Modelos No

Supervisado:

Actividad Facebook

Javier Eduardo Jaimes Velásquez[[1]](#footnote-1) Brandon Valencia Murillo[[2]](#footnote-2)

2024-07-11

Tabla de Contenidos

[1.1 Entiendo el agrupamiento no supervisado. 1](#_Toc171591760)

[1.2 El modelo de agrupamiento 2](#_Toc171591761)

[1.3 K-means, como funciona. 6](#_Toc171591762)

[1.4 Determinar k, numero de clusters. 7](#_Toc171591763)

**1 Modelos No Supervisados - Actividad en Facebook**

Hay situación en donde el interés es encontrar patrones a nivel grupal mas allá de lo individual, y es el caso que se estudiara a lo largo del libro.

Es importante recordar que:

* las tareas de agrupamiento difieren un poco de las tareas de clasificación.
* el algoritmo de agrupamiento objeto de estudio es el llamado *k-means*.

## 1.1 Entiendo el agrupamiento no supervisado.

Este de algoritmo es de carácter no supervisado, en donde automáticamente se definen grupos (*clústeres*) de elementos similares. Es claro que estos grupos no se han definido previamente por tal razón estos algoritmos son esenciales a la hora de descubrir patrones, mas allá de hacer predicciones.

El principio principal de agrupamiento se basa en que los elementos de un grupo deben ser tan similares como se pueda de los demás, pero muy diferente del resto, es decir que en términos de definición de similitud la distancia puede ser un elemento clave.

1.1.1 Algunos ejemplos de criterios de similitud

* Agrupamiento por características sociodemográficas.
* Patrones de Comportamiento de uso de un sistema, comportamiento de compra en una plataforma de ventas.
* Tan simple como por columnas de un conjunto de datos con valores similares.

## 1.2 El modelo de agrupamiento

A diferencia de los modelos de predicción, los modelos de agrupamiento buscan crear nuevos datos a partir de la lógica definida de agrupamiento, pero en resumen el proceso no difiere de a mucho, excepto de que al super no supervisado implica que los datos se agrupan siguiendo por ejemplo heurísticas definidas para ello.

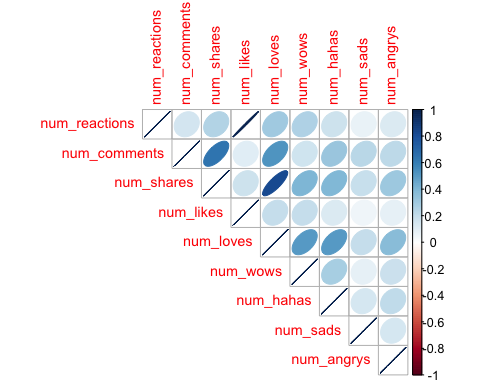
1.2.1 Descripción del conjunto de datos

Ahora se realiza un poco de limpieza sobre el conjunto de los datos:

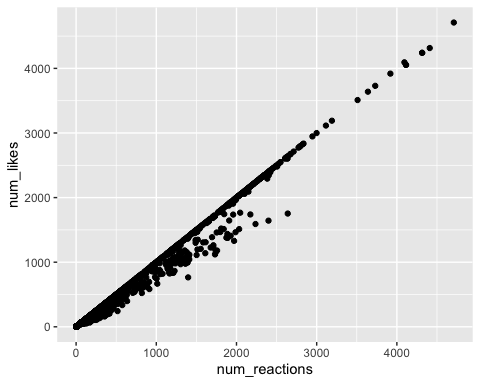
|  |  |  | Statistic (N=7,050) |
| --- | --- | --- | --- |
| num\_reactions | Mean (SD) |  | 230.1 (462.6) |
| Median (IQR) |  | 59.5 (202.0) |
| Range |  | 0.0 - 4,710.0 |
| num\_comments | Mean (SD) |  | 224.4 (889.6) |
| Median (IQR) |  | 4.0 (23.0) |
| Range |  | 0.0 - 20,990.0 |
| num\_shares | Mean (SD) |  | 40.0 (131.6) |
| Median (IQR) |  | 0.0 (4.0) |
| Range |  | 0.0 - 3,424.0 |
| num\_likes | Mean (SD) |  | 215.0 (449.5) |
| Median (IQR) |  | 58.0 (167.8) |
| Range |  | 0.0 - 4,710.0 |
| num\_loves | Mean (SD) |  | 12.7 (40.0) |
| Median (IQR) |  | 0.0 (3.0) |
| Range |  | 0.0 - 657.0 |
| num\_wows | Mean (SD) |  | 1.3 (8.7) |
| Median (IQR) |  | 0.0 (0.0) |
| Range |  | 0.0 - 278.0 |
| num\_hahas | Mean (SD) |  | 0.7 (4.0) |
| Median (IQR) |  | 0.0 (0.0) |
| Range |  | 0.0 - 157.0 |
| num\_sads | Mean (SD) |  | 0.2 (1.6) |
| Median (IQR) |  | 0.0 (0.0) |
| Range |  | 0.0 - 51.0 |
| num\_angrys | Mean (SD) |  | 0.1 (0.7) |
| Median (IQR) |  | 0.0 (0.0) |
| Range |  | 0.0 - 31.0 |

1.2.1.1 Variables duplicadas

Es importante determinar si algunas variables están correlacionadas entre sí, esto para dejarlas por fuera de nuestro modelo de agrupamiento:



Entiendo que, la variable *num\_reactions* y *num\_likes*, tienen un coeficiente de correlación de 1, haría sentido que al menos una de ella no fuera parte de nuestro modelo. Antes de eliminarlas del todo, se procede a verificar su nivel de dispersión por medio de un gráfico de variación:



Al comprobarse la correlación de forma gráfica procedemos a limpiar la variable de nuestro conjunto datos.

En el siguiente paso, estandaricemos los valores siguiendo el método z*-score*. En este proceso se buscan que todas las variables sean llevadas en una escala usando el método *scale* de R, en dicho caso la media será de y la desviación estandar de .

## 1.3 K-means, cómo funciona.

Por lo general lo algoritmos de agrupamiento se pueden dividir en dos grupos, los basados en una métrica de distancia o los basado en una función de agrupamiento. Y a su vez podemos también clasificarlos dependiendo del método, los basados en métodos jerárquicos, los basados en particiones y lo basados densidad. K-means está basado en particiones, entre sus ventajas están:

* Principios básicos que son de fácil entendimiento.
* Altamente configurable, adaptable para acomodarse a diversas situaciones.
* Su desempeño es bueno bajo la mayoría de situaciones de la vida real.

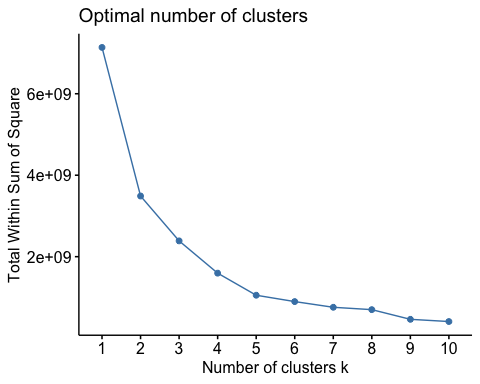
Entre sus desventajas:

* No tan sofisticado como algunos de los otros métodos
* No tiene garantía de selección en situación optima de agrupamiento.
* Requiere esfuerzos en la elección del numero , de clusters. - Y no es ideal para ciertas cosas como aquellos donde la alta variacion en la densidad del modelo es clave.

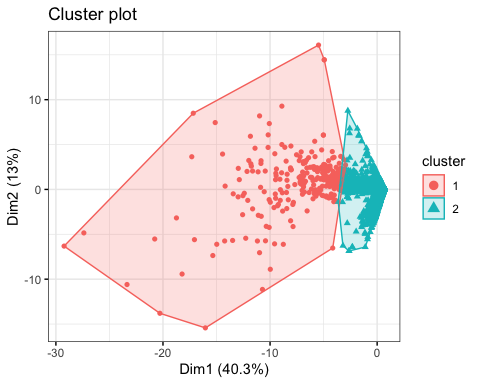
En simples términos: *k-means* asigna un número determinado de observaciones a un numero de clusters, donde fue determinado con antelacion. El algoritmo usa un proceso de heuristicas para encontrar una solucion local optima, basicamente inicia con su mejor supocision y compara diferentes agrupamientos hasta llegar al supuesto “mejor”.

## 1.4 Determinar k, numero de clústeres.

Empezaremos determinando el numero a partir del metodo del codo (*elbow*): aquí la idea es en como la homogeneidad y la heterogeneidad de los clústeres cambia para diferentes valores de :



Dado lo anterior, el numero debe ser igual a ., Con definido procedemos a crear el modelo utilizando el metodo *k-means* parte de la librería *stats*, y asignamos al dataframe *fb\_activity\_zs* el cluster asignado a cada observación. Por último visualización la agrupación creada por el método k-means:



1. Politécnico Grancolombiano, [jajaimes4@poligran.edu.co](mailto:jajaimes4@poligran.edu.co) [↑](#footnote-ref-1)
2. Politécnico Grancolombiano, [brvalencia6@poligran.edu.co](mailto:brvalencia6@poligran.edu.co) [↑](#footnote-ref-2)