

Intuición sobre Cluster Analysis

Análisis de Grupos

Luis R. Furlán

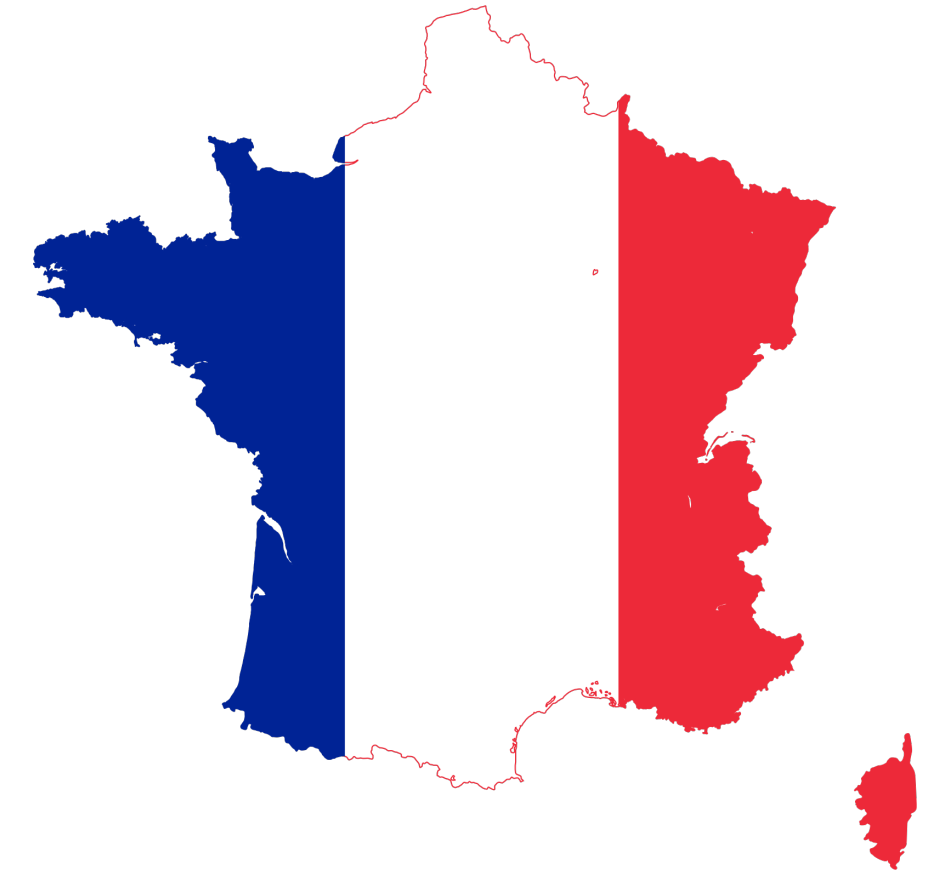
Agosto de 2021

Técnicamente hablando, el análisis de grupos es una técnica estadística multivariada que agrupa las observaciones, basadas en algunas de las características o variables que los describen

Intuitivamente, las observaciones en un conjunto de datos pueden ser divididos en grupos diferentes...y muchas veces esto puede ser de mucha ayuda

Ejemplo

Tenemos un conjunto de países



Ejemplo

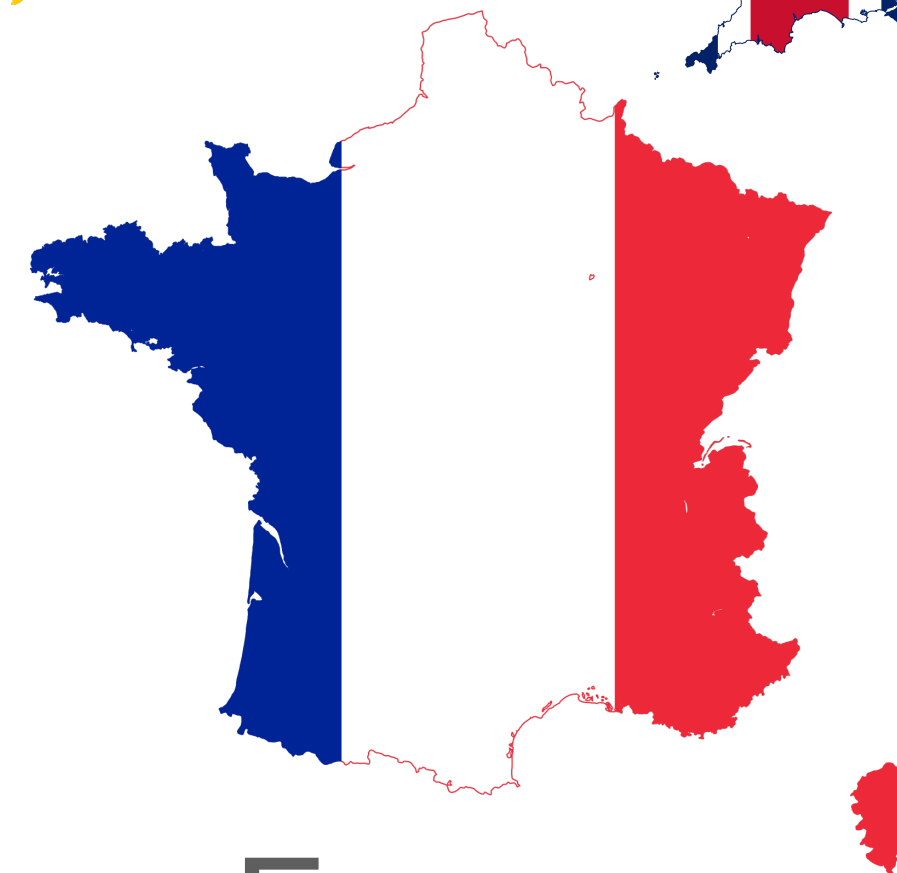
Imaginemos que hicimos un misterioso análisis y logramos formar 3 grupos

Cluster 1



Norteamérica

Cluster 2



Europa

Cluster 3



Australia

Ejemplo

¿Y si quisiéramos agruparlo en dos grupos? ¿Cuál sería el criterio?

Cluster 1



Hemisferio Norte

Cluster 2



Hemisferio Sur

Ejemplo

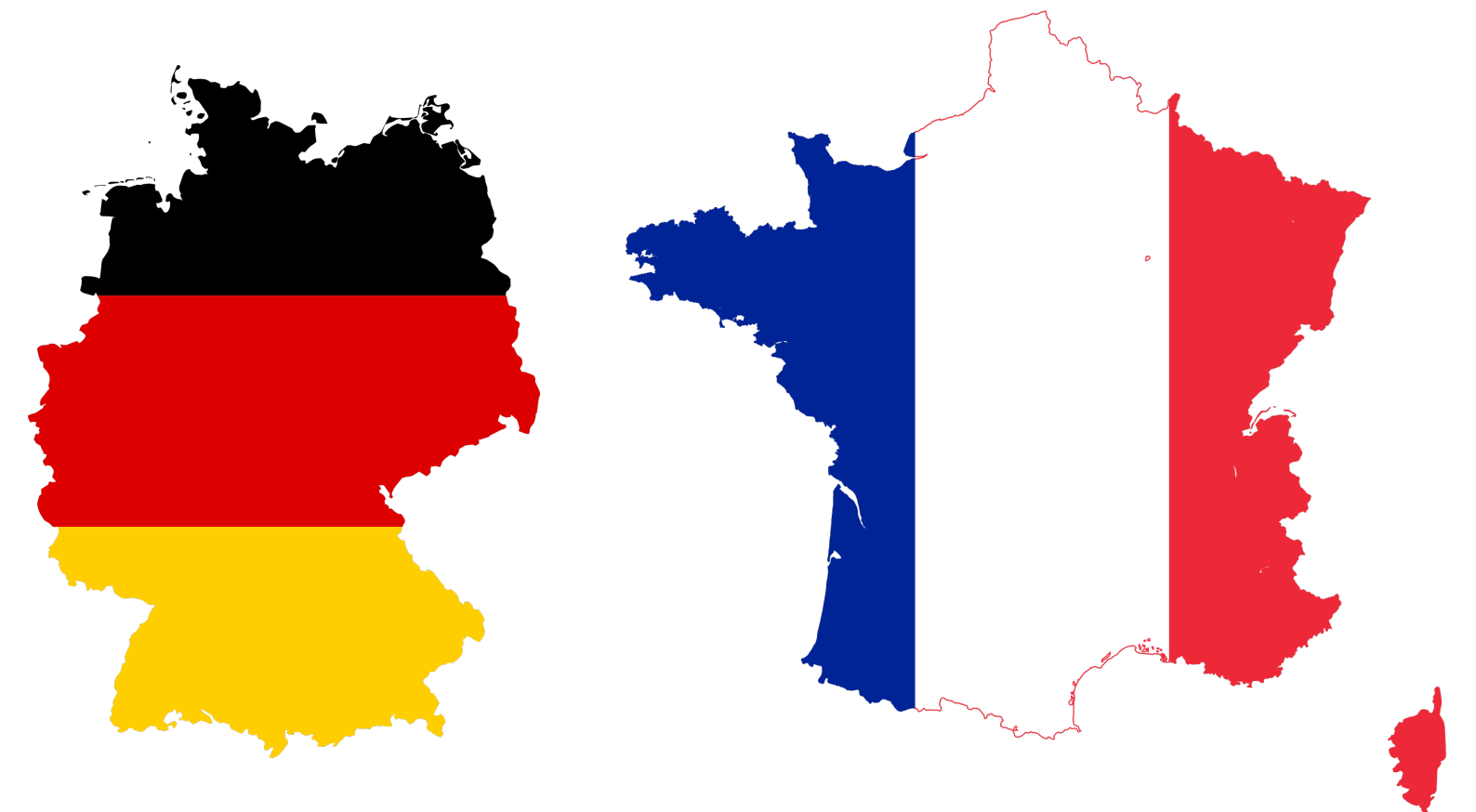
Parece que esto de Clustering es fácil...no tan rápido!!! ¿Cuál sería el criterio?

Cluster 1



Anglo Parlantes

Cluster 2



Otros idiomas

¿Cuál es la meta final del clustering?

La meta de clustering es la de maximizar la similitud de observaciones dentro de un grupo, y maximizar la disimilitud entre los grupos

¿Qué hemos visto hasta ahora?

- Qué es el Análisis de Grupos (Cluster Analysis)
- Qué es un Cluster
- Cuál es la meta final

¿Cuál es la diferencia entre Clustering y Clasificación?

- Hemos visto Regresión Lineal y Regresión Logística
- Dentro del contexto de ML estos son Regresión y Clasificación
- Estos se caen bajo el contexto de **aprendizaje supervisado**

Aprendizaje Supervisado

- Tenemos **datos etiquetados**
- Tenemos **entradas**
- Tenemos **valores correctos de salida**

MODELO (ENTRADAS) \longrightarrow SALIDAS \longrightarrow VALORES CORRECTOS DE SALIDA

Cluster Analysis

Aprendizaje no-supervisado

- Agrupamos los datos en varios grupos (clusters) pero no tenemos la menor idea si:
 - El número de grupos es correcto
 - Si son correctos
 - Si son útiles para algo

MODELO (ENTRADAS) \rightarrow SALIDAS \rightarrow ????

Es mas, lo que obtenemos es algo que nosotros mismos debemos **nombrar!**

Para resumir

CLASIFICACIÓN

Predecir una categoría de salida, dados datos de entrada

CLUSTERING

Agrupar datos basados en las similitudes entre ellos, y las diferencias de otros

Ya sabíamos que:



está en Norteamérica y que



está en Europa

Allí no hay ninguna Ciencia de Datos!!!

El análisis de grupos es muy útil

Se usa para

- Segmentación de mercados
- Segmentación de imágenes
- Muchas aplicaciones más

Análisis de grupos (Clusters)

Lo que veremos:

- Varios problemas de agrupamiento
- Cómo realizar el análisis de grupos (Cluster Analysis)
- Cómo encontrar el número óptimo de grupos
- Cómo identificar las características apropiadas
- Cómo interpretar los resultados

Nuestras herramientas principales: pandas y sklearn