Clustering Jerárquico

Clustering Jerárquico Aglomerativo

PASO 1: Hacer que cada punto sea un propio cluster. ➡ Así tendremos N clusters

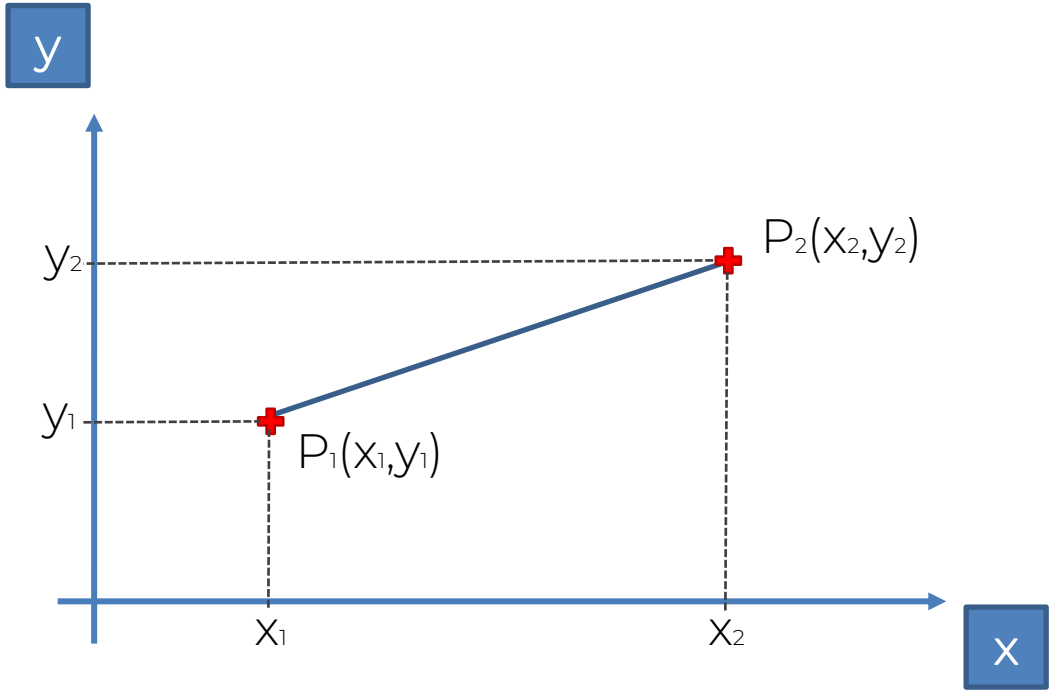
PASO 2: Elegir los dos puntos más cercanos y juntarlos en un único cluster ➡ N-1 clusters

PASO 3: Elegir los dos clusters más cercanos y juntarlos en un único cluster ➡ N - 2 clusters

PASO 4: Repetir el **PASO 3** hasta solo tener un único cluster

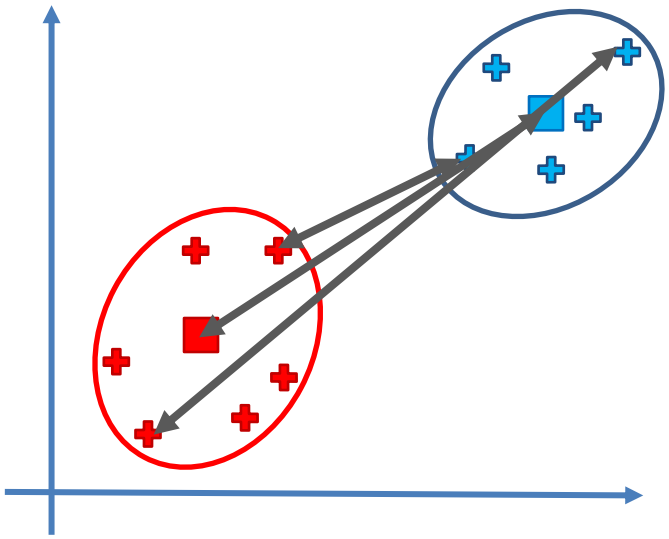
FIN

Distancia Euclídea



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

Distancia entre Clusters

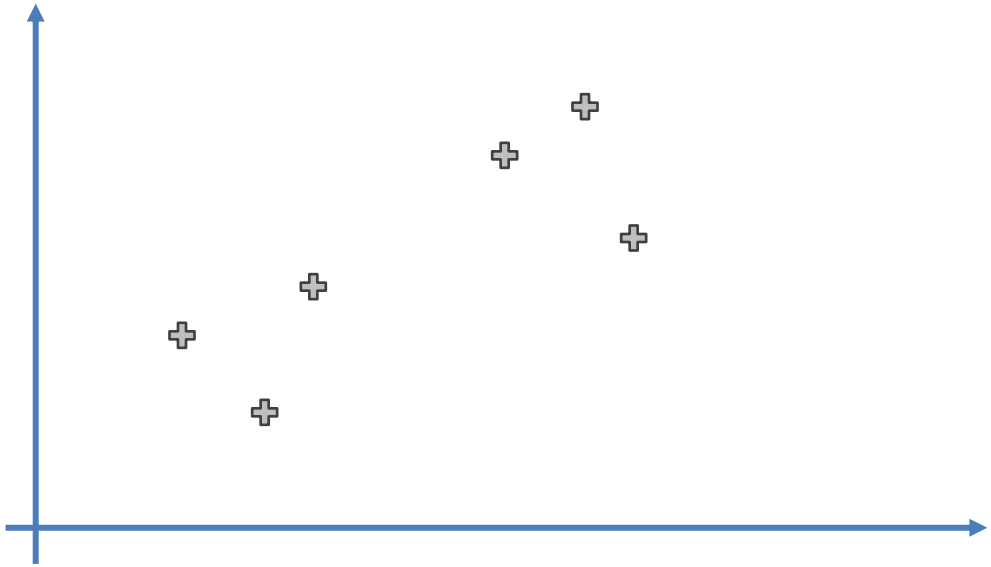


Distancia entre dos Clusters:

- Opción 1: Puntos más cercanos
- Opción 2: Puntos más alejados
- Opción 3: Distancia media
- Opción 4: Distancia entre sus baricentros

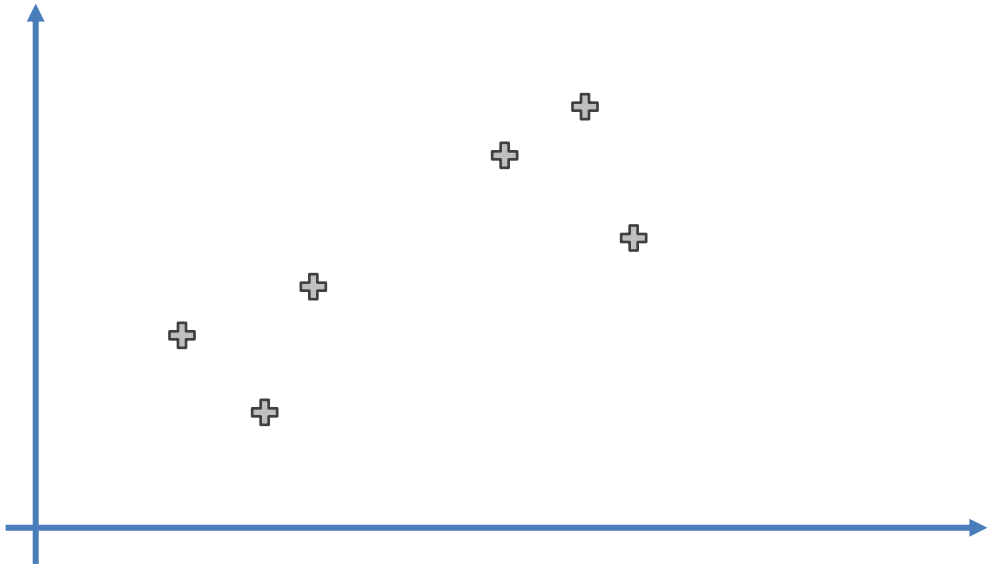
Clustering Jerárquico Aglomerativo

Consideremos el siguiente data set de N = 6 puntos



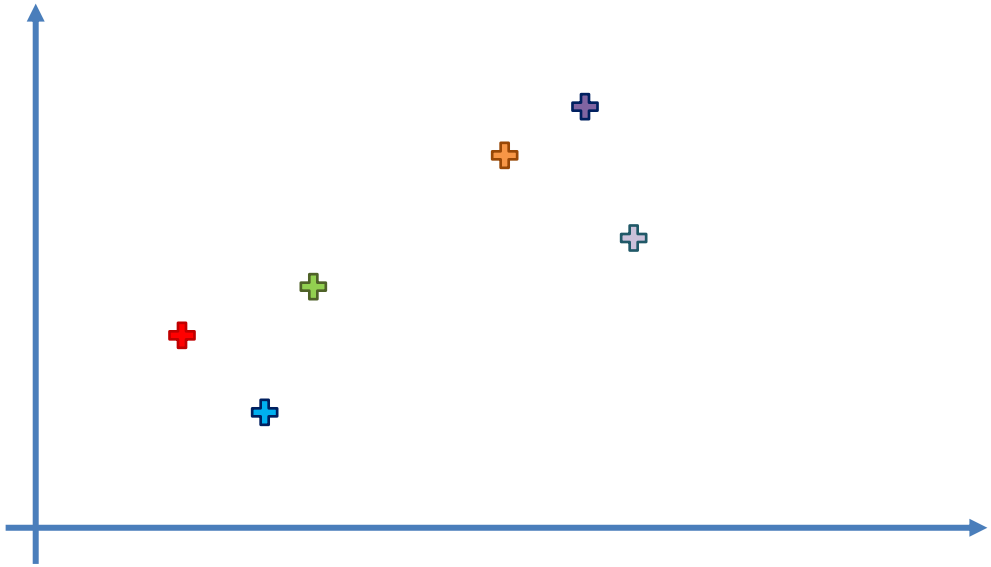
Clustering Jerárquico Aglomerativo

PASO 1: Hacer que cada punto sea un propio cluster. ➡ Así tendremos 6 clusters



Clustering Jerárquico Aglomerativo

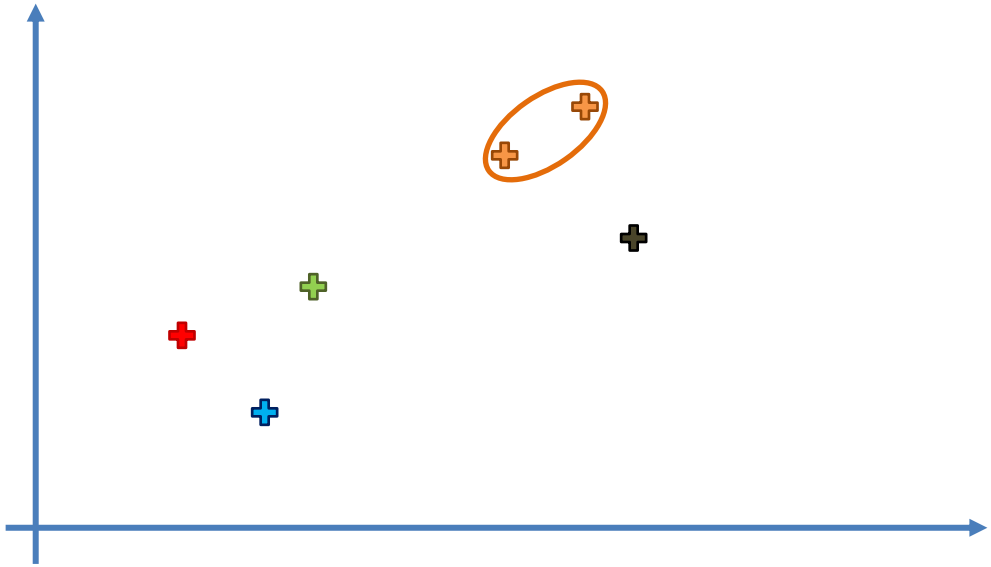
PASO 1: Hacer que cada punto sea un propio cluster. ➡ Así tendremos 6 clusters



Clustering Jerárquico Aglomerativo

PASO 2: Elegir los dos puntos más cercanos y juntarlos en un único cluster

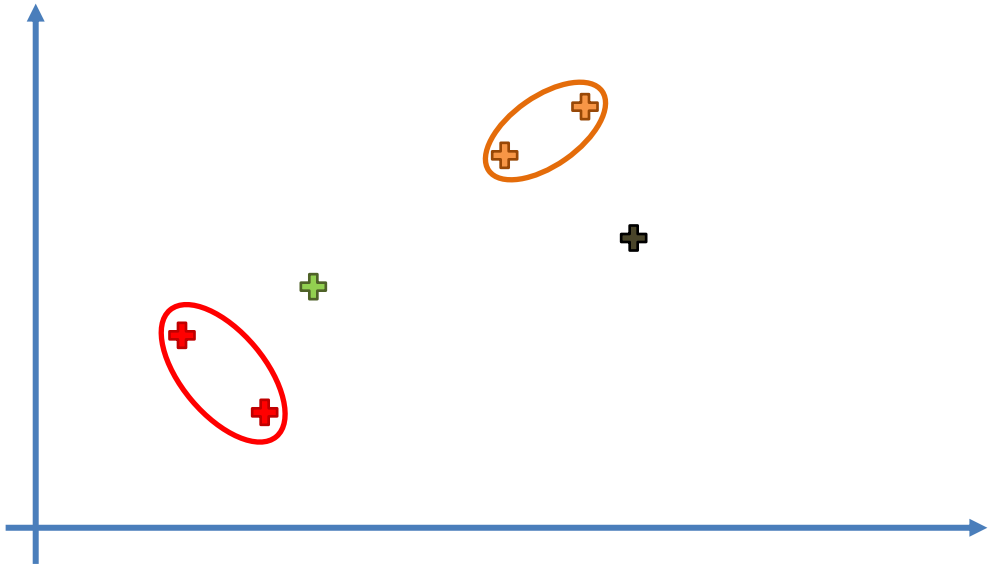
➡Así nos quedan 5 clusters



Clustering Jerárquico Aglomerativo

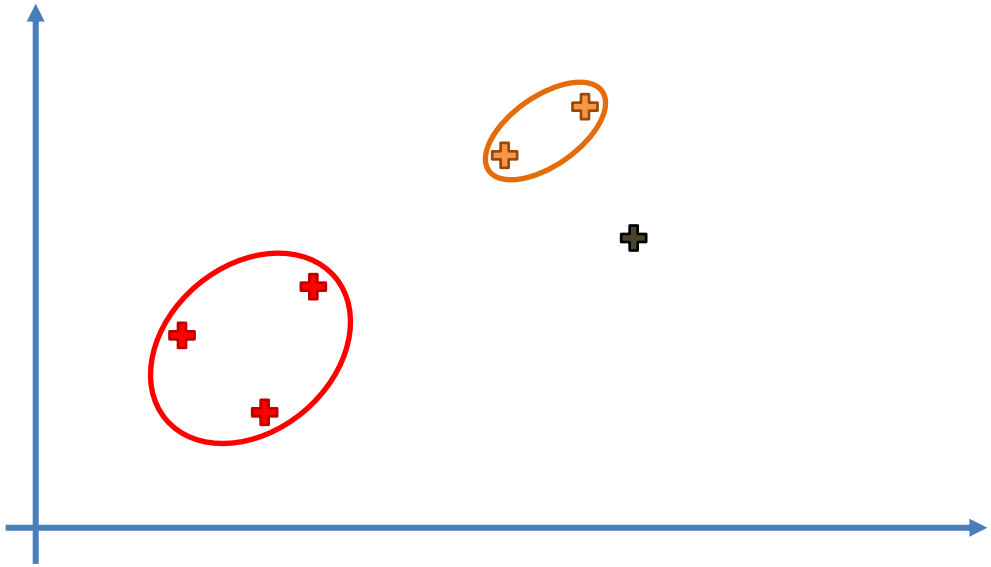
PASO 3: Elegir los dos clusters más cercanos y juntarlos en un único cluster

➡Así tenemos 4 clusters



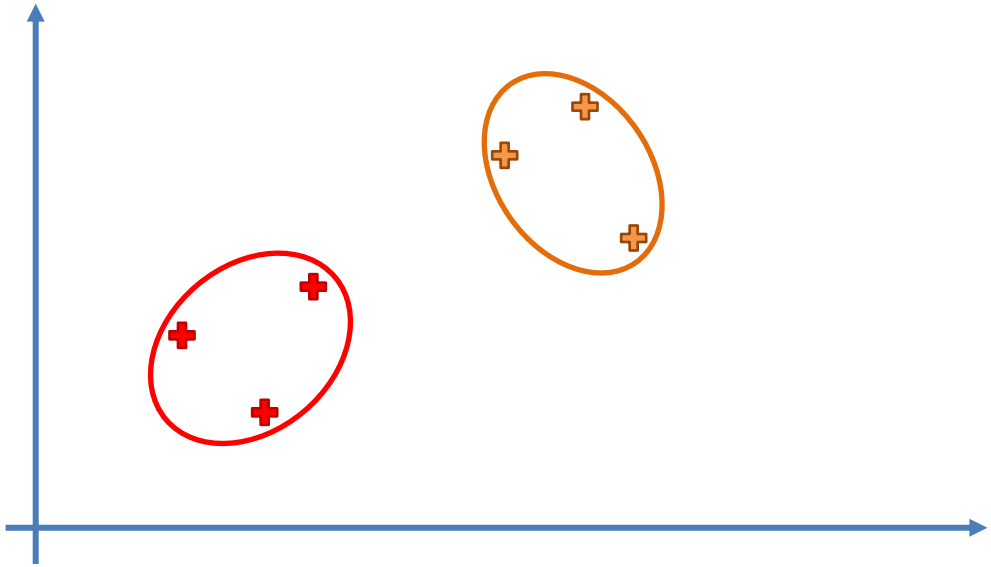
Clustering Jerárquico Aglomerativo

PASO 4: Repetir el PASO 3 hasta que quede un solo cluster



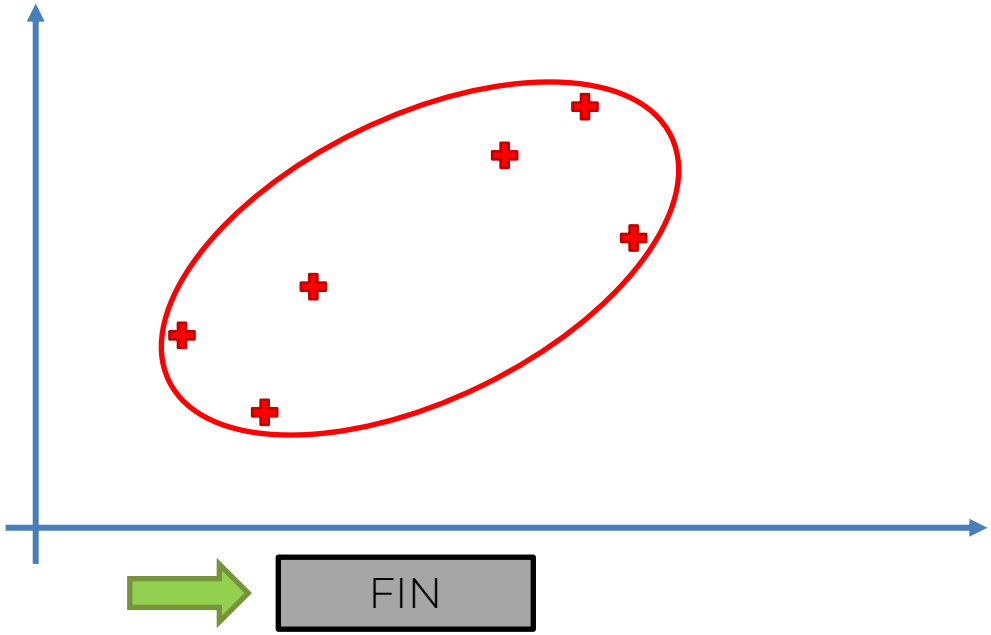
Clustering Jerárquico Aglomerativo

PASO 4: Repetir el PASO 3 hasta que quede un solo cluster



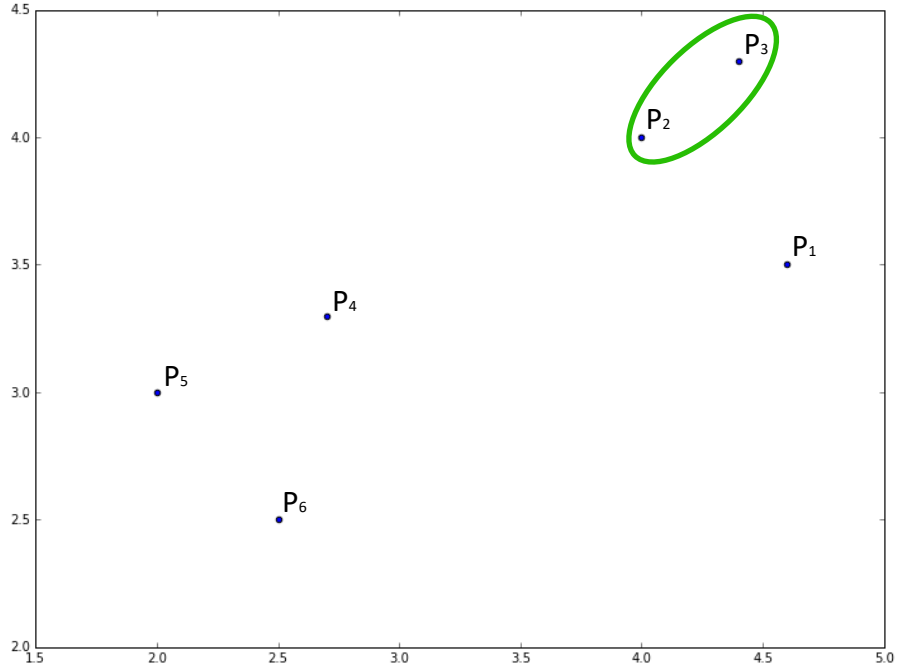
Clustering Jerárquico Aglomerativo

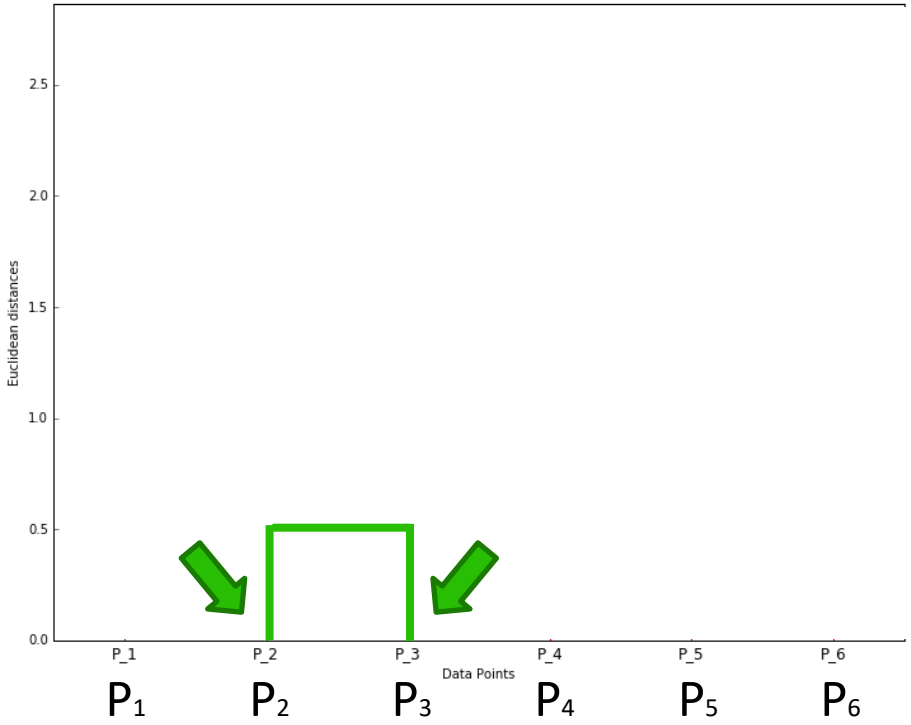
PASO 4: Repetir el PASO 3 hasta que quede un solo cluster



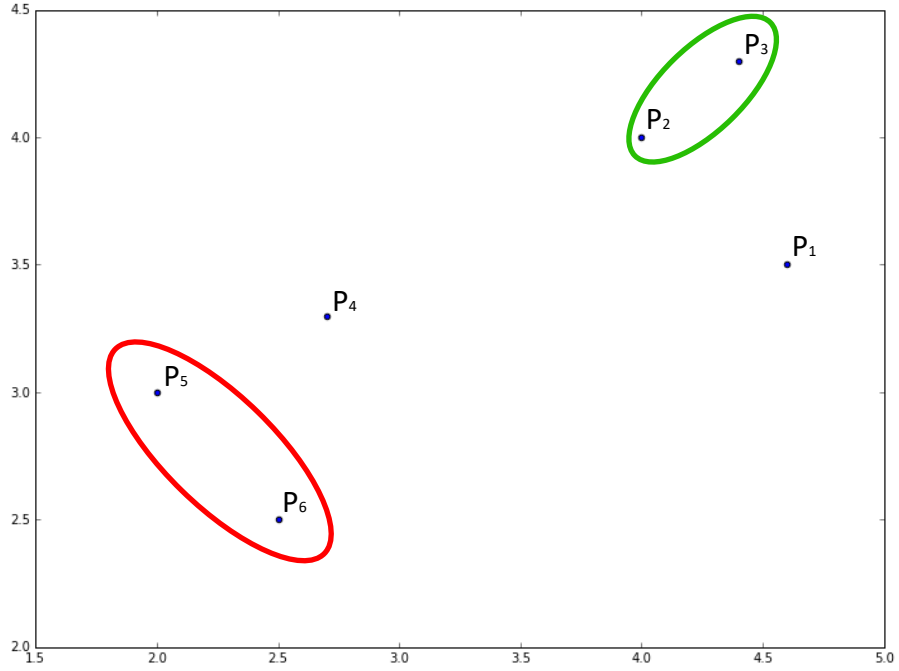
Idea del Clustering Jerárquico: ¿Cómo funcionan los dendrogramas?

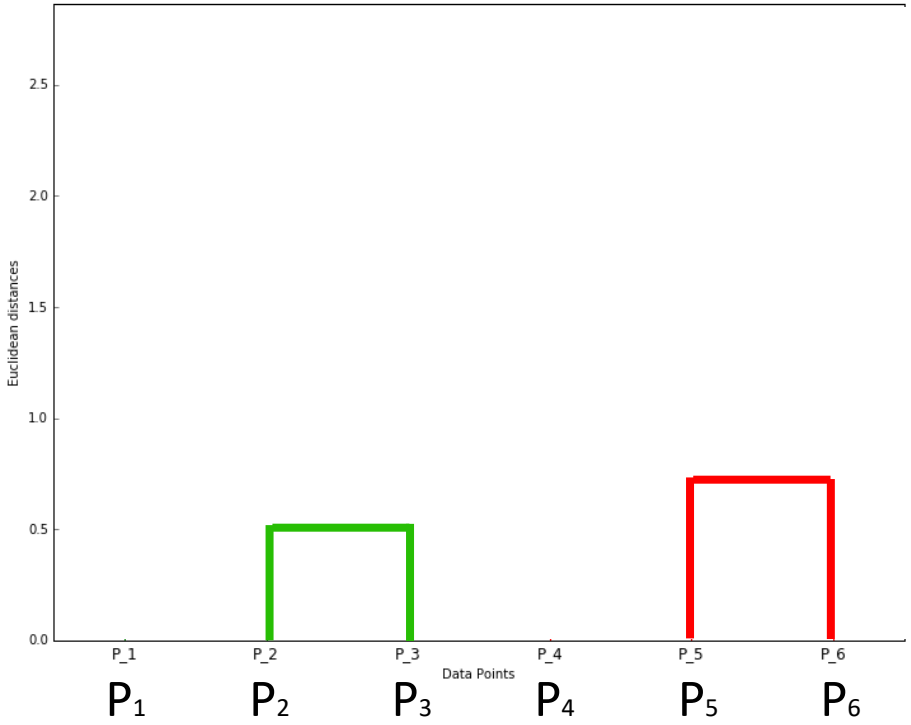
¿Cómo funcionan los dendrogramas?



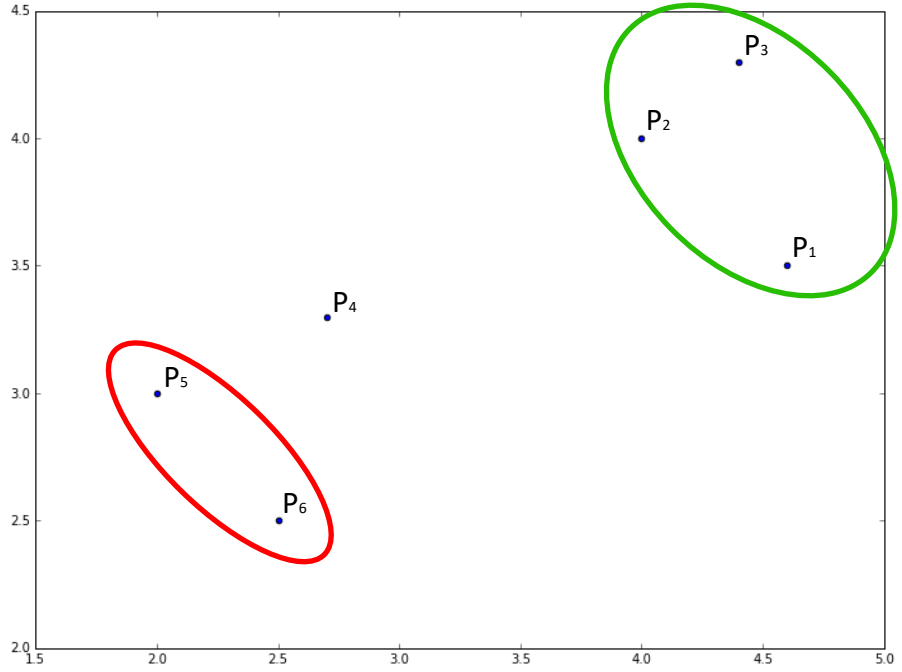


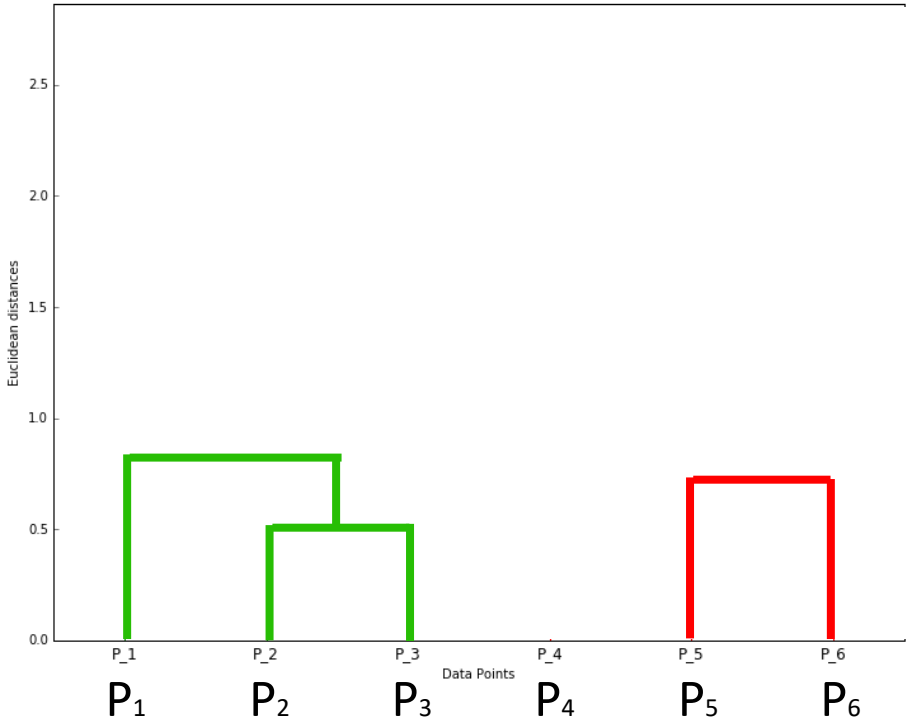
¿Cómo funcionan los dendrogramas?



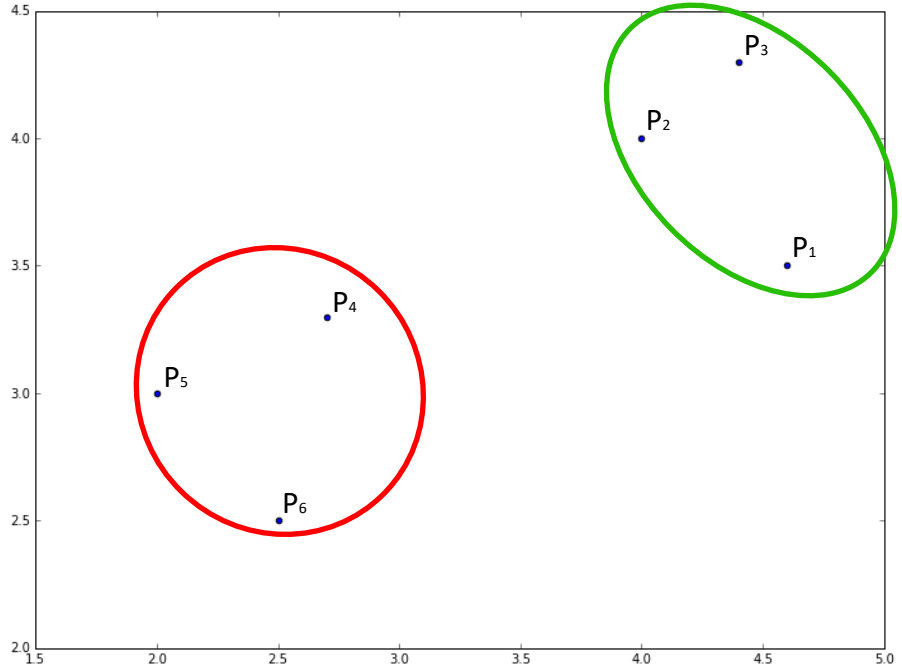


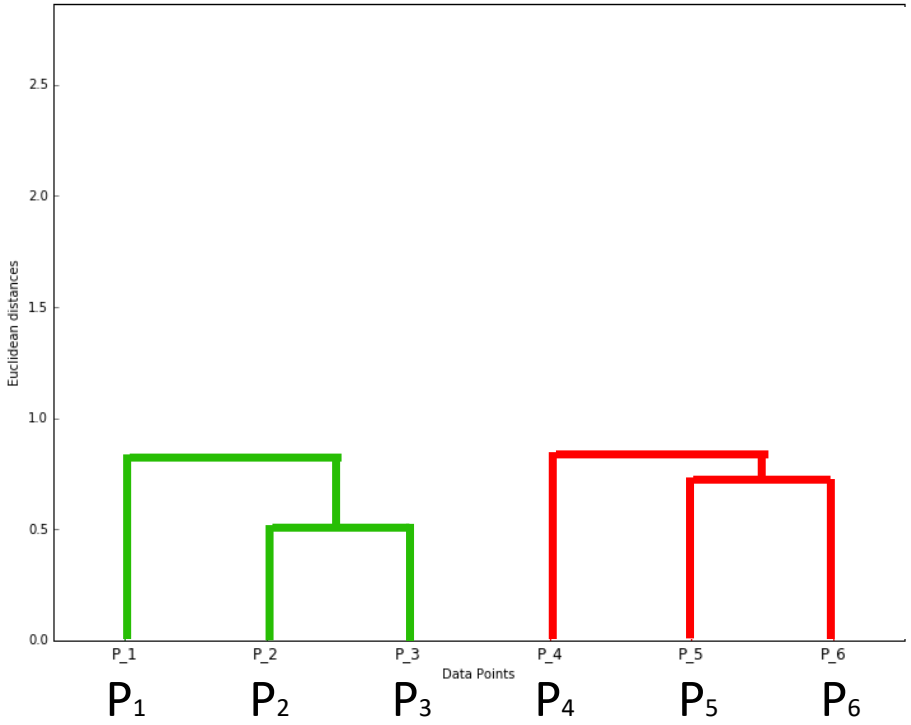
¿Cómo funcionan los dendrogramas?



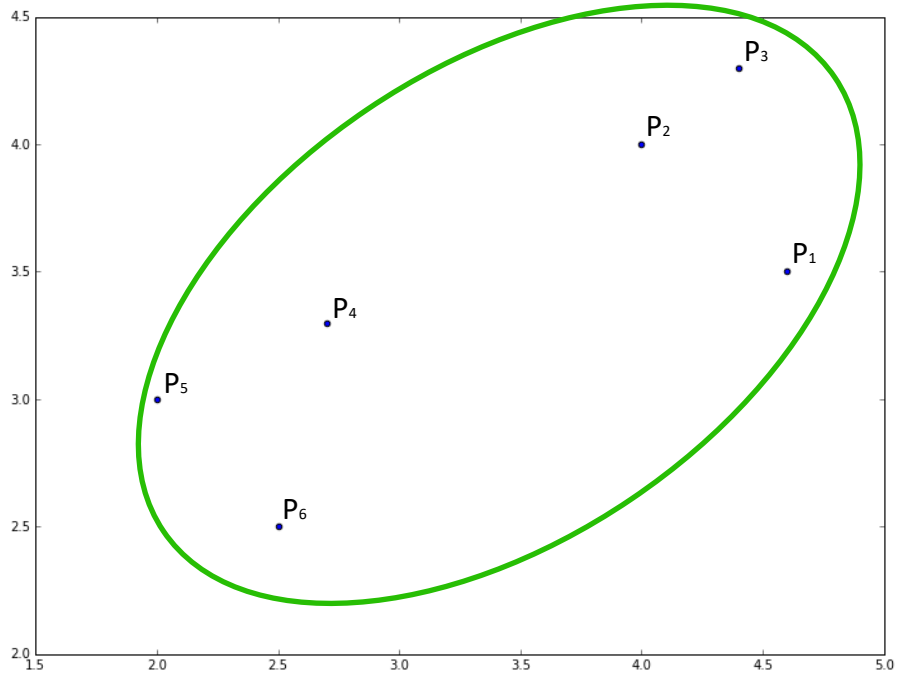


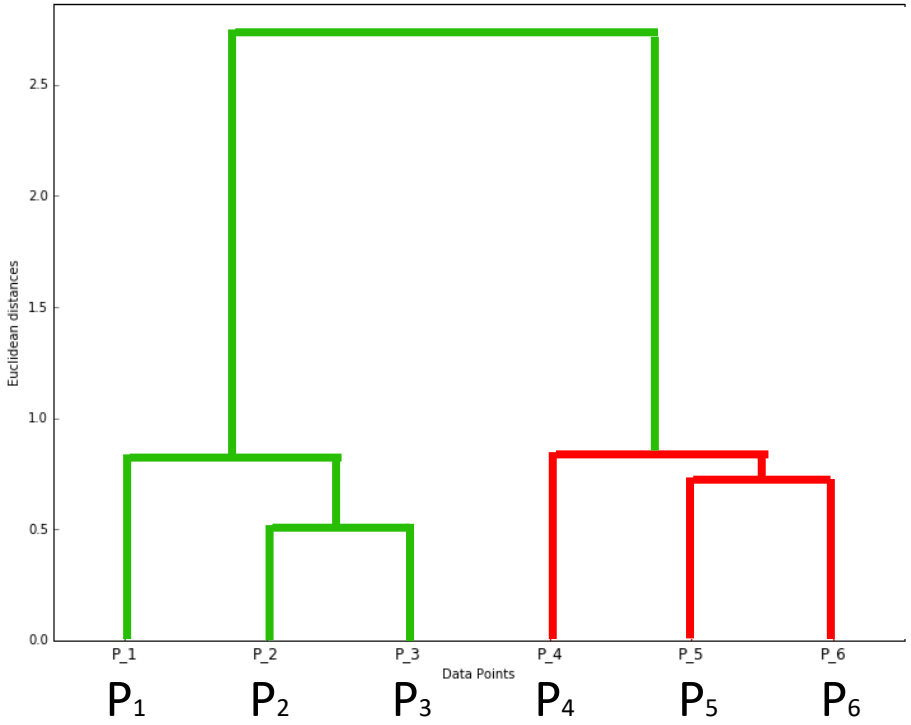
¿Cómo funcionan los dendrogramas?



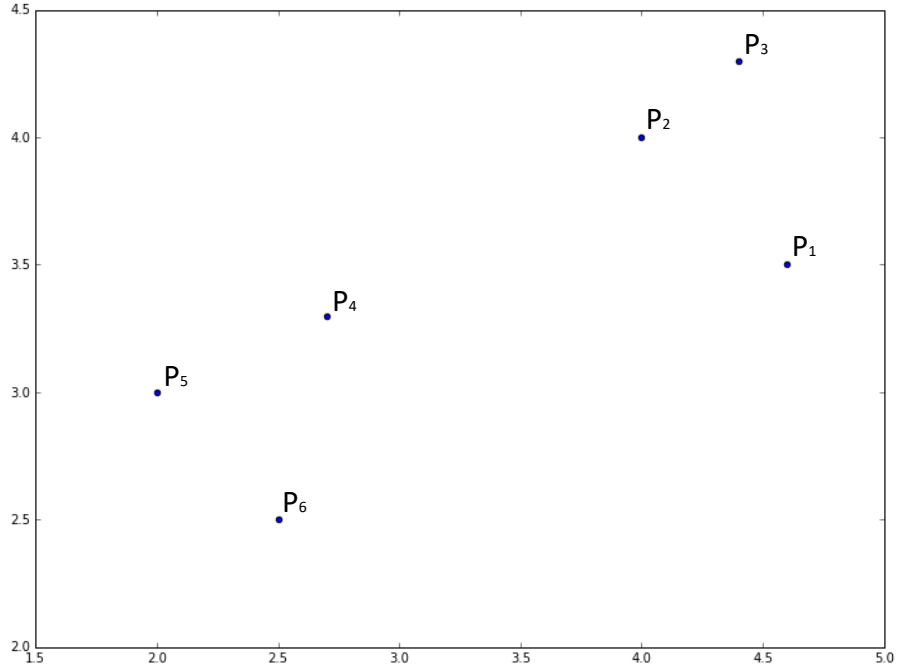


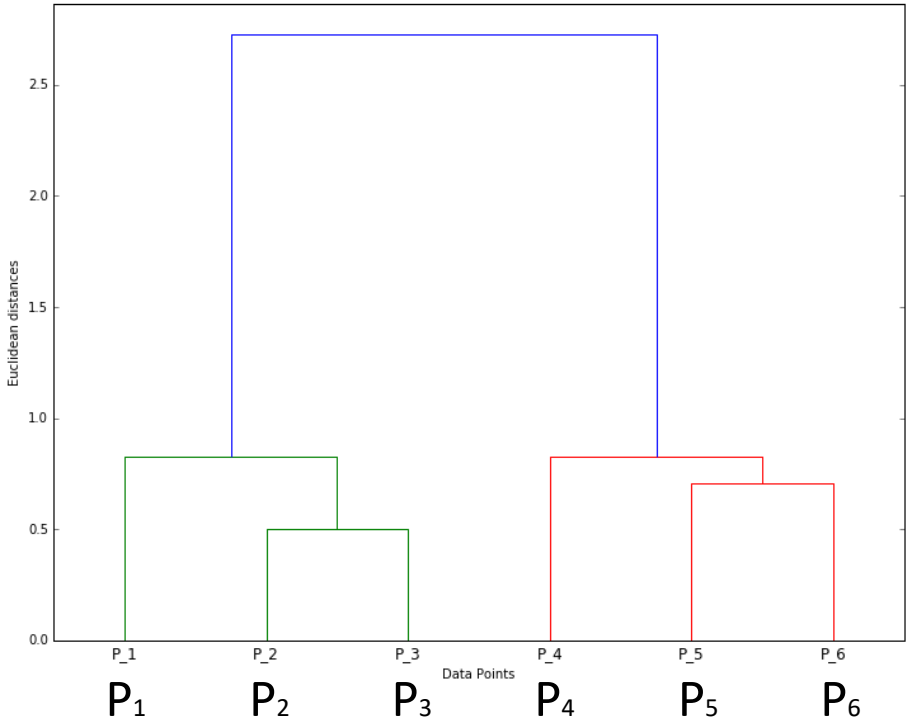
¿Cómo funcionan los dendrogramas?





¿Cómo funcionan los dendrogramas?





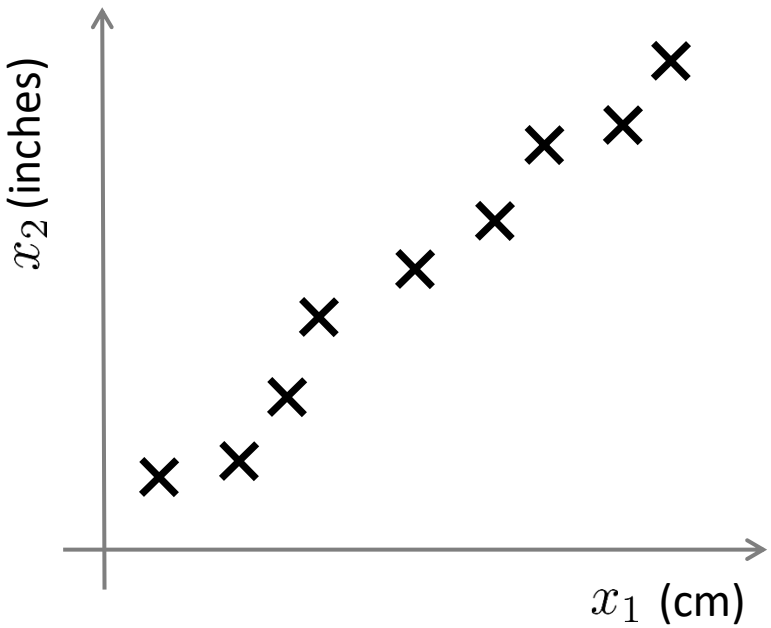
Reducción Dimensional (PCA)

Motivation

- Clustering
  - Una manera de resumir valores complejos de entender en una segmentación y llevarlos a puntos en los que podamos entenderlos visualmente.
- Dimensionality reduction
  - Una forma de simplificar la alta dimensionalidad de los datos.
  - Mejorar el performance de modelos en problemas de Big Data.

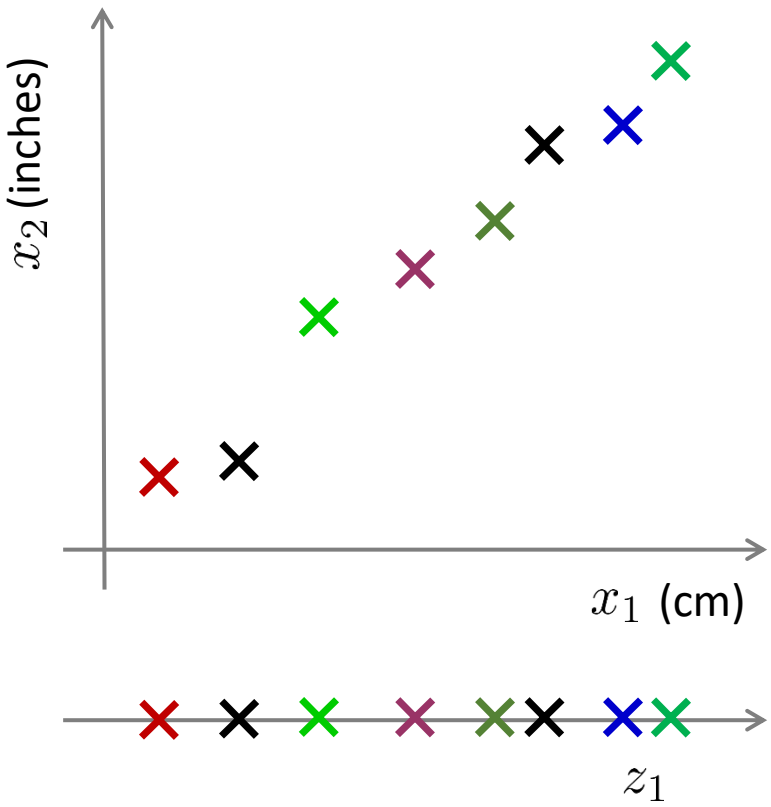
- Dado un grupo de puntos en *d* dimensiones
- Los convertimos en puntos de datos de $r<d$ dimensions
- Con la minima pérdida de información

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

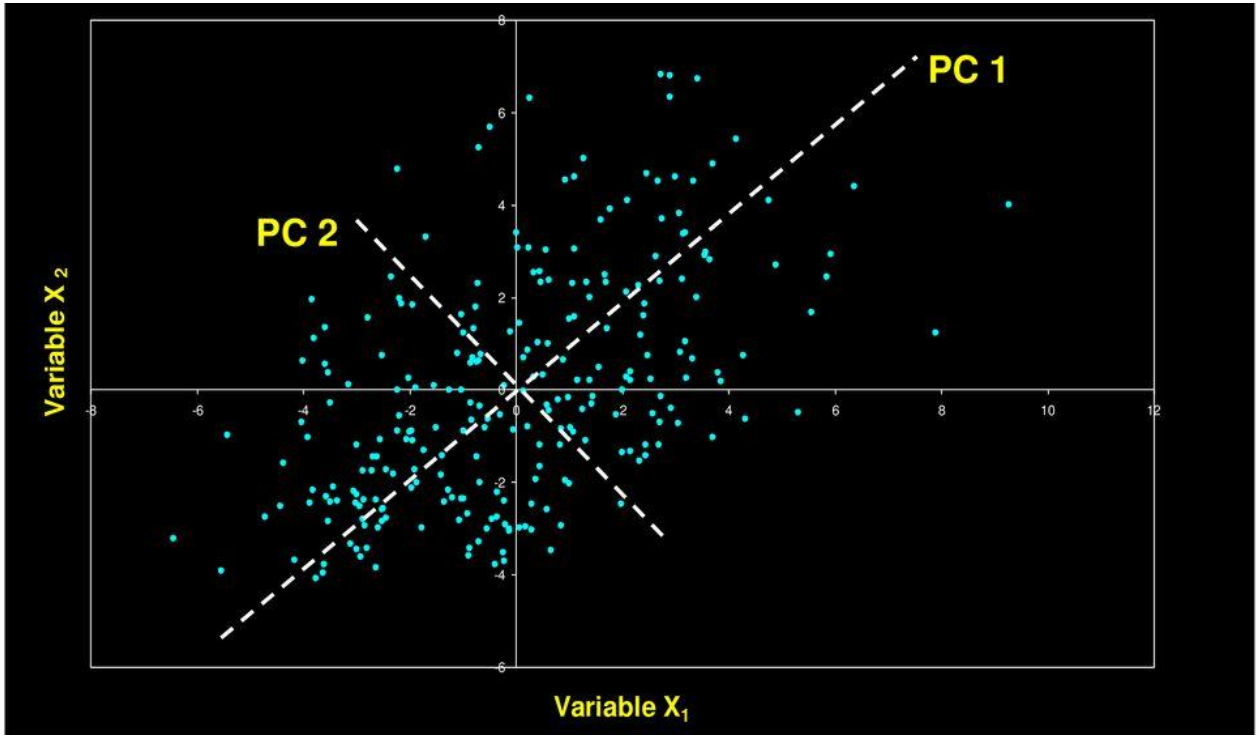
$$x^{(1)} \rightarrow z^{(1)}$$

$$x^{(2)} \rightarrow z^{(2)}$$

$$\vdots$$

$$x^{(m)} \rightarrow z^{(m)}$$

PCA encuentra las dimensiones de máxima varianza



PCA encuentra las dimensiones de máxima varianza

Eigenvector and Eigenvalue

$$Av = \lambda v$$

A: Matriz de Covarianza de X

v: Eigenvector or characteristic vector

λ: Eigenvalue or characteristic value

- ***The zero vector can not be an eigenvector***
- ***The value zero can be eigenvalue***

Metología para encontrar los “p” Componentes Principales

PASO 1: Aplicar escalado de variables a la matriz de características X, formada por m variables independientes.

PASO 2: Calcular la matriz de covarianzas de las m variables independientes de X.

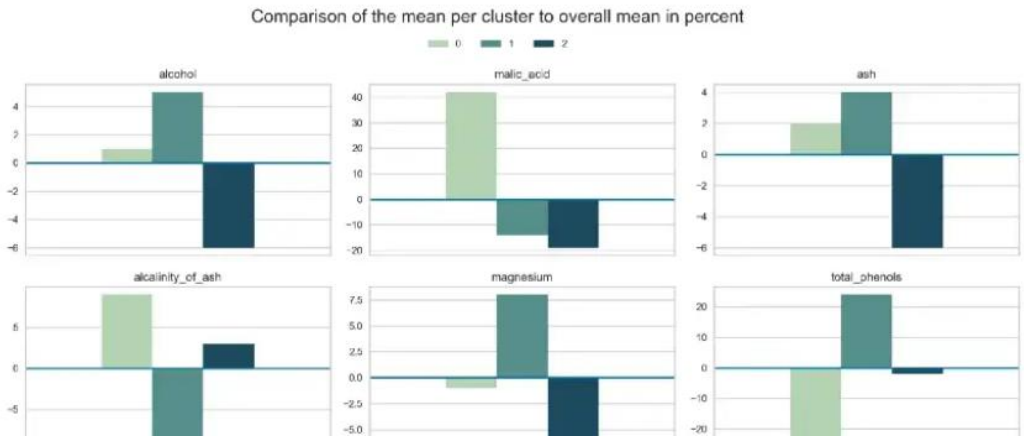
PASO 3: Calcular los valores y vectores propios de la matriz de covarianzas.

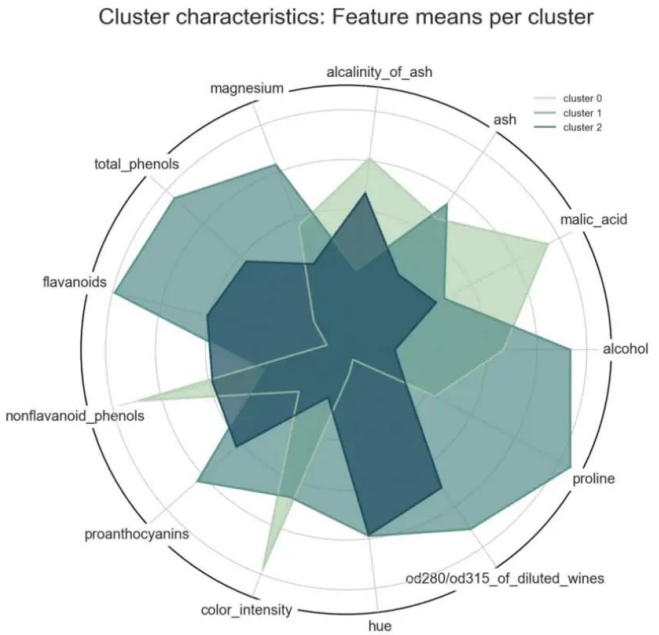
PASO 4: Elegir un porcentaje P de varianza explicada y elegir los $p \leq m$ valores propios más grandes

PASO 5: Los p vectores propios asociados a estos p valores más grandes son las componentes principales.
El espacio m-dimensional del dataset original se proyecta al nuevo subespacio p-dimensional de características, aplicando la matriz de proyecciones (que tiene los p vectores propios por columnas).

Visualización de características de Cluster

Existen diversas formas para visualizar las características predominantes en cada cluster, esto forma parte del proceso de “Perfilamiento”.





* https://towardsdatascience.com/best-practices-for-visualizing-your-cluster-results-20a3baac7426

VAMOS AL CÓDIGO!!