



Klustering



RESTOR\...

Agenda

- Klustering - K-means
- Hierarkisk klustering

Kod

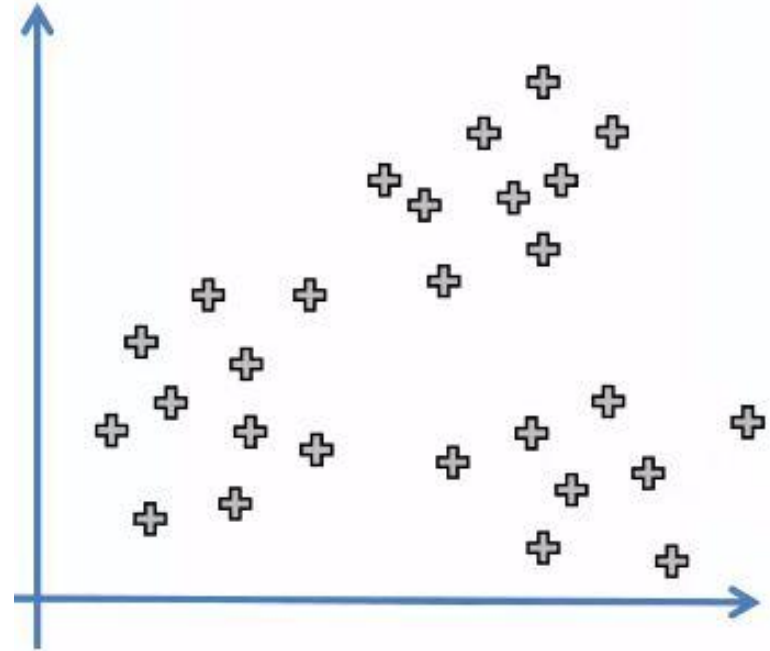
- Klustering

Klustering - Grupper

Till skillnad från *supervised* ML - inga "svar"

Hur grupperar man

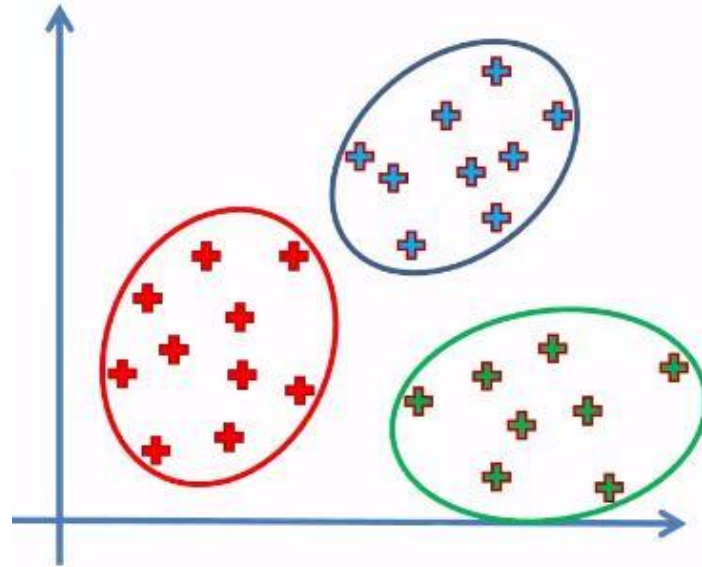
How to cluster



RESTOR\...

Klustring -

Intuitivt: "Klumpar" som hör ihop



Klustering - Algoritm

Princip för K-Means algoritmen

STEP 1: Choose the number K of clusters



STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



STEP 3: Assign each data point to the closest centroid → That forms K clusters



STEP 4: Compute and place the new centroid of each cluster



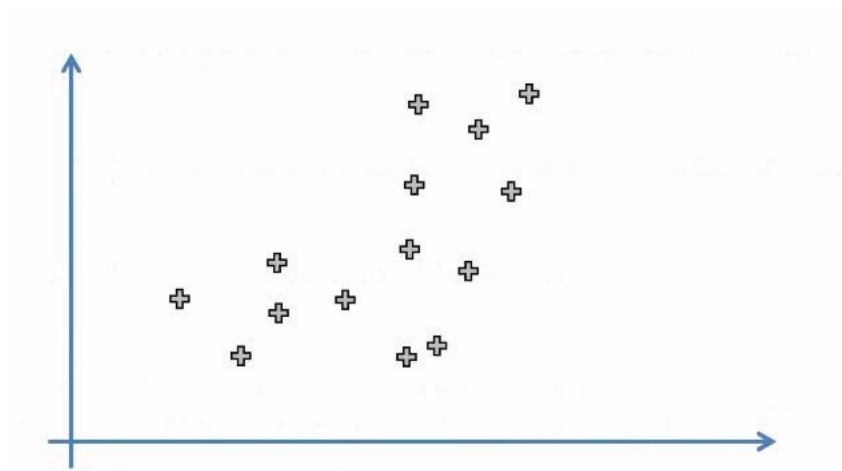
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



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

Klustering - Algoritm Exempel

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

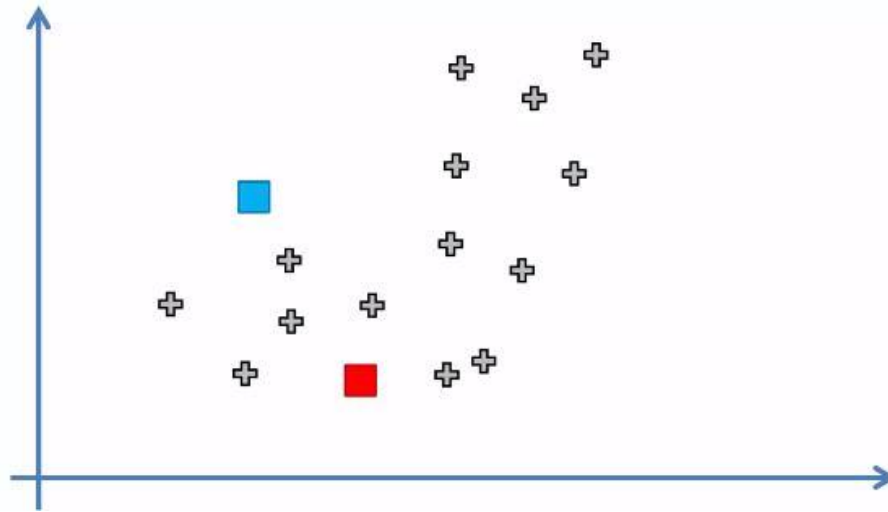


Klustering - Algoritm steg 1&2

Steg 1: Välj antal kluster: 2

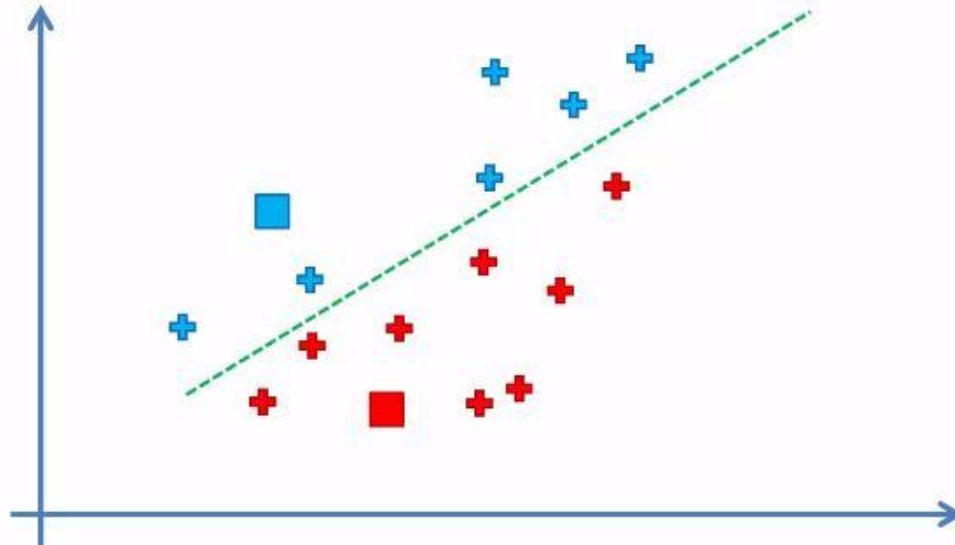
Steg 2: Placera ut 2 (antal kluster) punkter slumpmässigt.

Dessa kallas centroider.



Klustering - Algoritm steg 3

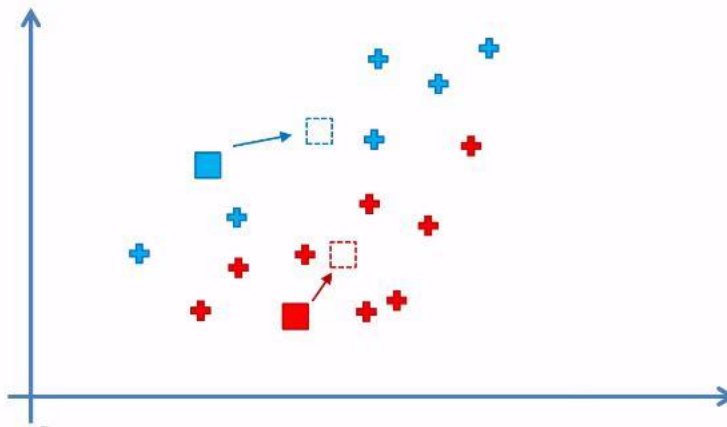
Steg 3: Tilldela nu varje datapunkt den klass som centroiden har som ligger närmast



RESTOR\...

Klustring - Algoritm steg 4

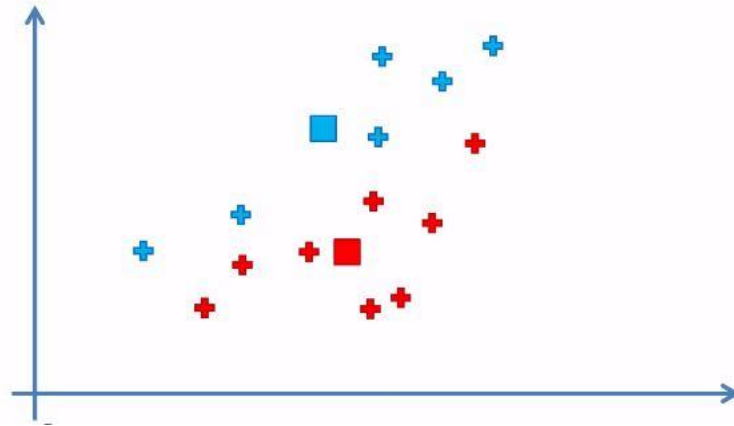
Steg 4: Flytta nu centroiden för varje kluster till respektive tyngdpunkt.



Klustering - Steg 3

Upprepa steg 3.

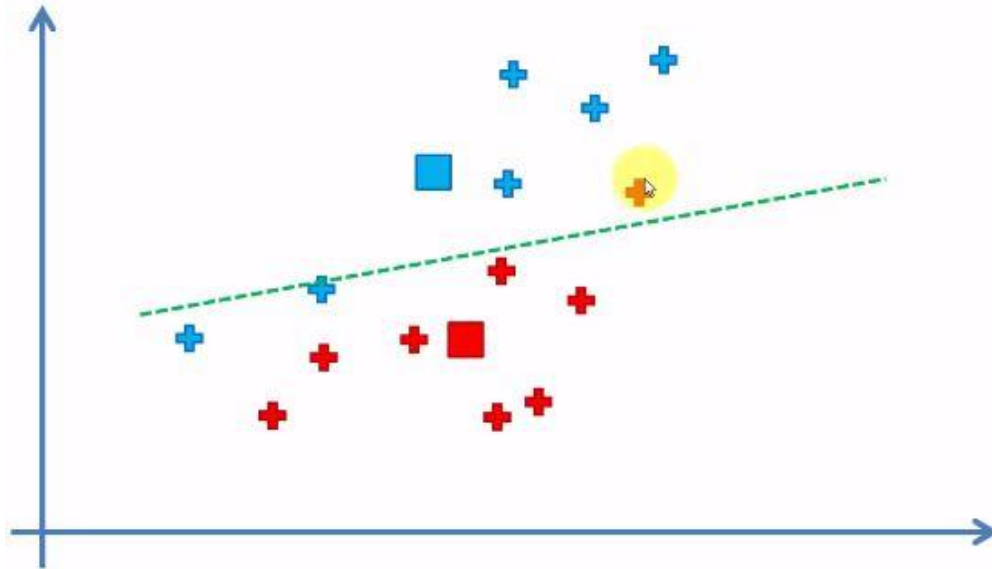
När centroiderna är flyttade, förberedd för att åter tilldela varje punkt en klass.



RESTOR\...

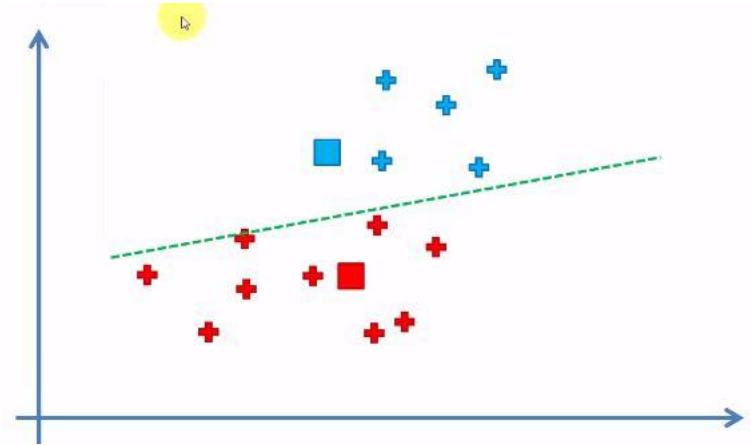
Klustring - Steg 3

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



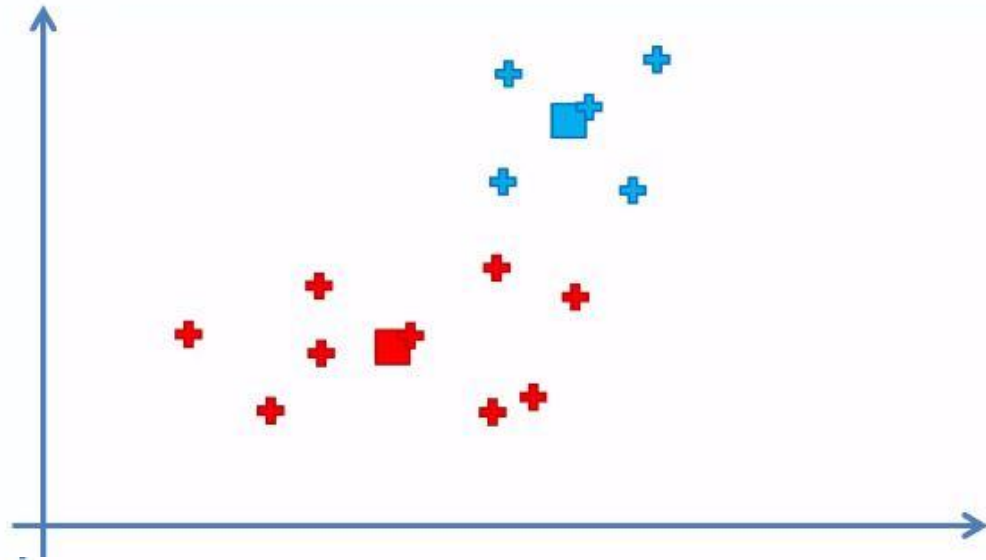
Klustring - Algoritm steg 5

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



Klustering - Algoritma step 4

Repetera step 4

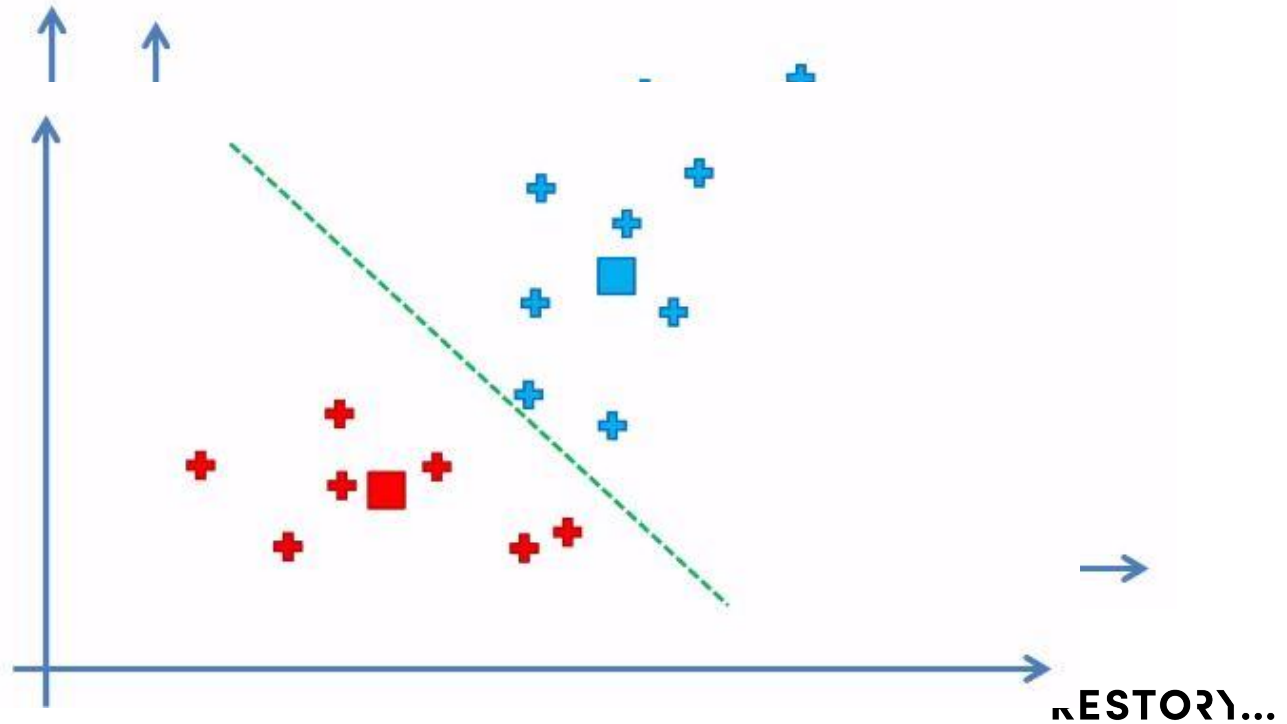


RESTART...

Klustering - Algorithm step 5 & 4

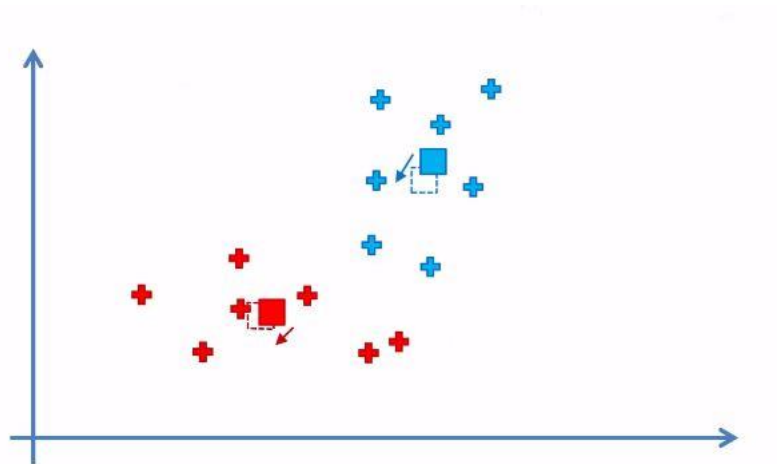
Repetera step 5

Repetera step 4



Klustring - Algoritm

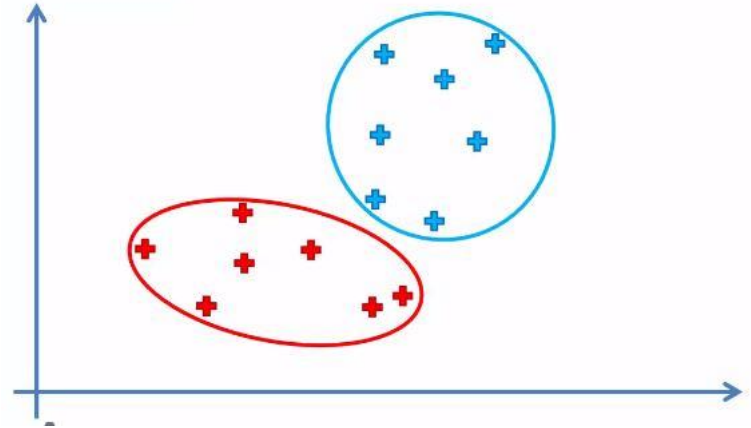
När ingen punkt längre ändrar klass
mellan steg 4 och steg 5 är modellen
färdig!



RESTOR\...

Klustering - Klar

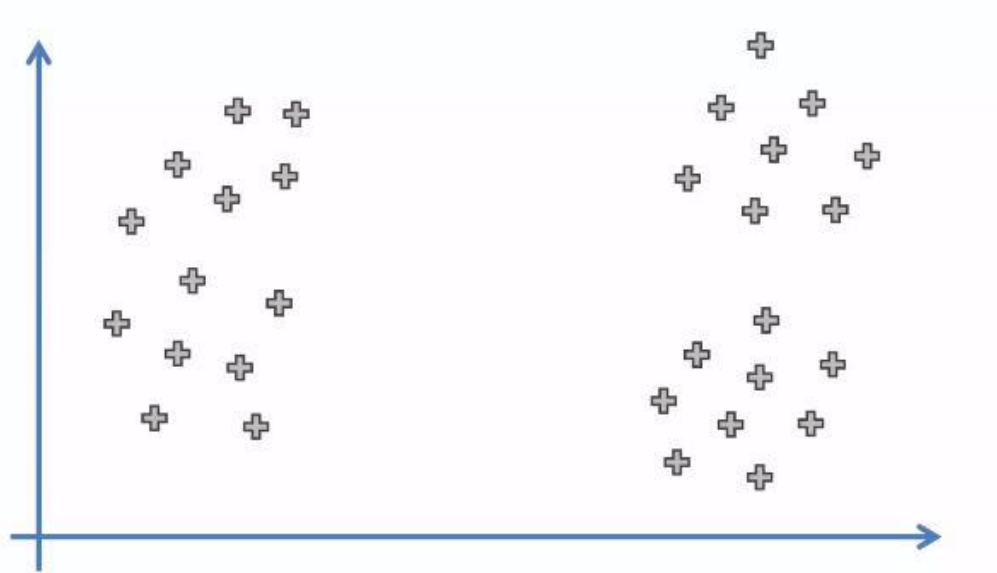
Slutgiltig klustering



RESTOR\...

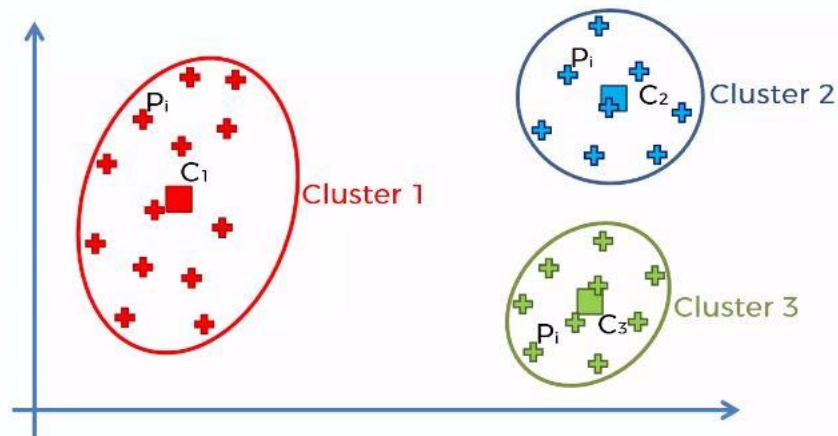
Klustring - Val av K

Hur väljer man bäst antal kluster K?



Klustring - Val av K

Vi behöver ett mått som beskriver klustringens effektivitet!



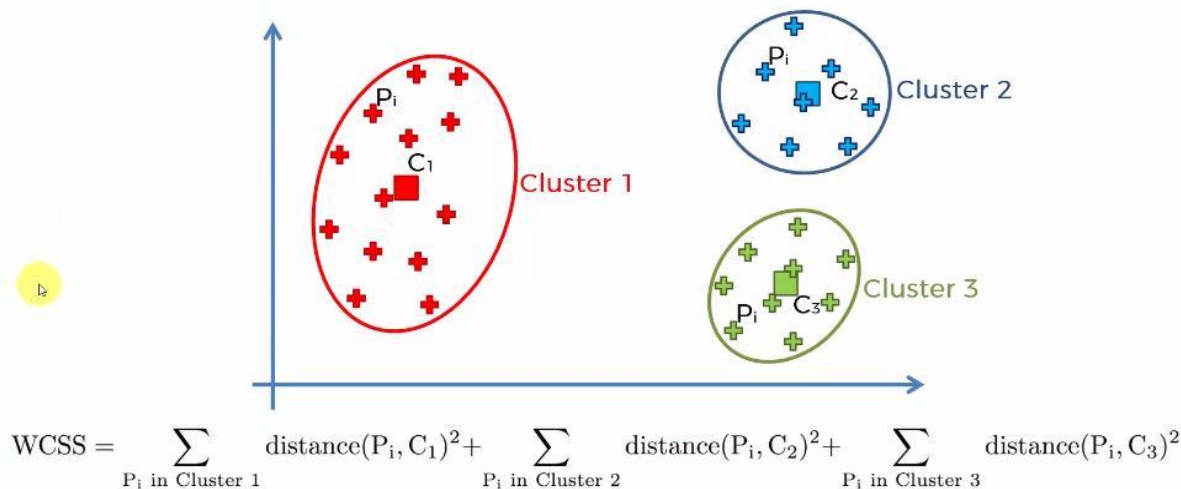
Klustering - Val av K

WCSS - Within Cluster Sum of Squares

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

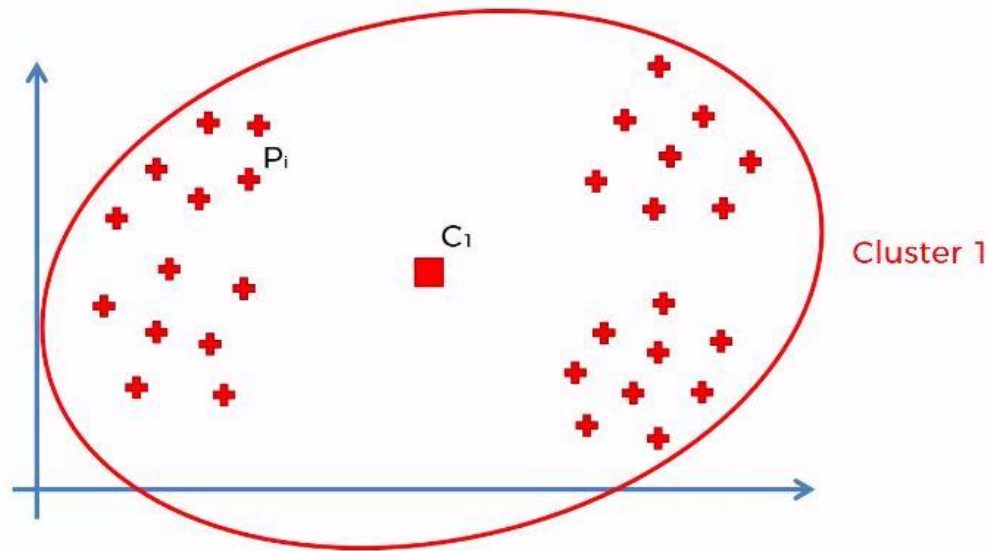
Klustering - Val av K

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



Klustering - Val av K

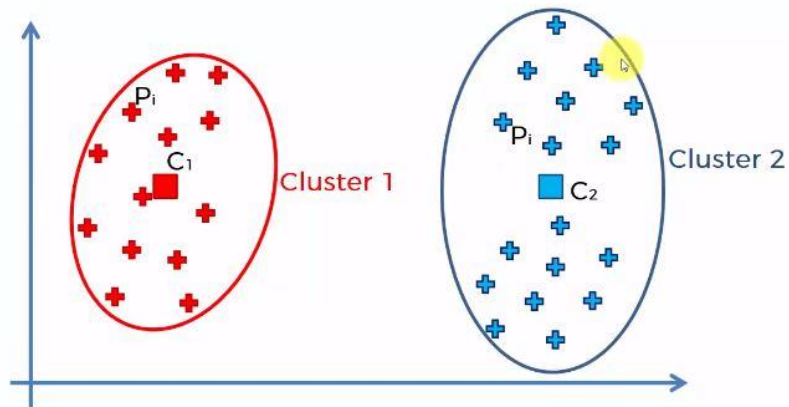
Om vi bara har ett kluster?



$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$$

Klustering - Val av K

Om vi har två kluster



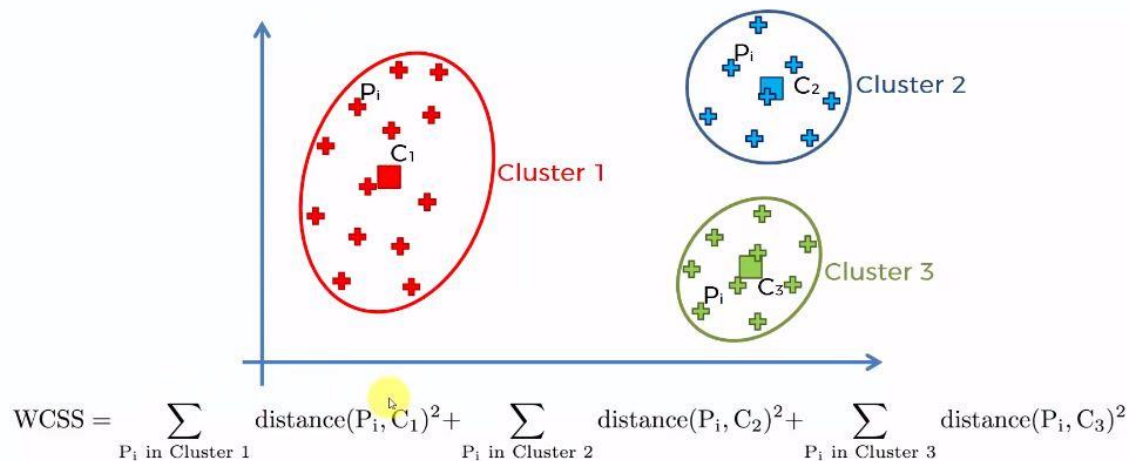
$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

Klustring - Val av K

Om vi har tre kluster?

Vad händer med WCSS när K växer?

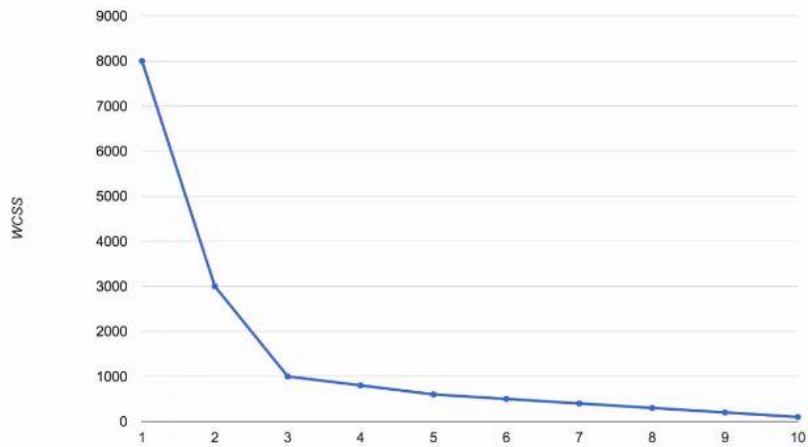
Hur litet kan WCSS bli?



Klustering - Val av K

Armbågsmetoden/ Elbow method

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



Hierarkisk klustering

Hierarkisk Klustering - Agglomerativa metoden

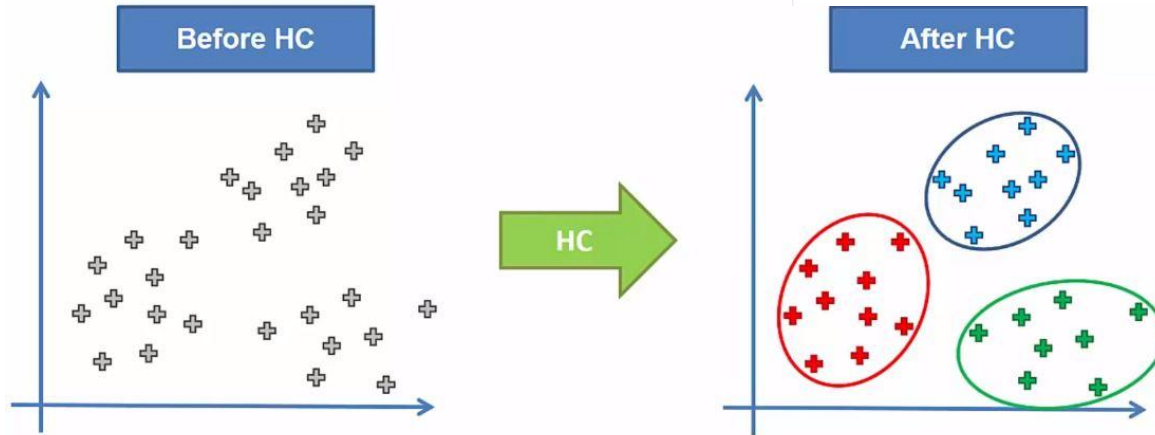
Definitions of *agglomerative*

1. adjective clustered together but not coherent

synonyms: agglomerate, agglomerated, clustered

collective

forming a whole or aggregate

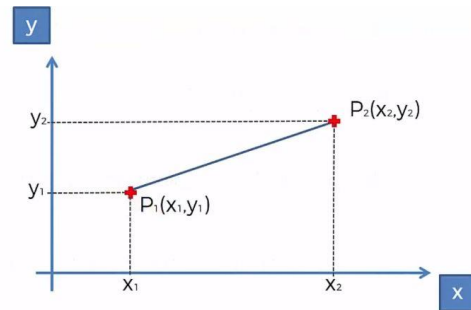


RESTORY...

Hierarkisk Klustring -

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

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



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

STEP 1: Make each data point a single-point cluster ➡ That forms N clusters



STEP 2: Take the two closest data points and make them one cluster ➡ That forms N-1 clusters



STEP 3: Take the two closest clusters and make them one cluster ➡ That forms N - 2 clusters

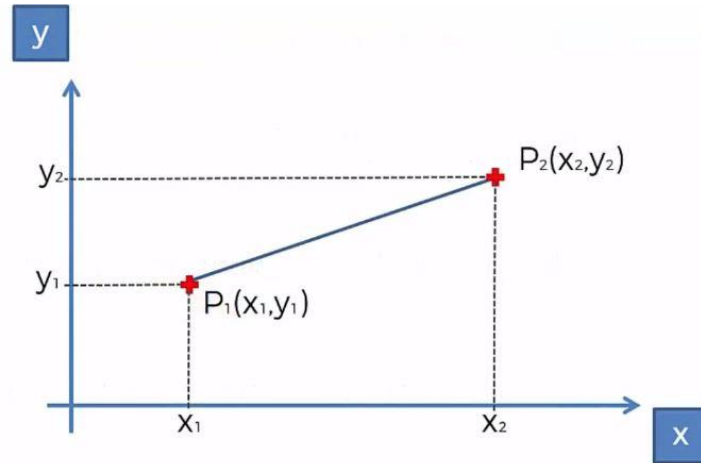


STEP 4: Repeat STEP 3 until there is only one cluster



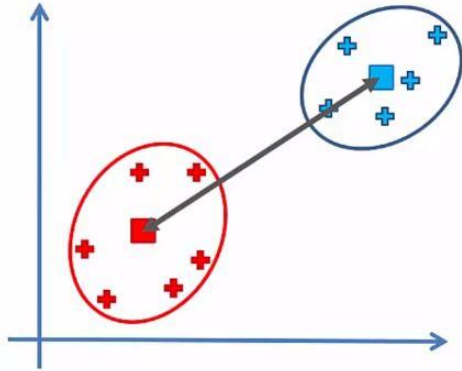
FIN

Hierarkisk Klustering - Avstånd



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

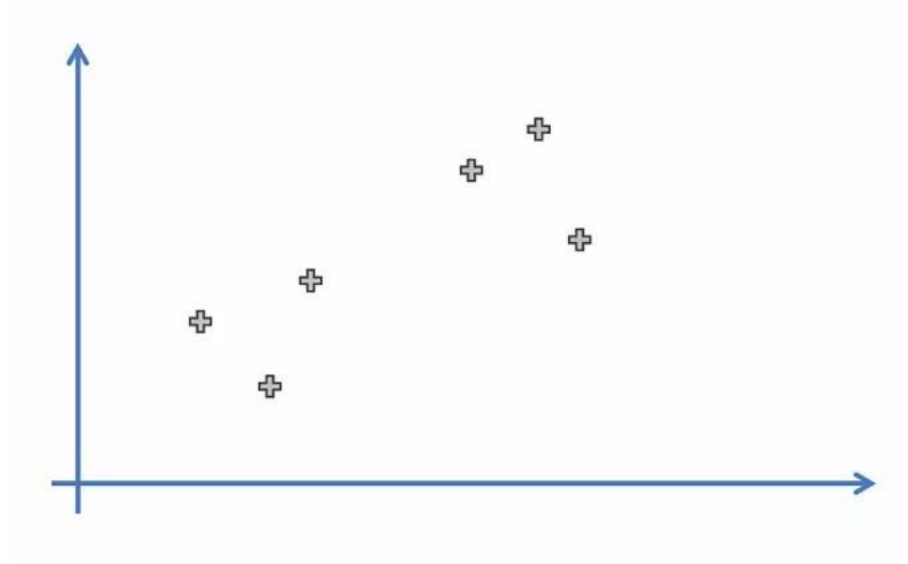
Hierarkisk Klustering - Avstånd



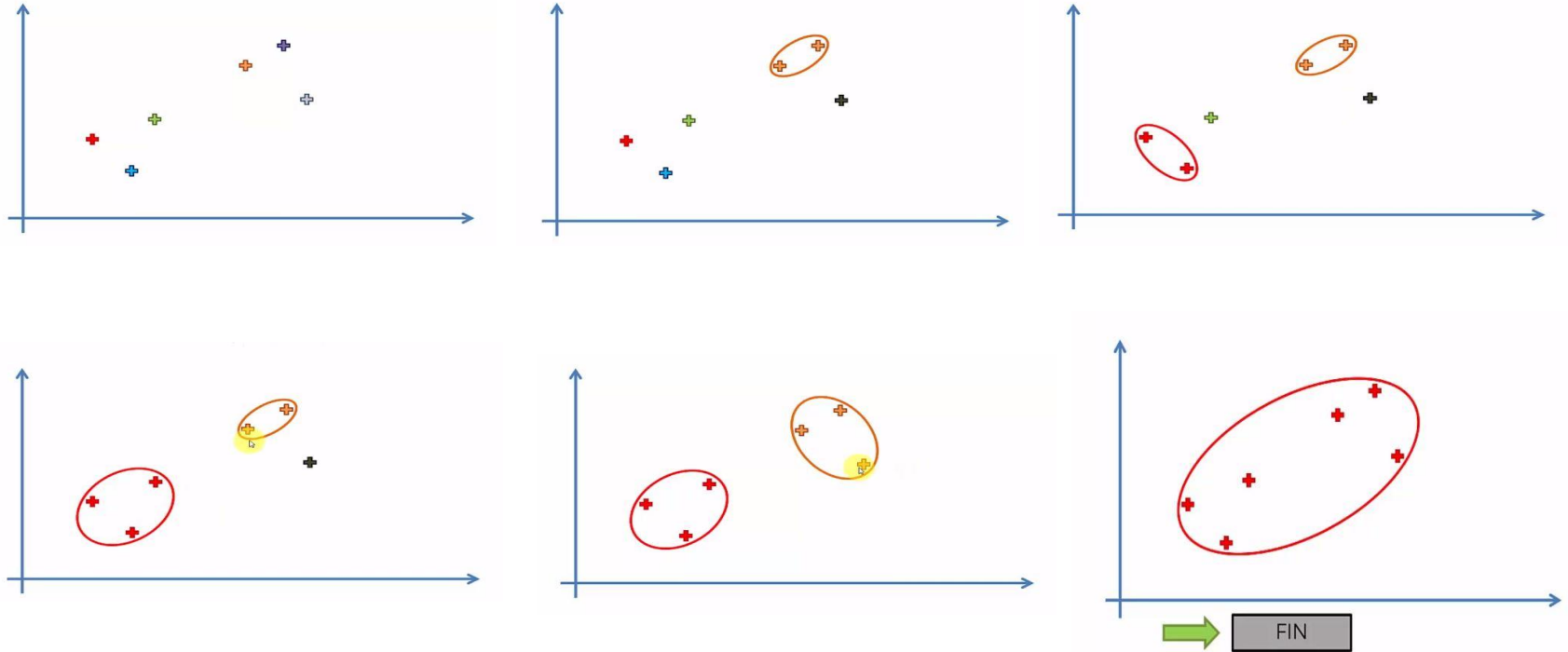
Distance Between Two Clusters:

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

Hierarkisk Klustering - Exempel



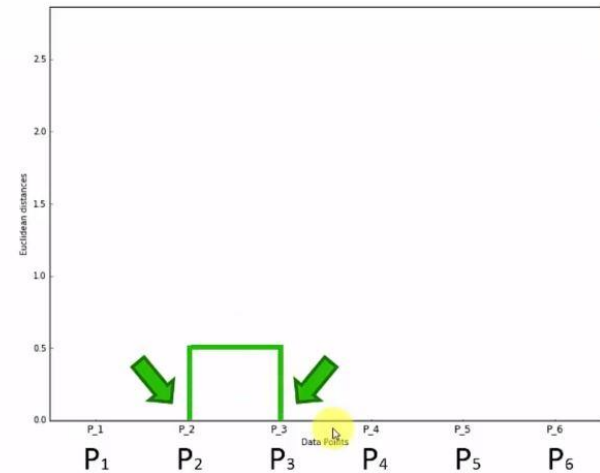
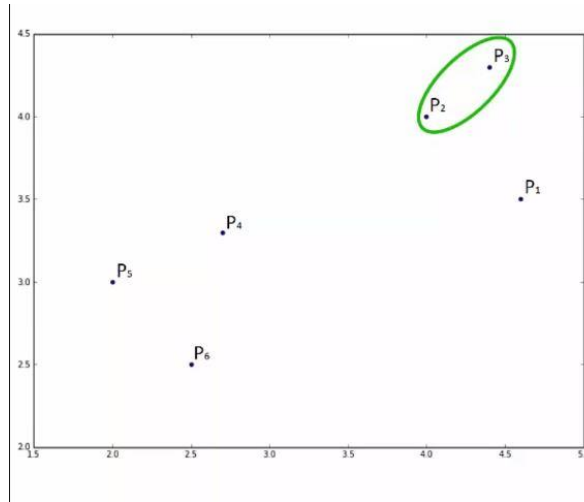
Hierarkisk Klustring - Exempel



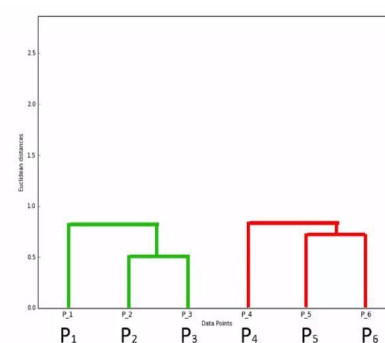
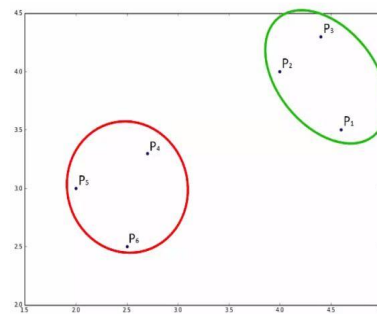
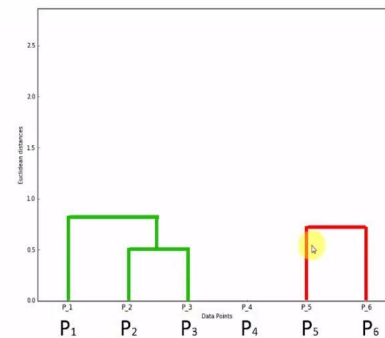
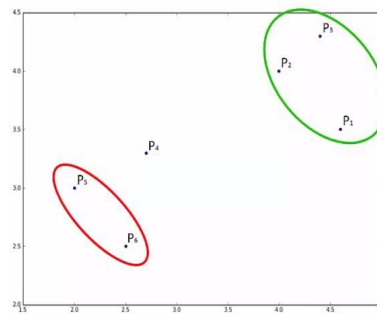
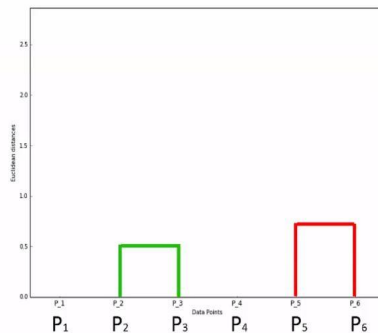
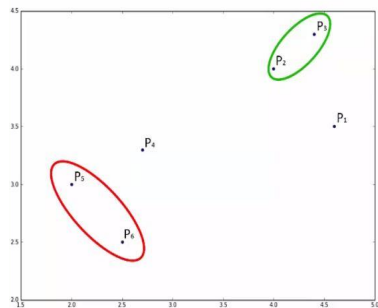
RESTRÖY...

Hierarkisk Klustring - Dendrogram

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



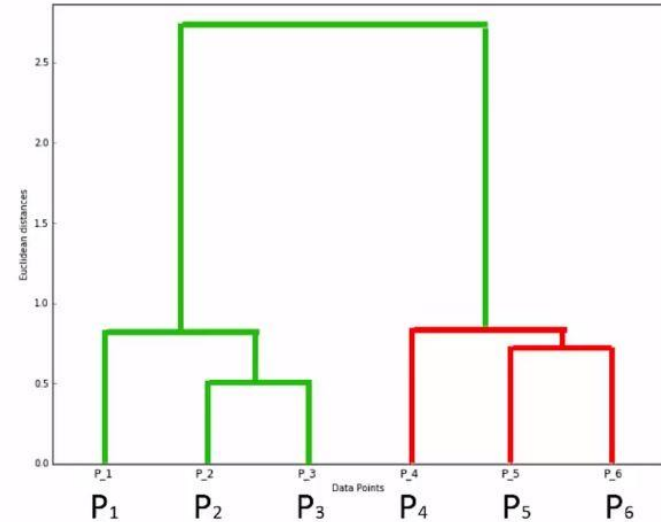
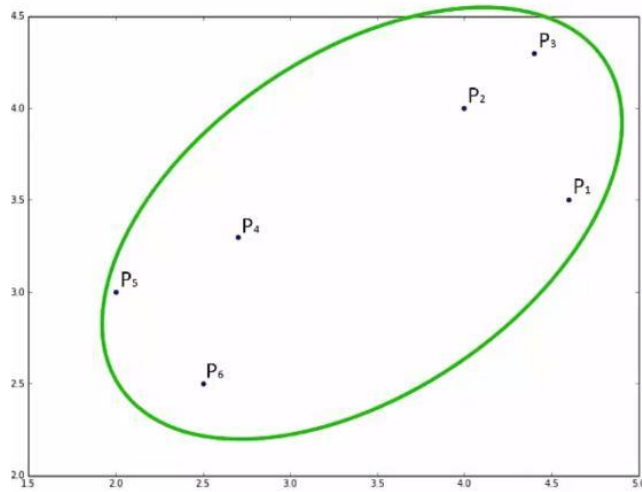
Hierarkisk Klustering - Dendrogram



RESTRY...

Hierarkisk Klustring - Dendogram

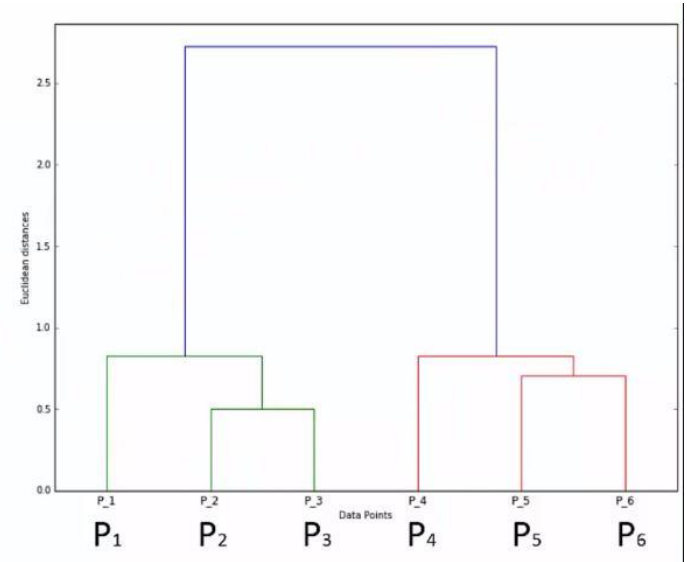
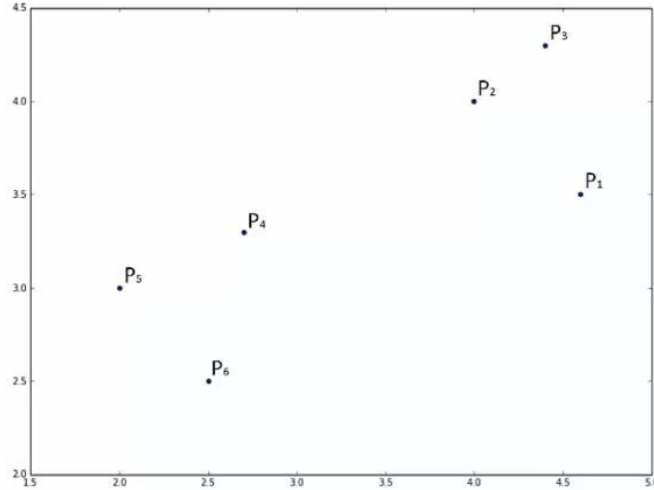
Till slut tillhör alla punkter ett kluster



RESTOR...

Hierarkisk Klustring - Dendrogram

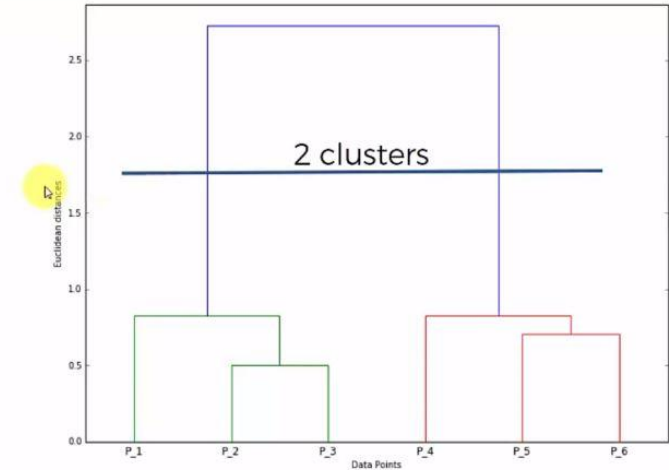
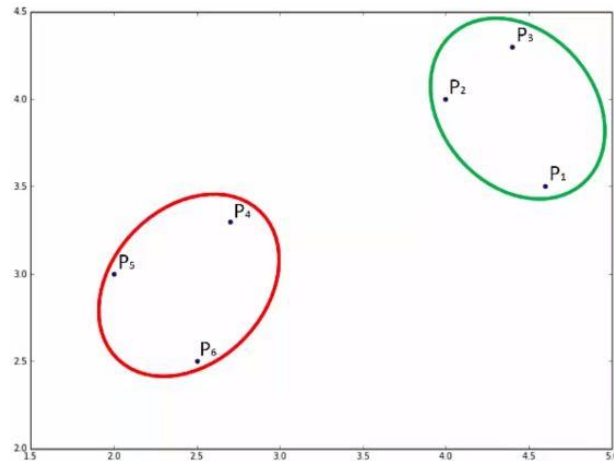
Ritat med plot algoritm



RESTORY...

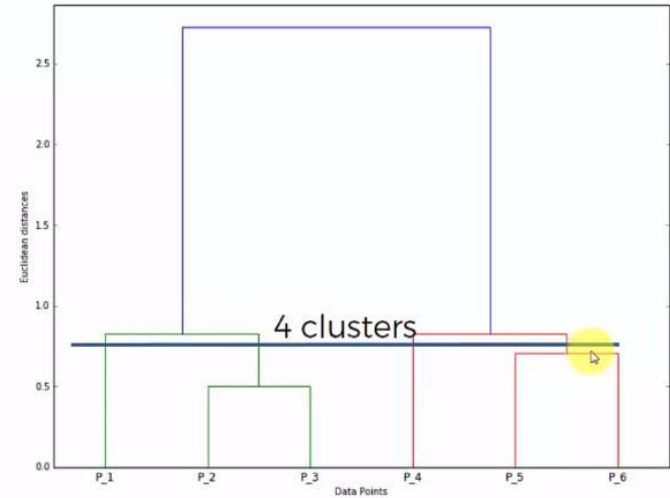
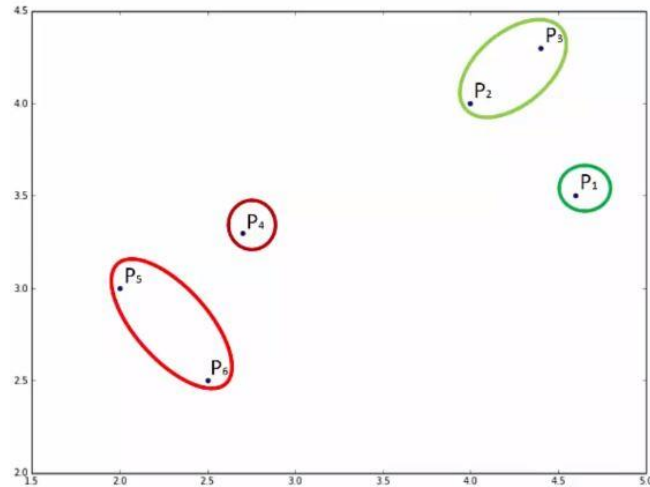
Hierarkisk Klustring - Dendogram

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



RESTRÖY...

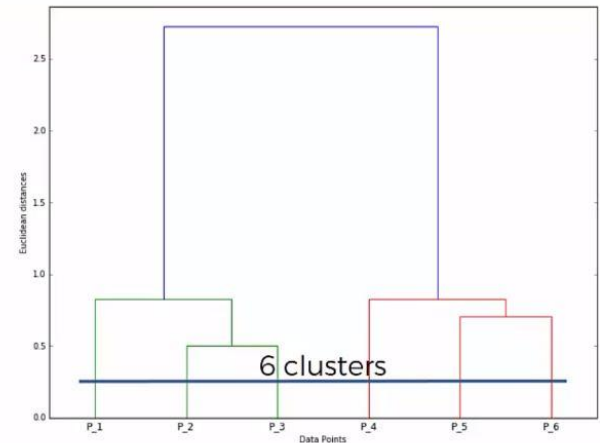
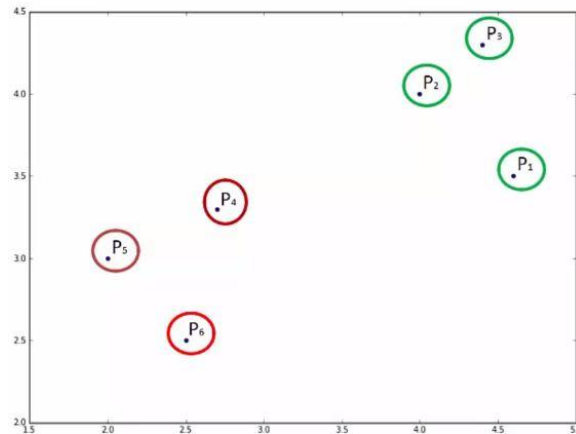
Hierarkisk Klustering - Dendogram



RESTORY...

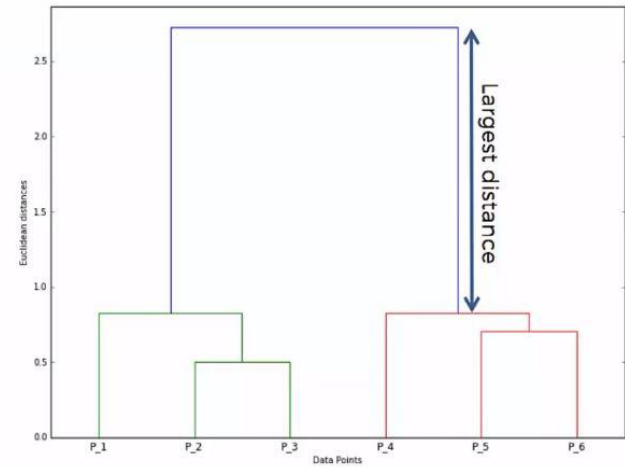
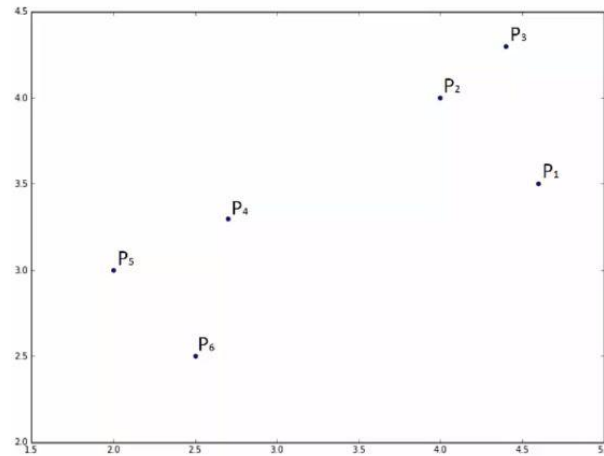
Hierarkisk Klustring - Dendogram

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



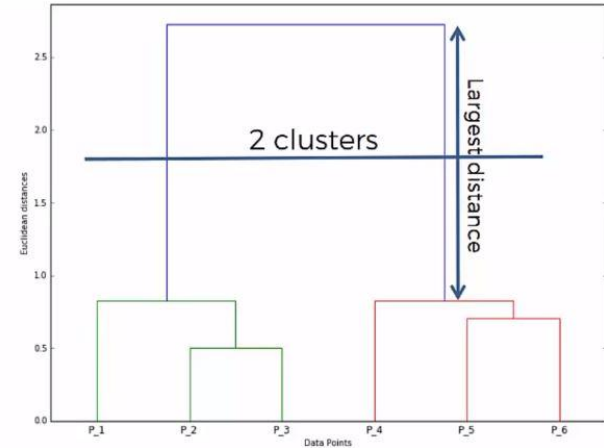
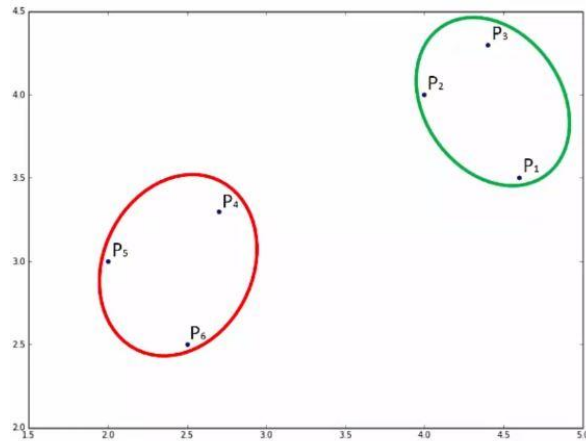
Hierarkisk Klustring - Dendogram

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



RESTORY...

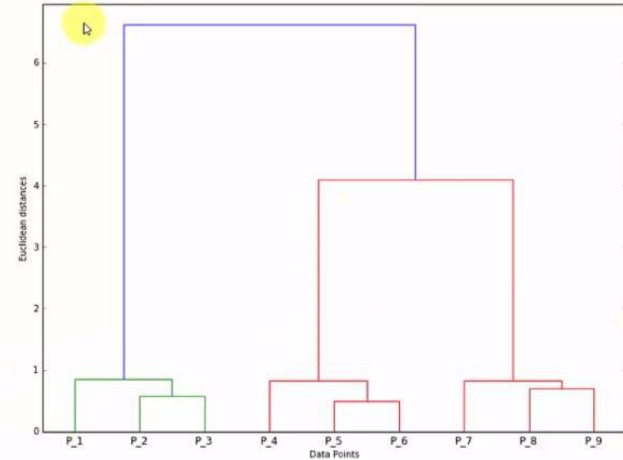
Hierarkisk Klustering - Dendogram



RESTORY...

Hierarkisk Klustring - Dendrogram

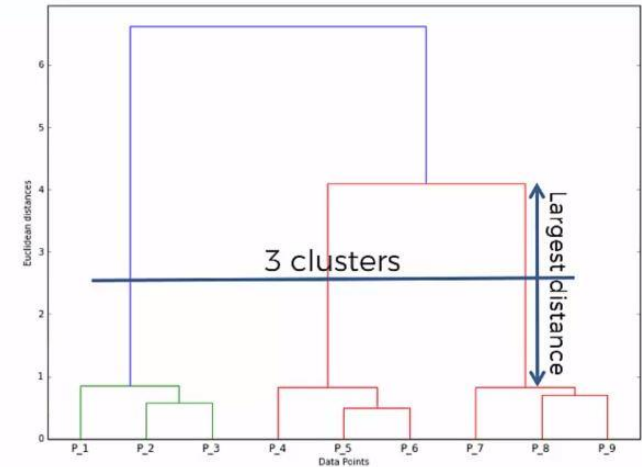
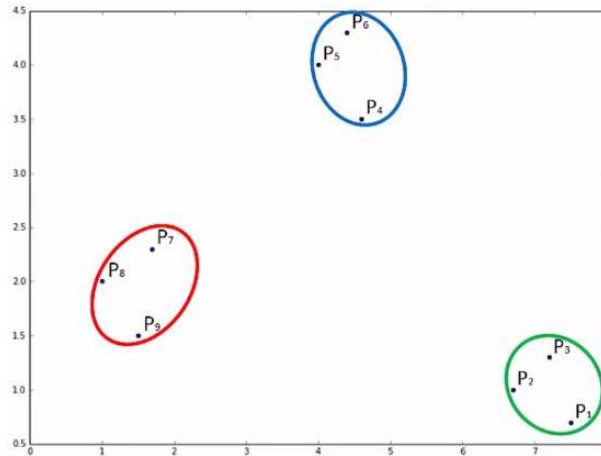
Vilken klusterindelning blir bäst enligt dendrogramet?



Hierarkisk Klustring - Dendrogram

Vilken klusterindelning blir bäst enligt dendrogramet?

3 st kluster



Sammanfattning

Länkar