

Técnicas de Clustering

Las tres técnicas más utilizadas en clustering

¿Qué vas a ver en esta lección?

- La filosofía del clustering
- Las tres técnicas más utilizadas
 - Hierarchical
 - K-means
 - GMM

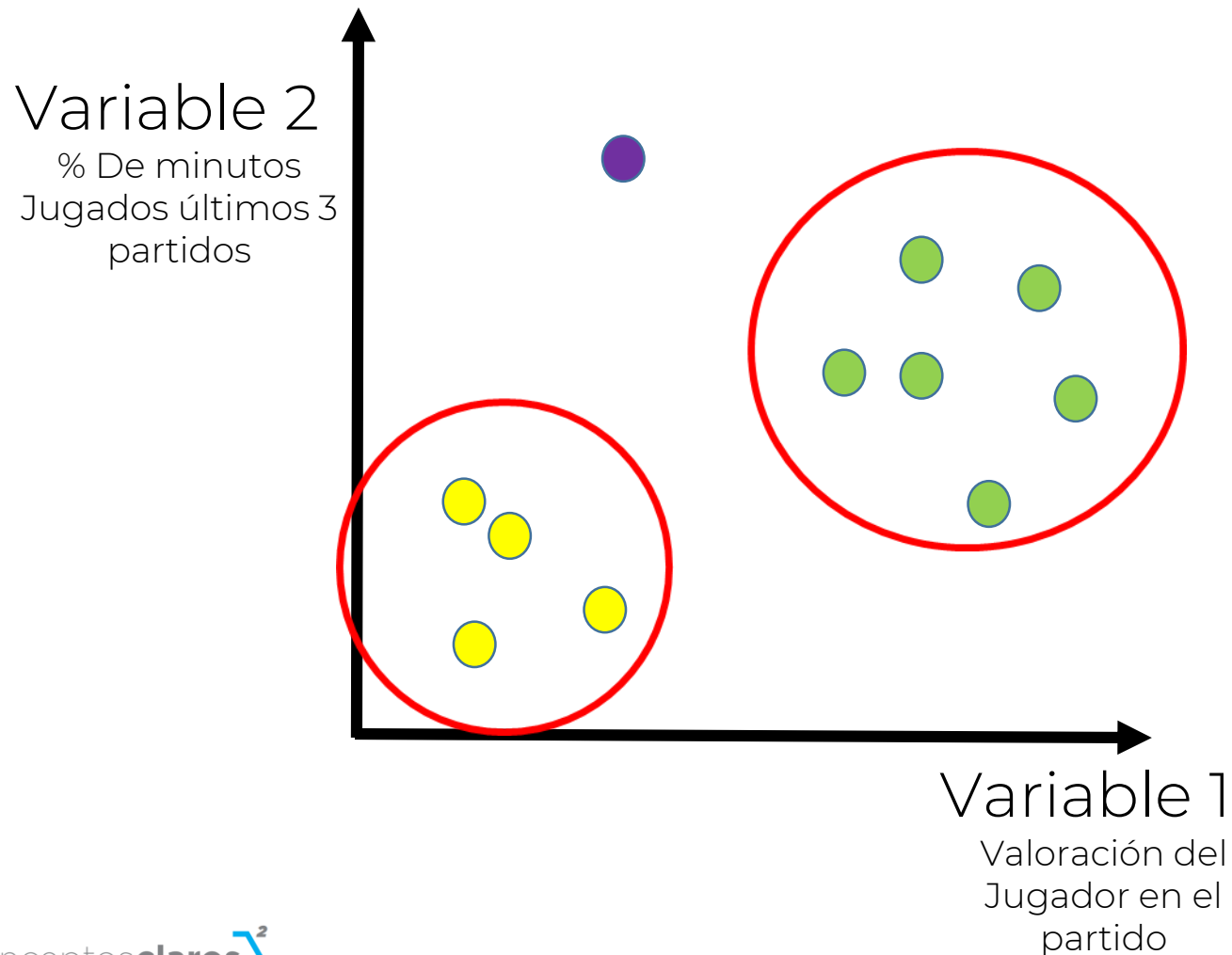
El mapa de aplicaciones prácticas de ciencia de datos



La filosofía del clustering

Reconocimiento de patrones para crear grupos

Clustering (no supervisado)



Objetivo:

A partir de características (features)

Construir grupos que sean lo suficientemente distintos

“Queremos crear 3 grupos de jugadores. Los malos, normales y buenos según el %de minutos jugados los últimos 3 partidos y su valoración”

Esquema del clustering

VARIABLES
DE
ENTRADA

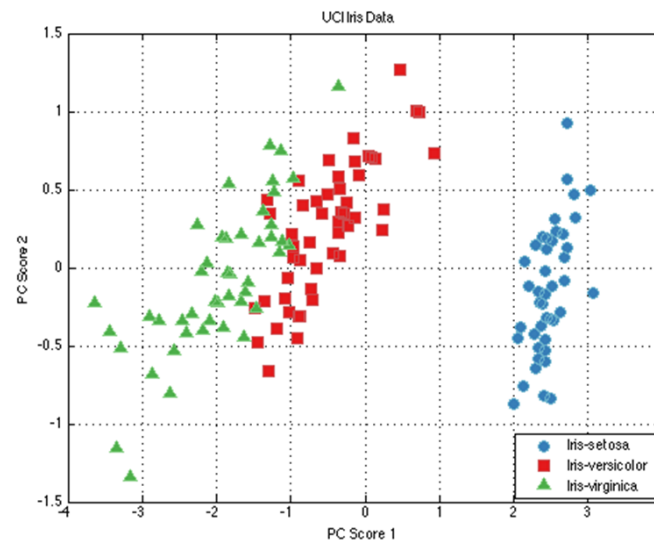


CLUSTERING



VARIABLE
DE
SALIDA

DIMENSIONES FLORES:
PETAL LENGTH
SEPAL LENGTH
PETAL WIDTH
SEPAL WIDTH



CLASE – ESPECIE DE FLOR

NO SUPERVISADO

CREO UNA NUEVA VARIABLE DE
SALIDA QUE ES UNA CLASE

El proceso

- Estandarizar los datos para evitar problemas con las unidades
- Encontrar una matriz de distancias que te permita saber lo lejos que están los puntos entre si – dissimilarity matrix (en GMM se utilizan modelos gaussianos)
- Utilizar un método para cortar en grupos según las distancias
 - Hierarchical
 - K-means
 - GMM

Clustering jerárquico o hierarchical

Cómo encontrar grupos con el clustering jerárquico

La matriz de distancias – dissimilarity matrix

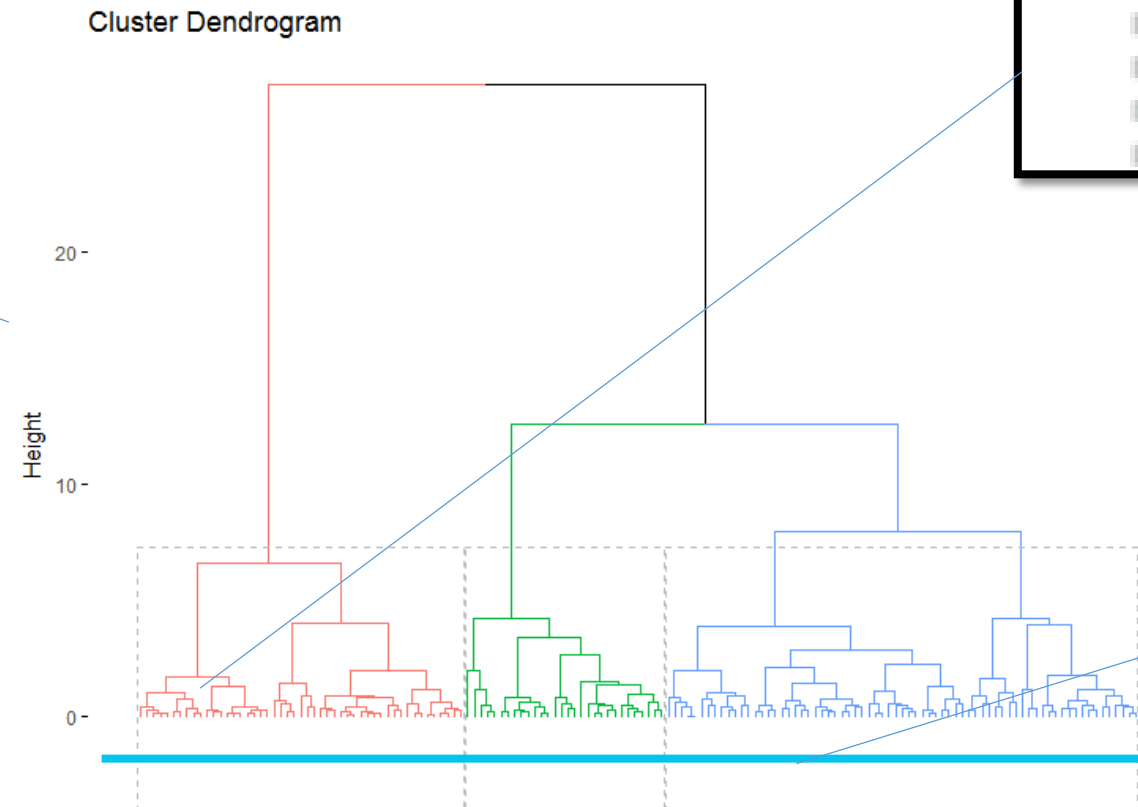
- La matriz de distancias – dissimilarity matrix
- Matriz cuadrado que nos indica cuanto de lejos están las observaciones teniendo en cuenta todas las variables
- Para calcular estas variables deberíamos tener las misma escala entre variables. Estandarizar es importante en el clustering

Names	Formula
Euclidean distance	$\ a - b\ _2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\ a - b\ _2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\ a - b\ _1 = \sum_i a_i - b_i $
Maximum distance	$\ a - b\ _\infty = \max_i a_i - b_i $
Mahalanobis distance	$\sqrt{(a - b)^T S^{-1} (a - b)}$ where S is the Covariance matrix

Crear un dendrograma a partir de la matriz de distancias

- La matriz de distancias – dissimilarity matrix

Distancia que
tenemos entre
grupos de
observaciones

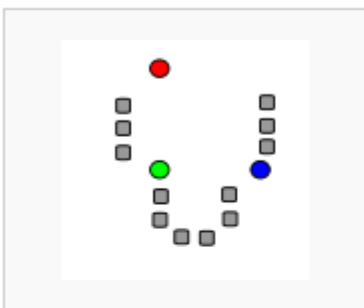


Clustering k-means

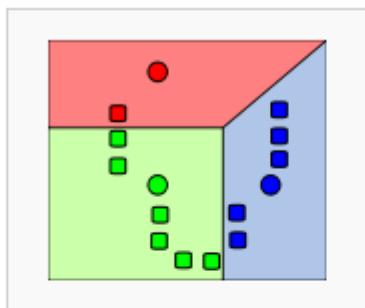
Cómo encontrar grupos con las medias de los grupos

K-means clustering

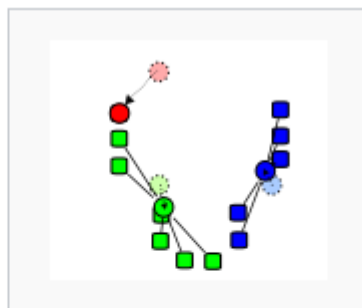
- Clustering basado en el centroide de los puntos
- Si $k=3$ el algoritmo empieza con 3 centroides y va colocando los otros puntos dónde más cerca están del centro y más lejos de los otros centroides



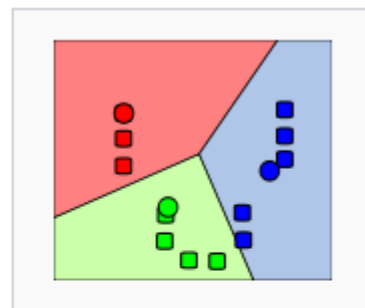
1) k centroides iniciales (en este caso $k=3$) son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).



2) k grupos son generados asociándole el punto con la media más cercana. La partición aquí representa el [diagrama de Voronoi](#) generado por los centroides.



3) EL [centroide](#) de cada uno de los k grupos se recalcula.



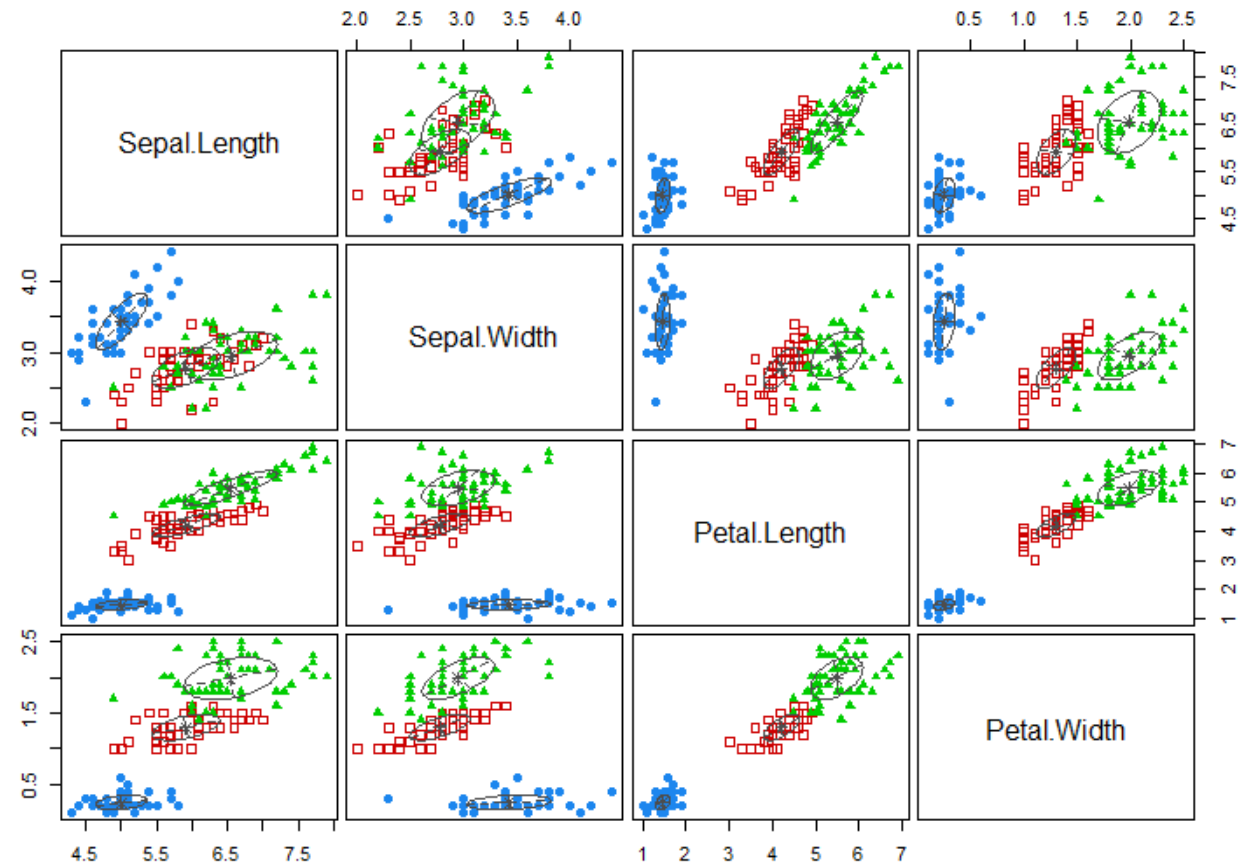
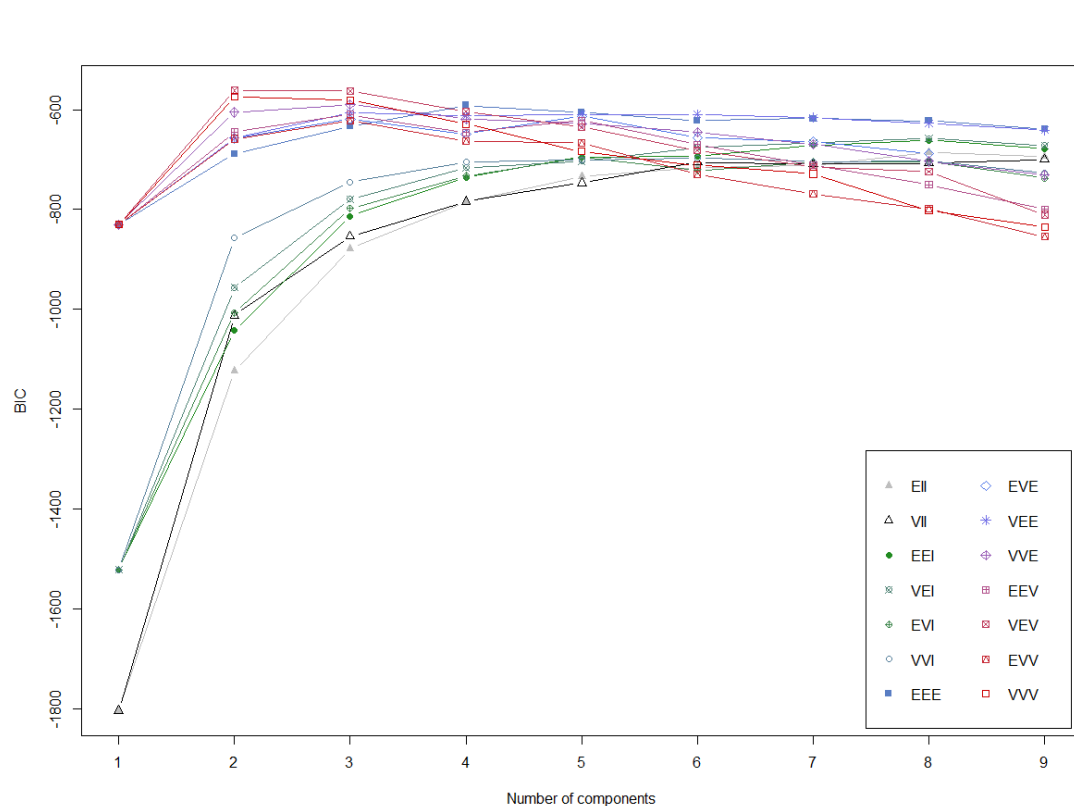
4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

Clustering con modelos gaussianos

GMM – Gaussian Mixture Models

GMM – Gaussian Mixture Models

- Clustering basado modelos gaussianos calculados a partir de las variables de entrada



Take away

El resumen de la lección

Lo más importante de la lección

- El clustering ayuda a crearnos grupos que se comportan de forma similar
- Tenemos tres técnicas interesantes:
 - Las mejores son k-means y GMM
 - Prueba de las dos y decide cual te funciona en tu caso 😊

Tú turno

Un pequeño ejercicio de clustering

Tú turno

- Descarga la hoja de trabajo que te he preparado
- Sigue los pasos para trabajar el clustering en los datos de ejemplo
- ¡A por ello!