

EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Aprendizaje no supervisado.

Objetivo: En esta última sección del curso cambiaremos el enfoque hacia aprendizaje no supervisado con el objetivo de estudiar las técnicas más usuales: reducción de dimensiones y agrupación de datos. En particular veremos algunas técnicas que se pueden usar en este contexto.

Lectura recomendada: Capítulo 12 de [1].

1. INTRODUCCIÓN

- Hasta ahora nos hemos concentrado en **aprendizaje supervisado**.
- Esto es, consideramos que tenemos una colección observaciones $\{(x_i, y_i)\}_{i=1}^n$.
- El objetivo es encontrar

$$y = f(x) + \epsilon. \quad (1)$$

- Ahora, estudiaremos técnicas de **aprendizaje no supervisado**. Sólo tendremos atributos (características) $x_i \in \mathbb{R}^p$.

1.1. Objetivos de aprendizaje no supervisado

- Descubrir cosas interesantes acerca de nuestras observaciones.
- La discusión se concentrará en:
 1. Métodos de reducción de dimensiones (visualización, o pre-procesamiento).
 2. Métodos de agrupación (visualización o pre-procesamiento).

1.2. Retos del aprendizaje no supervisado

- Es un mecanismo mucho más subjetivo (mayor comunicación con expertos en el área).
- Identificar patrones cuando no es claro el qué se está buscando.

1.3. Ventajas

- Es mucho más sencillo obtener **datos sin etiquetas** que obtener datos con un objetivo claro.
- Aplicaciones mucho más abiertas:
 - Encontrar grupos de pacientes con alguna característica en común.
 - Segmentación de clientes de acuerdo a patrones de compra.
 - Segmentación de creadores de contenido en redes sociales.

2. MÉTODOS DE REDUCCIÓN DE DIMENSIONES

- Buscan representaciones de menor dimensionalidad. Usualmente con el objetivo de crear visualizaciones.
- Encontrar patrones en un espacio de menor dimensión.

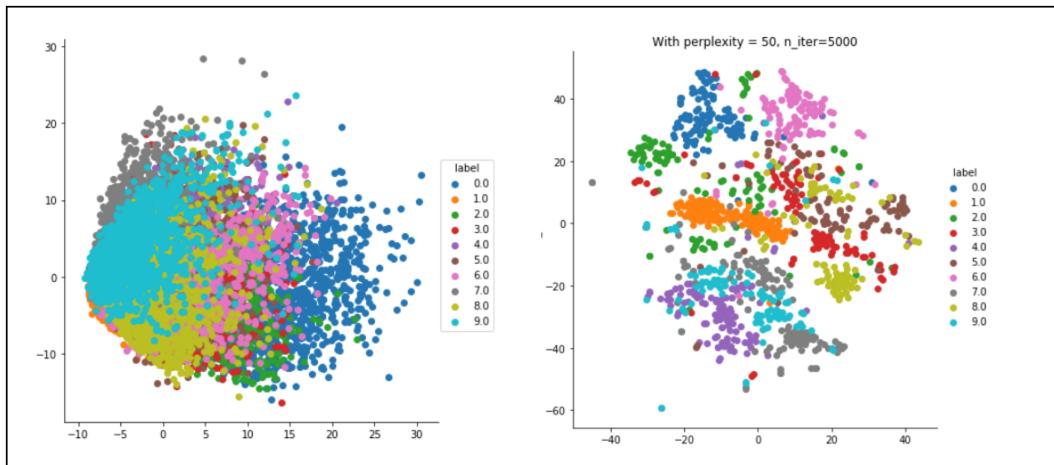


FIGURA 1. Los métodos de reducción de dimensiones pueden ser lineales (izquierda) o no-lineales (derecha).

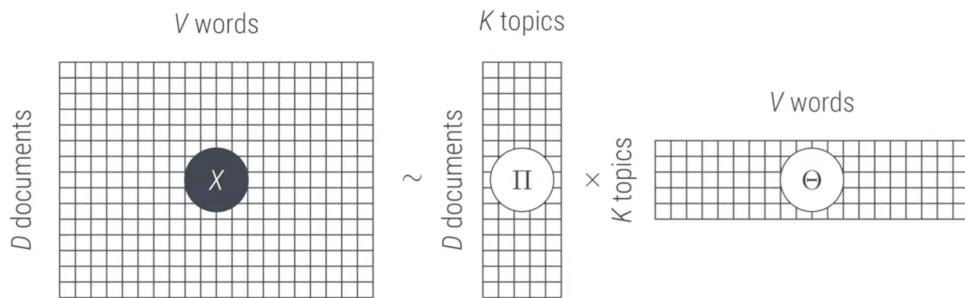


FIGURA 2. Modelo de tópicos sobre documentos en términos de contenido. Imagen tomada del curso de Probabilistic ML de Phillip Hennig.

- Factorización de matrices: sistemas de recomendación o modelado de tópicos.
- Aprendizaje de variedades: los datos pueden tener una estructura de subespacio.
- Arquitecturas de auto-codificación: Modelos basados en redes neuronales.
- Detección de datos atípicos: Máquinas de soporte vectorial

2.1. Observaciones

Los métodos utilizan distintos supuestos para poder construirse y necesitan validarse en la práctica o hacer un estudio de validación de supuesto.

3. ANÁLISIS DE COMPONENTES PRINCIPALES

- PCA es capaz de crear una representación de menor dimensión.
- Con PCA se puede crear un espacio de atributos que podría ser utilizados para tareas supervisadas.

3.1. El modelo

- Podemos pensar en PCA como un mecanismo iterativo para crear atributos.

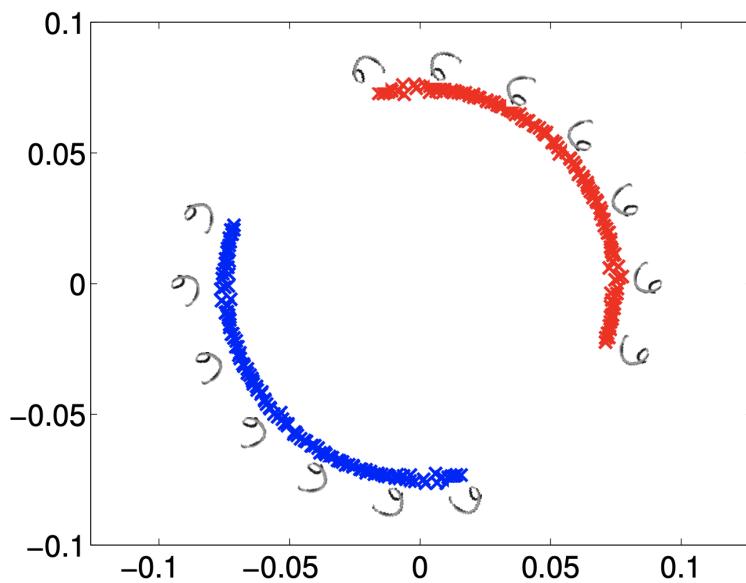


FIGURA 3. Imagen tomada de la escuela de verano en procesos Gaussianos 2014. Universidad de Sheffield.

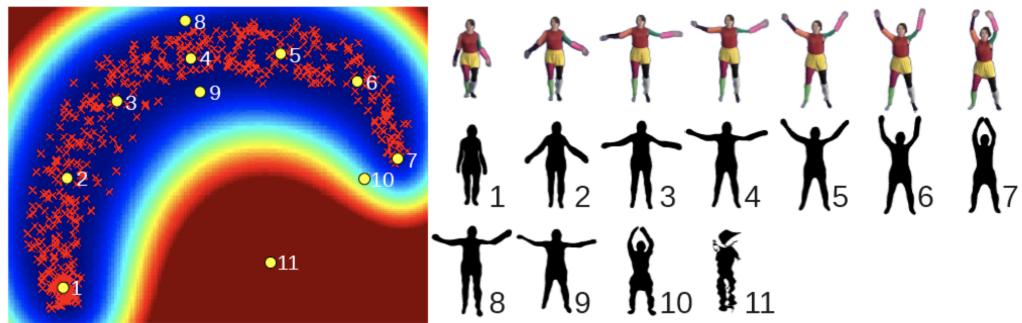


FIGURA 4. Imagen tomada de [2]. Modelo de seguimiento de poses.

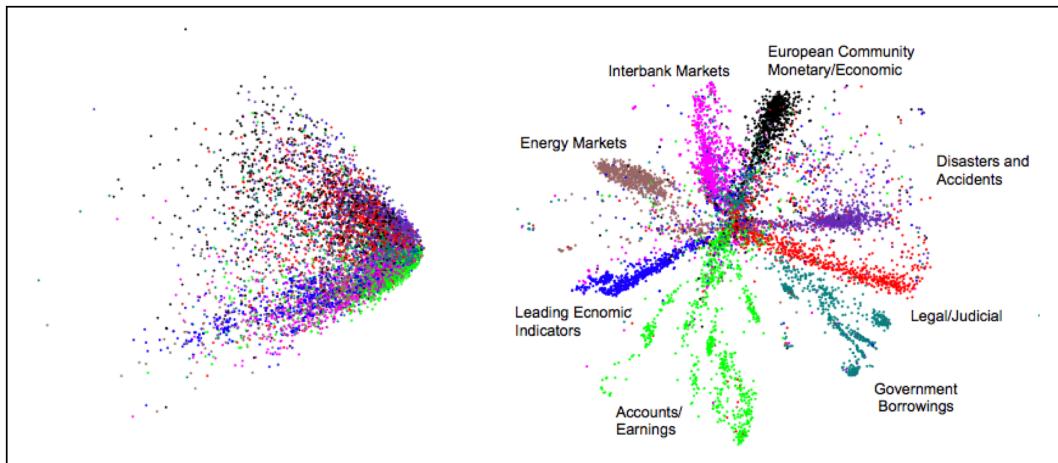


FIGURA 5. Se pueden usar transformaciones no-lineales para encontrar cierta estructura semántica (análisis de documentos). Imagen tomada de [3].

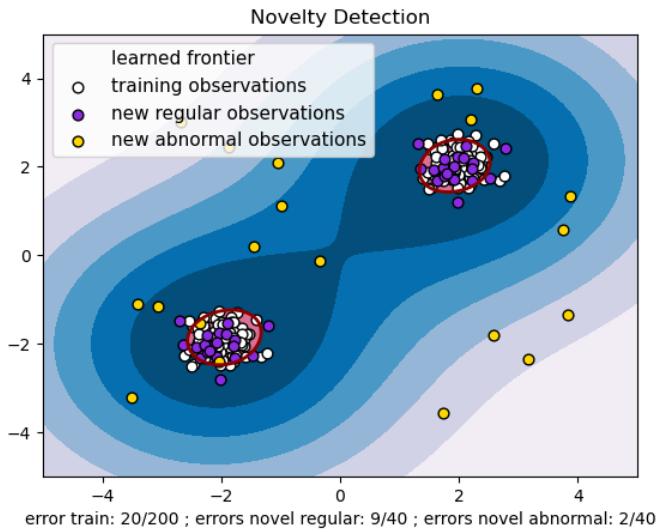


FIGURA 6. Detección de patrones, imagen tomada de los ejemplos de scikit-learn.

- El primer atributo que construimos a partir de X_1, \dots, X_p es buscar

$$Z_1 = \phi_{11}X_1 + \dots + \phi_{p1}X_p, \quad (2)$$

en donde Z_1 tiene la mayor varianza posible.

- Por simplicidad pedimos que $\sum_j \phi_{j1}^2 = 1$.
- En la literatura se llaman a los coeficientes $\phi_{11}, \dots, \phi_{p1}$ **cargas** o *loadings*.

3.2. Proceso

- Supongamos que tenemos un conjunto de datos que podemos organizar en $X \in \mathbb{R}^{n \times p}$.
- Supondremos que todas las columnas han sido **centradas** (es decir, el promedio de cada columna es igual a 0).
- Buscamos

$$z_{i1} = \phi_{11}x_{i1} + \dots + \phi_{p1}x_{ip}, \quad (3)$$

para cada $i = 1, \dots, n$.

- Nota que la **media** de las z_{i1} es cero y podemos escribir la **varianza** en términos de $\frac{1}{n} \sum_i z_{i1}^2$.

3.3. Formulación

- Buscamos resolver el problema

$$\max_{\phi_1 \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_{ij} \right)^2, \quad \text{sujeto a} \quad \sum_{j=1}^p \phi_{j1}^2 = 1. \quad (4)$$

- Este problema se puede resolver utilizando la **descomposición espectral** de la matriz X .
- Definimos Z_1 como el primer **componente principal**.

3.4. Geometría

- El vector de cargas ϕ_1 define la dirección en el espacio de atributos originales en la que los datos varían más.
- Si proyectamos en esta dirección, entonces recuperamos los *scores* z_{11}, \dots, z_{n1} .

3.5. Proceso: más componentes

- El segundo componente principal será una combinación lineal de los atributos, tal que tenga **máxima varianza** y que sea ortogonal a Z_1 .
- Por lo tanto los *scores* toman la forma

$$z_{i2} = \phi_{12}x_{i1} + \dots + \phi_{p2}x_{ip}, \quad (5)$$

donde el vector ϕ_2 es el vector de cargas del segundo componente principal.

3.6. Solución

- La restricción de ortogonalidad sobre los componentes nos permite encontrar las direcciones ϕ_1, ϕ_2, \dots como los vectores propios por la derecha de X .
- Las varianzas de los componentes están dados por $\frac{1}{n} \times \lambda_i^2$ donde λ_i son los valores propios.
- A lo más, hay $\min(n - 1, p)$ componentes principales.

3.7. Ejemplo:

- Tenemos datos del número de arrestos por cada 100,000 habitantes por distintos tipos de crímenes: **Assault**, **Murder**, and **Rape**. También tenemos la proporción de población que vive en zonas urbanas, **UrbanPop**.
- El registro es de 50 ciudades en EUA. Por lo tanto $n = 50$ y $p = 4$.

3.8. Interpretación

- El primer componente principal define la línea en un espacio de p dimensiones que es la más cercana a nuestras n observaciones (bajo distancia Euclídea).
- Esta idea se extiende naturalmente a buscar hiper-planos en más dimensiones.
- Por ejemplo, con los dos componentes recuperamos el plano más cercano a los datos.

3.9. Escala en los atributos

3.10. Proporción de varianza explicada

- Nos interesa medir la proporción de varianza de cada componente para entender la importancia o la capacidad de resumen de cada componente principal.
- La **varianza total** en el conjunto de datos se define como

$$\sum_{j=1}^p \mathbb{V}(X_j) \approx \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2, \quad (6)$$

la **varianza explicada** por cada componente principal la estimamos

$$\mathbb{V}(Z_m) \approx \frac{1}{n} \sum_{i=1}^n z_{im}^2. \quad (7)$$

- Se puede probar que

$$\sum_{j=1}^p \mathbb{V}(X_j) = \sum_{m=1}^M \mathbb{V}(Z_m), \quad (8)$$

con $M = \min(n - 1, p)$.

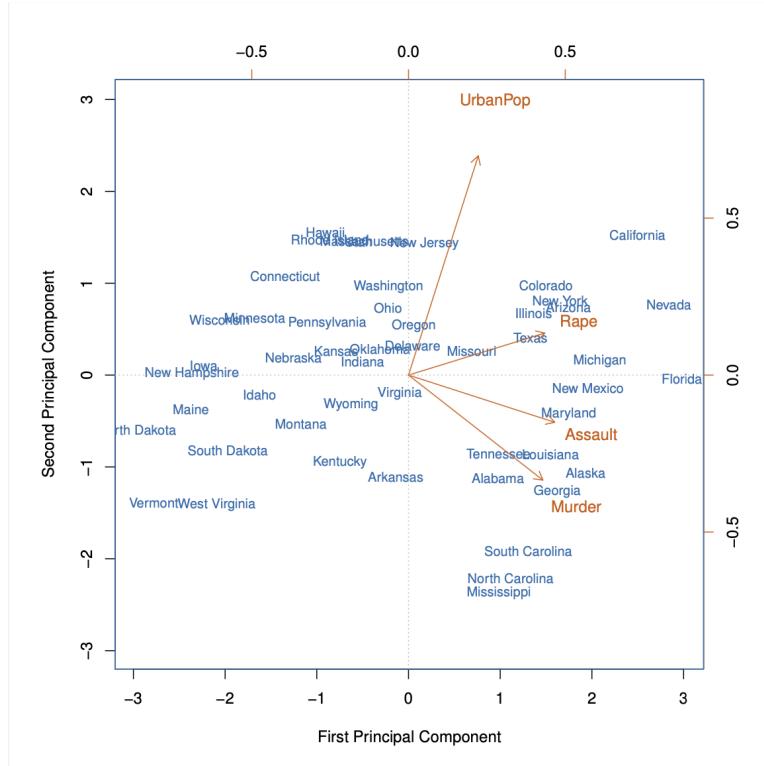


FIGURA 7. Imagen tomada de [1]. Gráfico tipo biplot que muestra los scores (observaciones) y los loadings (atributos).

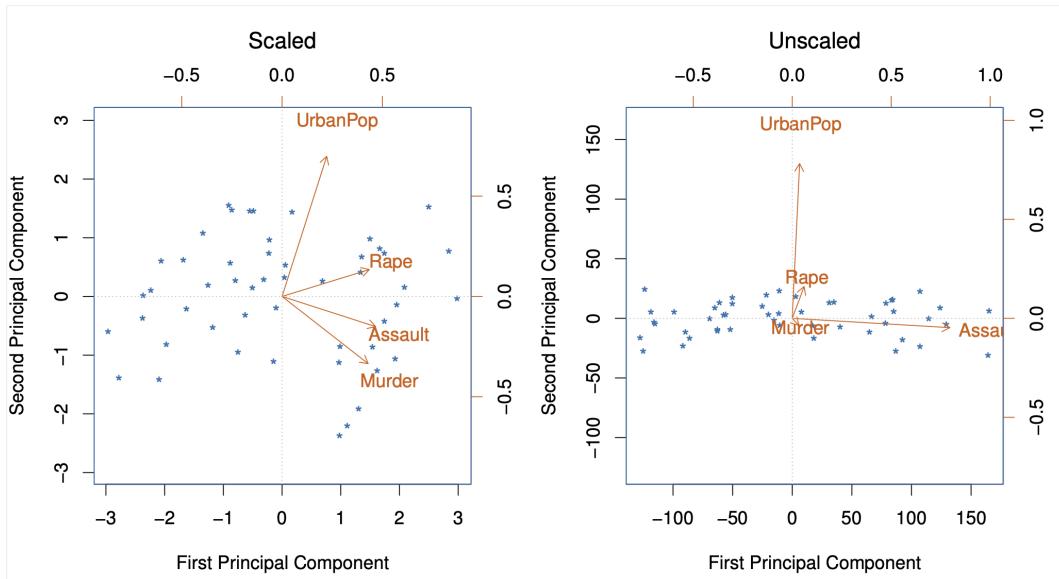


FIGURA 8. Imagen tomada de [1].

- Por lo tanto podemos calcular la proporción de varianza explicada como

$$\frac{\mathbb{V}(Z_m)}{\sum_{j=1}^p \mathbb{V}(X_j)}. \quad (9)$$

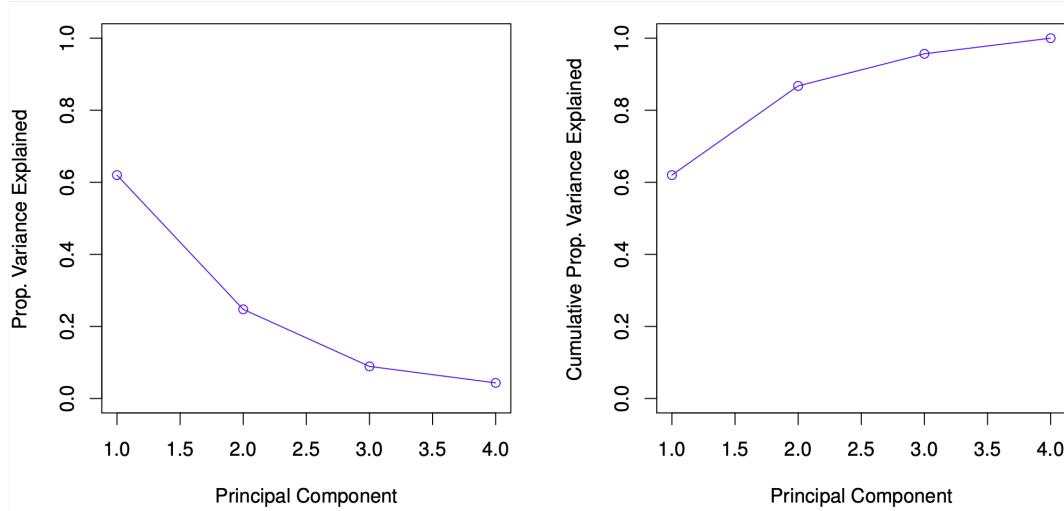


FIGURA 9. Imagen tomada de [1].

3.11. ¿Cuántos componentes habrá que utilizar?

Depende.

4. SEGMENTACIÓN

- Análisis de conglomerados, segmentación, agrupación, ó *clustering* es una técnica que busca encontrar grupos de observaciones.
- Las observaciones **dentro** de los grupos se deben de parecer lo mas posible.
- Las observaciones **entre** grupos deben de ser lo mas distintas posible.
- ¿Cómo medimos *similaridad* o *diferencias*?
 - Con conocimiento de dominio y específico a la aplicación que estamos tratando.

4.1. Conglomerados para segmentación de mercados

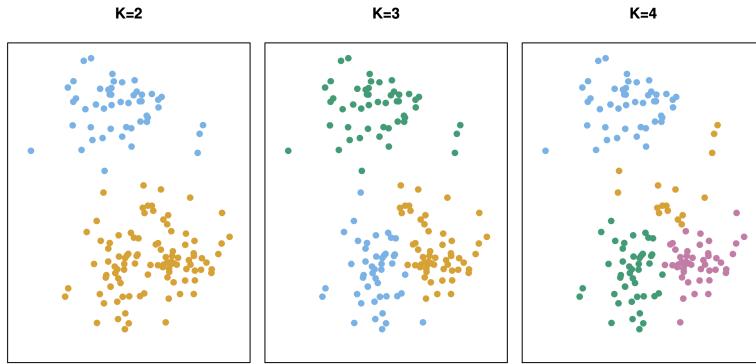
- Identificar grupos de personas que sean mas receptivas a un tipo particular de anuncios o que sean mas propensas a comprar un producto.

4.2. Métodos de segmentación (ejemplos)

- *K medias*, busca identificar los grupos.
- *Clustering jerárquico*, no sabemos cuántos grupos necesitamos.
- Modelo de mezclas Gaussianas.

5. K MEDIAS

Con un conjunto de datos ficticios de $n = 150$ y $p = 2$, obtenemos los siguientes grupos utilizando valores distintos para K .



5.1. ¿Cómo funciona?

Sean C_1, \dots, C_K los conjuntos de índices a los que pertenece cada instancia. Estos conjuntos satisfacen:

1. $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$.
2. $C_j \cap C_k = \emptyset$ para toda $j \neq k$.

- Un buen algoritmo de K medias es aquel que tiene una **variación dentro de grupos pequeña**.
- La métrica $\text{WCV}(C_k)$ representa qué tan diferentes son las observaciones dentro del grupo k .
- Buscamos resolver:

$$\min_{C_1, \dots, C_K} \sum_{k=1}^K \text{WCV}(C_k). \quad (10)$$

- Usualmente usamos distancia Euclideana

$$\text{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2, \quad (11)$$

donde $|C_k|$ denota la cuántas observaciones pertenecen al grupo k .

5.2. Algoritmo

1. Asignar al azar un grupo a cada observación.
2. Repetir hasta que las asignaciones no cambien:
 - a) Para cada grupo, encontrar el centro.
 - b) Asignar a cada observación el centro mas cercano.
 - c) Asignar el grupo al que pertenece.

5.3. Propiedades del algoritmo

- El algoritmo tiene garantía de disminuir la función objetivo en cada iteración.
- Nota que

$$\frac{1}{|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2 = 2 \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2, \quad (12)$$

donde \bar{x}_k es el centro del grupo k .

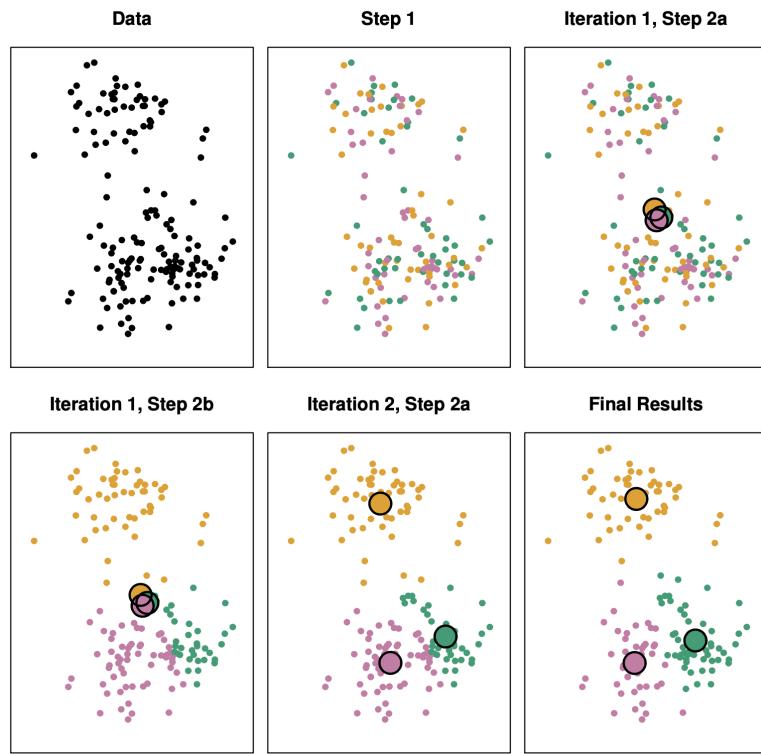


FIGURA 10. Imagen tomada de [1].

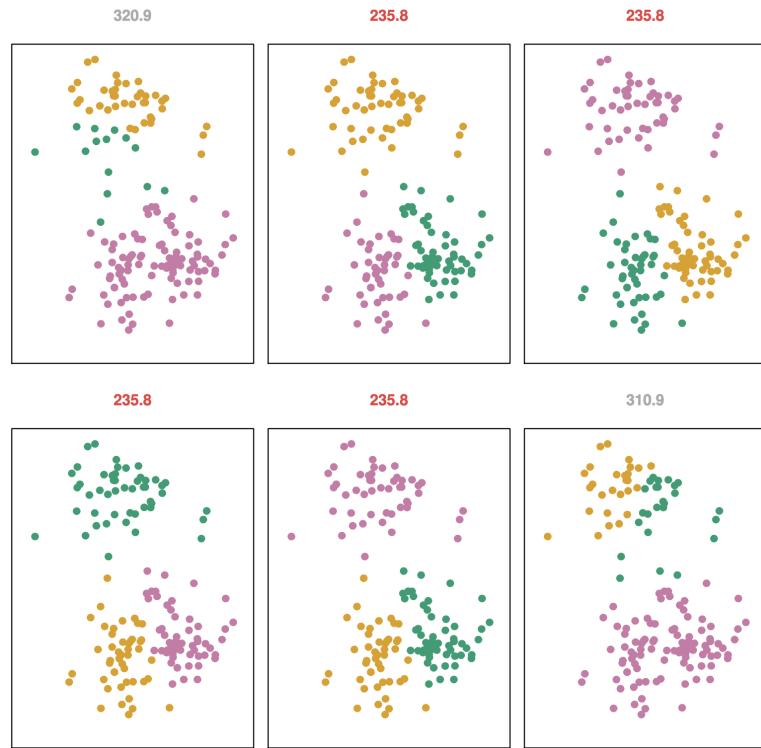


FIGURA 11. Imagen tomada de [1].

6. AGRUPACIÓN JERÁRQUICA

- No necesitamos determinar el número de grupos que necesitamos.
- El procedimiento es iterativo en donde los grupos se van juntando poco a poco.

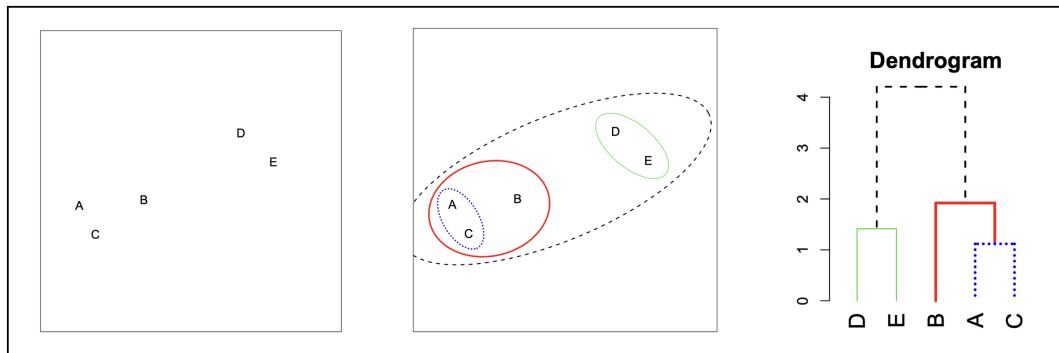


FIGURA 12. Imagen tomada de [1].

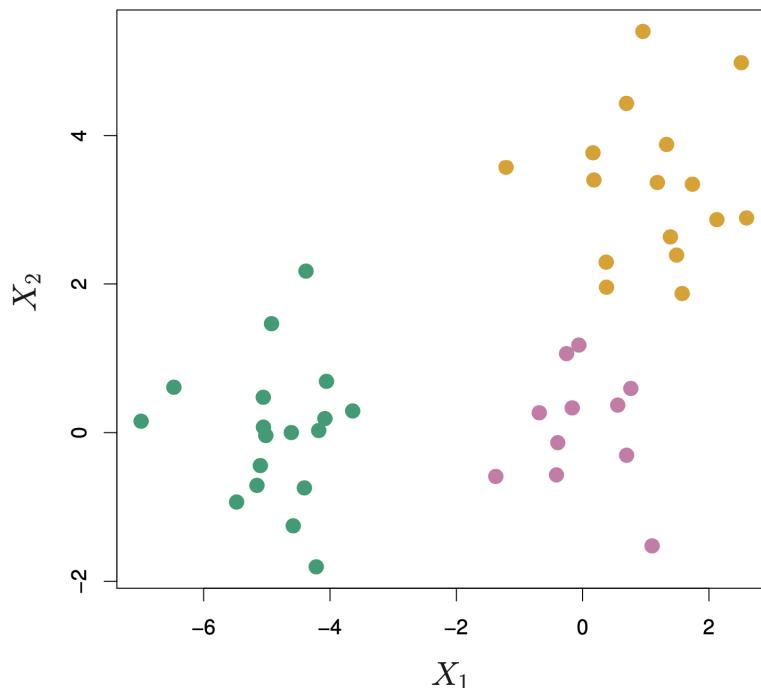


FIGURA 13. Imagen tomada de [1].

6.1. Puntos a considerar

- Escalas en los atributos.
- Escoger la métrica.
- Escoger el número de *clusters*.
- Escoger los atributos que se usarán.

REFERENCIAS

- [1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. [1](#), [6](#), [7](#), [9](#), [10](#), [11](#)

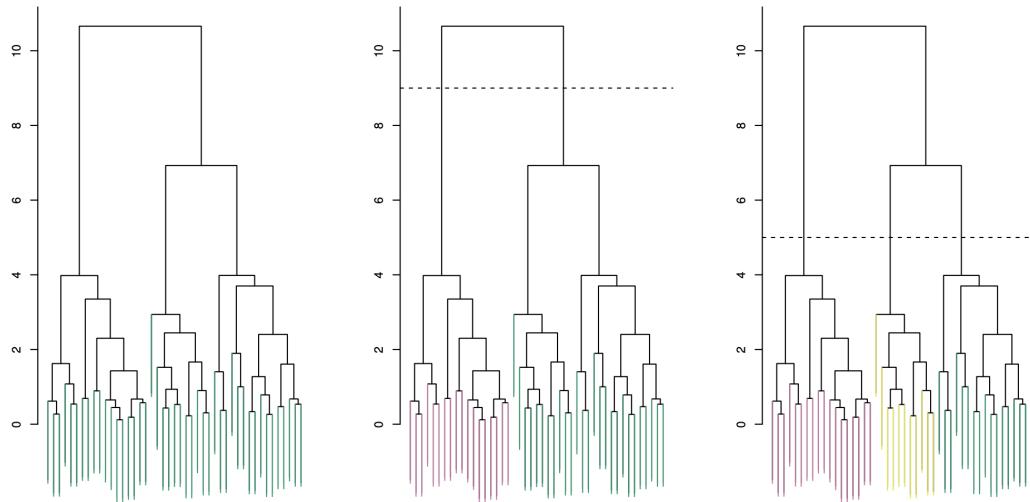


FIGURA 14. Imagen tomada de [1].

- [2] V. A. Prisacariu and I. D. Reid. PWP3D: Real-Time Segmentation and Tracking of 3D Objects. *International Journal of Computer Vision*, 98(3):335–354, jul 2012. ISSN 1573-1405. . 3
- [3] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, jul 2009. ISSN 0888613X. . 3