierarkisk Klustring - Dendrogram

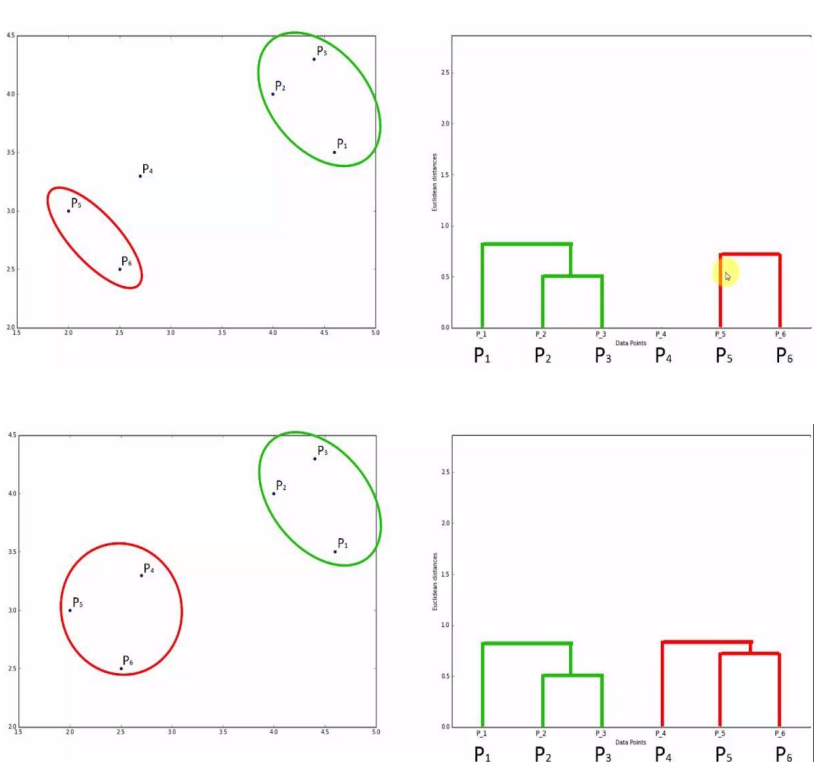
Sammanbind de två punkterna närmast till ett kluster. Markera avståndet som höjd i dendrogramet.





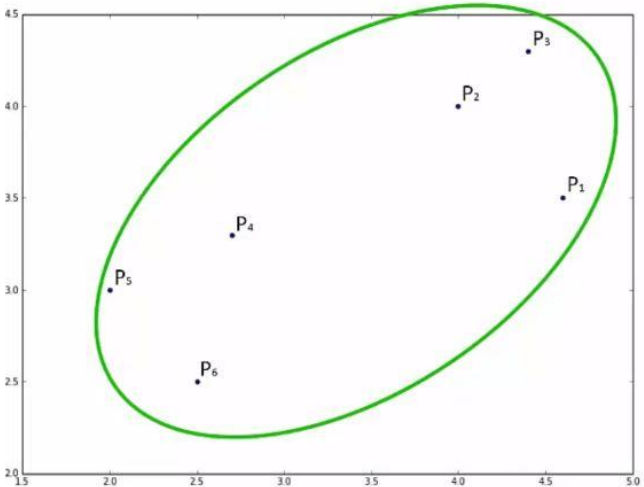
Hierarkisk Klustring - Dendrogram

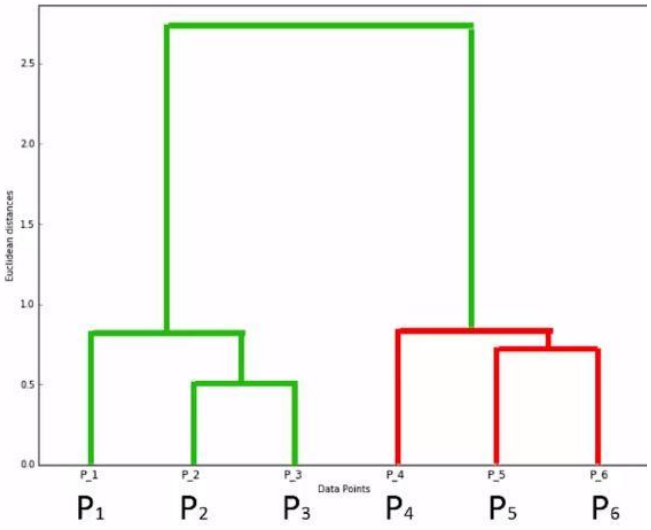




Hierarkisk Klustring - Dendogram

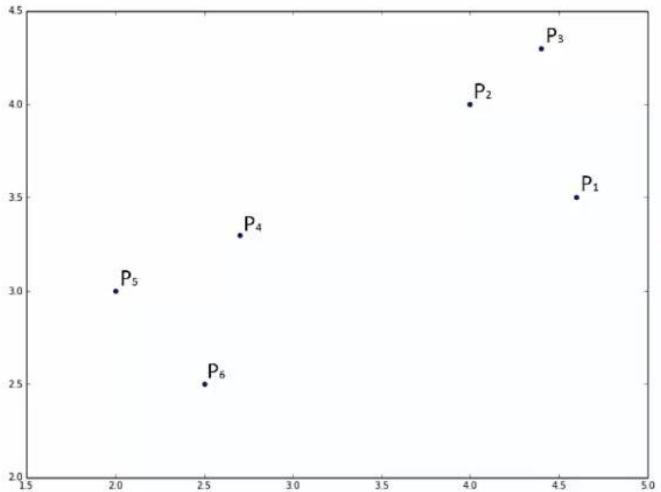
Till slut tillhör alla punkter ett kluster

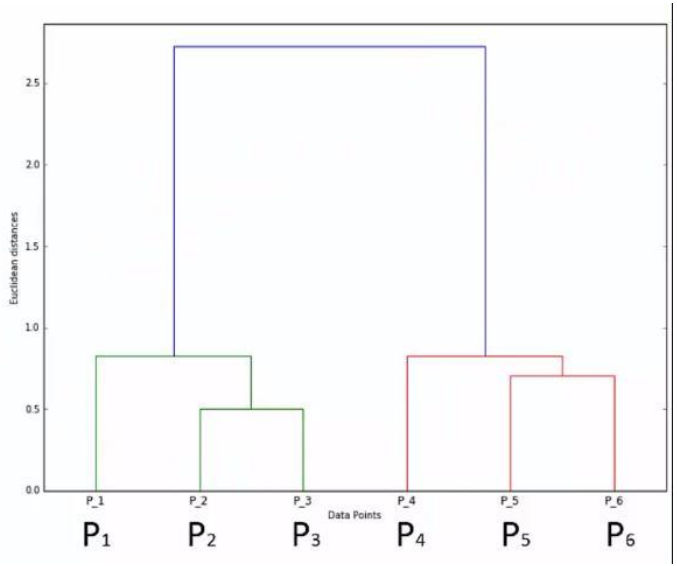




Hierarkisk Klustring - Dendogram

Ritat med plot algoritm

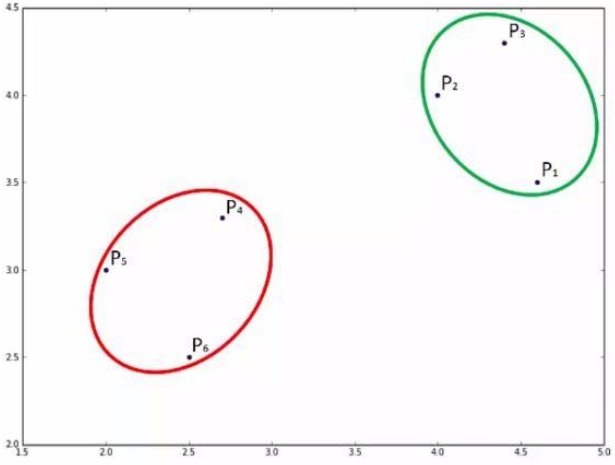


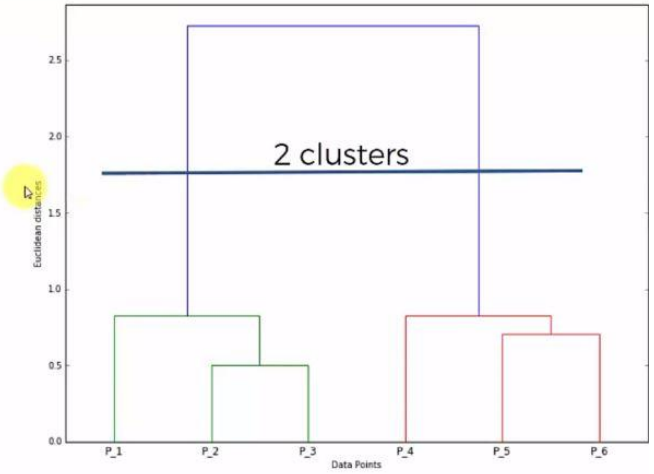


RESTORY...

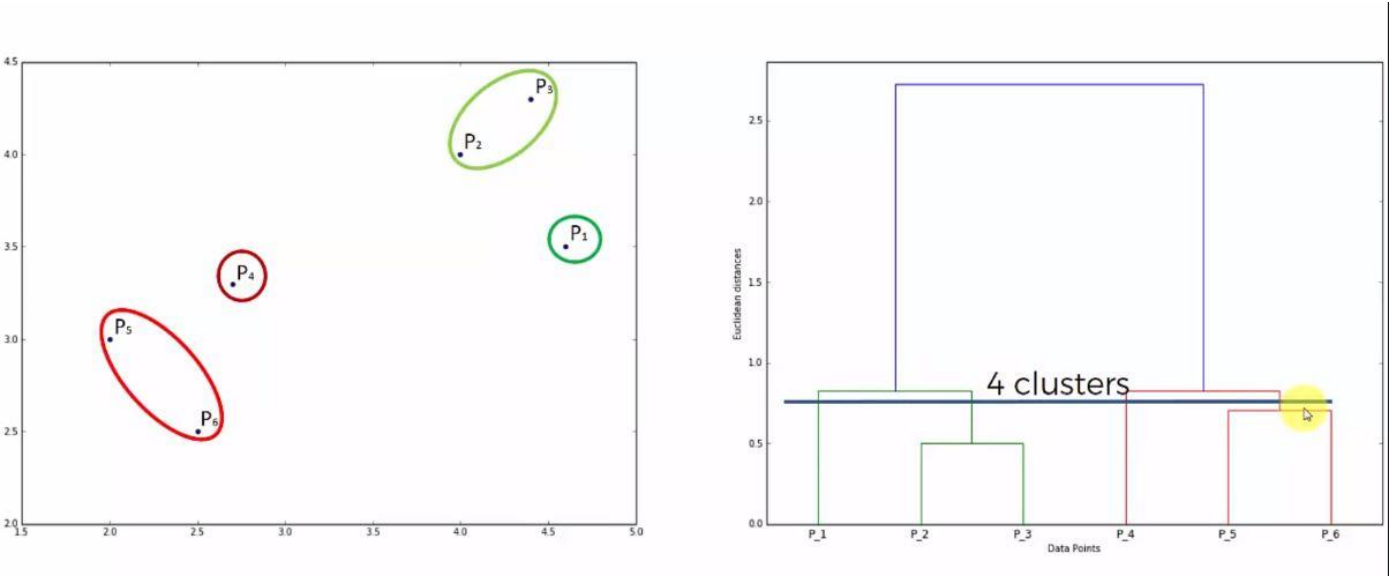
Hierarkisk Klustring - Dendogram

Drar man en vågrät linje skär den genom lika många lodräta linjer som nivån motsvarar.



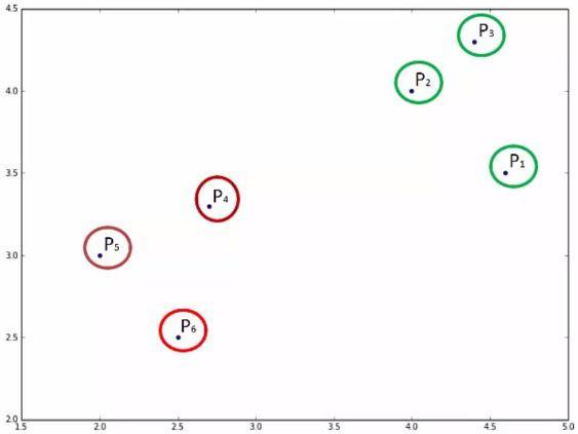


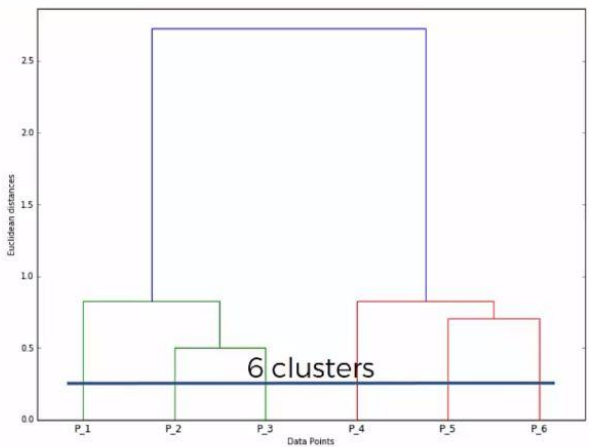
Hierarkisk Klustring - Dendogram



Hierarkisk Klustring - Dendogram

Drar man en vågrät linje längst ner motsvarar det att varje punkt är ett kluster.

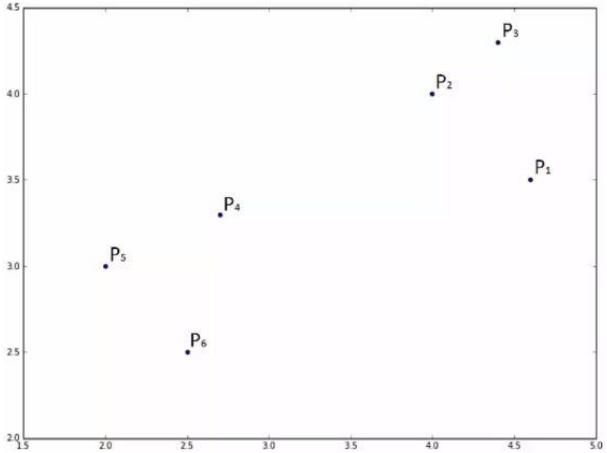


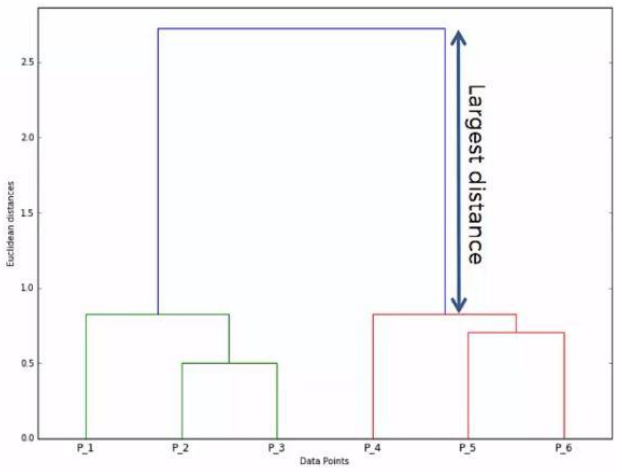


RESTORY...

Hierarkisk Klustring - Dendogram

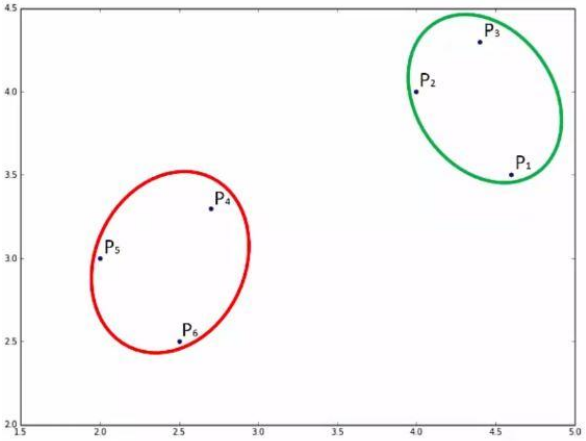
Den mest effektiva indelningen uppnås
där nivåerna är som mest skilda.

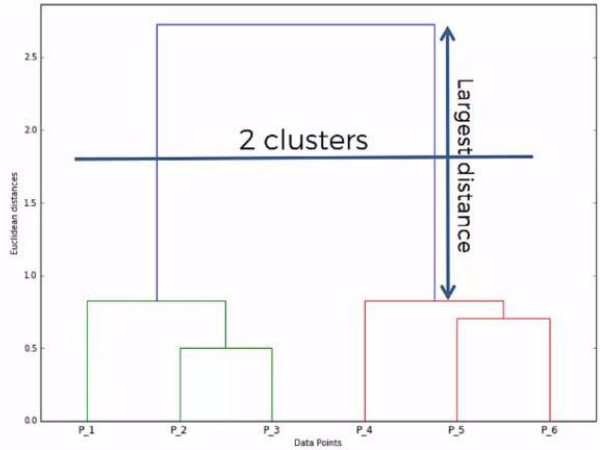




RESTORY...

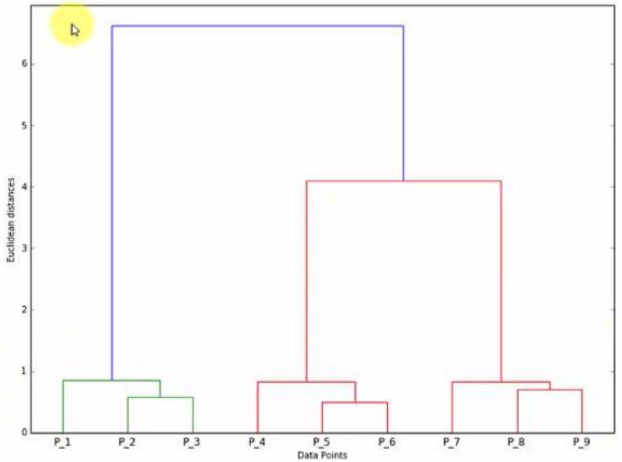
Hierarkisk Klustring - Dendogram





Hierarkisk Klustring - Dendogram

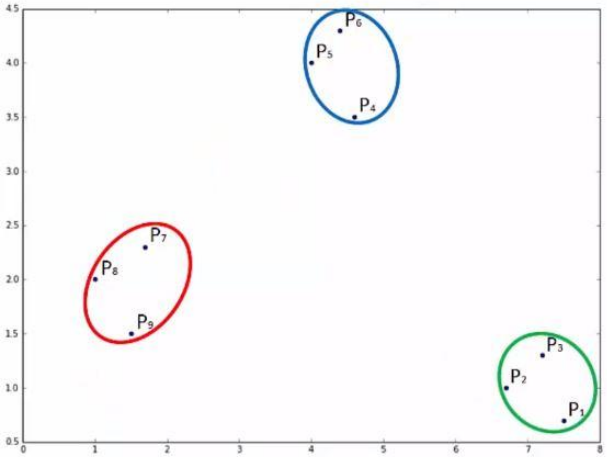
Vilken klusterindelning blir bäst enligt dendrogramet?

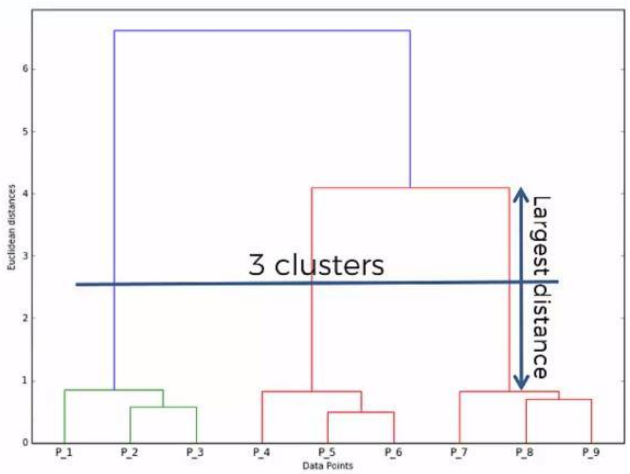


Hierarkisk Klustring - Dendogram

Vilken klusterindelning blir bäst enligt dendrogramet?

3 st kluster





Sammanfattning

Länkar