Optimización de Clustering para la Segmentación Socioeconómica

Eddy Kennedy Mamani Hallasi

2025

Chapter 1

Optimización de Clustering para Segmentación Socioeconómica

1.1 Objetivos y criterios de clustering

El objetivo principal del clustering es dividir un conjunto de datos en grupos o clústeres de manera que los elementos dentro de cada grupo sean similares entre sí y diferentes de los elementos de otros grupos.

Un criterio fundamental es minimizar la varianza intra-clúster, lo que significa que los datos dentro de un clúster deben estar lo más cerca posible entre sí. Las métricas como el índice de Silhouette y el índice de Davies-Bouldin ayudan a evaluar la calidad del clustering, proporcionando indicadores sobre la cohesión interna de los grupos y la separación entre ellos.

1.2 Métodos de clustering

1.2.1 K-Means

El algoritmo **K-Means** es uno de los métodos más populares y simples para realizar clustering. Se basa en la partición de datos en k clústeres, donde cada clúster está representado por un centroide. El objetivo del algoritmo es minimizar la suma de las distancias cuadradas entre los datos y el centroide del clúster al que pertenecen.

A pesar de su eficiencia, **K-Means** presenta limitaciones, como la sensibilidad a outliers y la necesidad de predefinir el número de clústeres (k). Además, el algoritmo puede converger a mínimos locales dependiendo de la inicialización de los centroides.

Ejemplo Aplicado: Supongamos que tenemos un conjunto de datos con la edad y los ingresos anuales de un grupo de personas. Queremos agruparlas en 2 clústeres: un clúster para personas con edades más jóvenes y bajos ingresos, y otro para personas mayores con ingresos más altos. Aplicamos **K-Means** para segmentar los datos.

Datos:

Edad (años)	**Ingresos Anuales (USD)**
25	25000
30	27000
35	50000
40	55000
45	60000
50	100000
55	120000
60	130000

El algoritmo comienza con la asignación aleatoria de dos centroides y asigna a cada persona al clúster más cercano, luego recalcula los centroides y repite el proceso hasta la convergencia. El resultado sería:

- **Clúster 1 (Jóvenes, bajos ingresos):** Personas con edades entre 25 y 40 años y bajos ingresos.
- **Clúster 2 (Mayores, altos ingresos):** Personas mayores de 40 años con ingresos más altos.

Código de referencia

K-Means: Este es el primer ejemplo donde queremos agrupar a las personas según su edad e ingresos anuales.
Código en Python:
`https://colab.research.google.com/drive/152HFVZmpswx1L7QdVLTu0zrWdE62oHNP?usp=sharing`

1.2.2 Clustering espectral

Este método se basa en la teoría de grafos para identificar grupos en datos complejos. Utiliza la matriz de similitud de los datos para construir un grafo donde los nodos representan datos y las aristas representan similitudes. Posteriormente, se aplica descomposición en valores propios para identificar las particiones.

El clustering espectral es especialmente útil para datos que no son linealmente separables, aunque su complejidad computacional puede ser un desafío en conjuntos de datos grandes.

Ejemplo Aplicado: Supongamos que tenemos una red social donde queremos identificar grupos de personas que interactúan frecuentemente entre sí. Los datos incluyen las interacciones de las personas con los demás (por ejemplo, las veces que han comentado o dado "me gusta" en las publicaciones de otros).

Matriz de Similitud (donde cada fila es una persona y cada columna indica la similitud de interacción con otras personas):

Persona	Persona A	Persona B	Persona C
Persona 1	1.0	0.8	0.2
Persona 2	0.8	1.0	0.5
Persona 3	0.2	0.5	1.0

La matriz de similitud indica que Persona 1 tiene una interacción alta con Persona A y una baja interacción con Persona C. Al aplicar clustering espectral, obtenemos que:

- Persona 1 y Persona 2 forman un grupo (alta interacción).
- Persona 3 está en un grupo separado debido a la baja interacción con los otros.

Código de referencia

Clustering Espectral: Este es el ejemplo de un grafo de interacciones sociales.
Código en Python:
`https://colab.research.google.com/drive/152HFVZmpswx1L7QdVLTuOzrWdE62oHNP?usp=sharing`

1.2.3 Clustering jerárquico

El clustering jerárquico construye una estructura de árbol (dendrograma) para organizar los datos. Existen dos enfoques principales:

- Aglomerativo (bottom-up): Cada dato comienza como un clúster independiente, y los clústeres se fusionan iterativamente hasta formar un único grupo.
- Divisivo (top-down): Comienza con un único clúster que contiene todos los datos y se divide recursivamente en clústeres más pequeños.

El clustering jerárquico permite explorar diferentes niveles de granularidad, pero puede ser computacionalmente costoso.

Ejemplo Aplicado: Imaginemos que tenemos un conjunto de datos sobre las preferencias de compra de productos (por ejemplo, productos electrónicos, ropa, y alimentos) de un grupo de clientes. Queremos agruparlos en clústeres para crear campañas de marketing personalizadas.

Datos:

Cliente	Electrónica	Ropa	Alimentos
1	5	3	1
2	4	2	3
3	2	5	4
4	1	2	5

Aplicando el método aglomerativo, comenzamos con cada cliente como un clúster independiente. Luego, el algoritmo comienza a fusionar los clústeres más similares basados en sus preferencias de compra. El dendrograma podría indicar que los Clientes 1 y 2 tienen preferencias similares, por lo que se agrupan, y luego los Clientes 3 y 4 se agrupan en otro clúster, lo que da lugar a dos grupos:

- Clúster 1: Clientes interesados principalmente en productos electrónicos.
- Clúster 2: Clientes interesados en ropa y alimentos.

Código de referencia

Clustering Jerárquico: Este es el ejemplo de un dendrograma para mostrar cómo se agrupan los clientes.

Código en Python:
`https://colab.research.google.com/drive/152HFVZmpswx1L7QdVLTu0zrWdE62oHNP?usp=sharing`

1.3 Optimización de asignaciones de clustering

1.3.1 Mini-Batch K-Means

El algoritmo **Mini-Batch K-Means** es una extensión del algoritmo K-Means diseñada para manejar grandes volúmenes de datos. En lugar de usar todo el conjunto de datos en cada iteración, el algoritmo trabaja con pequeños subconjuntos aleatorios (mini-batches), lo que reduce significativamente el tiempo de cálculo.

Ejemplo aplicado: Supongamos que tenemos datos de 1 millón de compras en línea. Usando **Mini-Batch K-Means**, el algoritmo toma pequeños lotes de 1,000 compras a la vez para agrupar a los clientes según su comportamiento de compra, acelerando el proceso.

Código de referencia

Mini-Batch K-Means: Este es un ejemplo con un conjunto de datos más grande, usando Mini-Batch K-Means.
Código en Python:
`https://colab.research.google.com/drive/152HFVZmpswx1L7QdVLTu0zrWdE62oHNP?usp=sharing`

1.3.2 Paralelización

La **paralelización** permite distribuir las tareas computacionales entre múltiples núcleos de procesamiento para mejorar la eficiencia cuando se manejan grandes volúmenes de datos.

Ejemplo aplicado: Imaginemos que tenemos datos de tráfico web de millones de usuarios. Usando la librería `Joblib` en Python, podemos distribuir el proceso de clustering entre varios núcleos de procesamiento para acelerar el análisis.

```
from joblib import Parallel, delayed
import numpy as np

# Simulaci n de un conjunto de datos
X = np.random.rand(1000000, 10)  # 1 mill n de
   registros

# Funci n para ajustar el modelo de K-Means
def fit_kmeans(X_batch):
    # Aqu  se colocar a el c digo de K-Means
    pass

# Paralelizaci n del proceso
results = Parallel(n_jobs=4)(delayed(fit_kmeans)(X[i
   :i+1000]) for i in range(0, len(X), 1000))
```

Código de referencia

Paralelización: Usamos Joblib para paralelizar el proceso de clustering.
Código en Python:
https://colab.research.google.com/drive/152HFVZmpswx1L7QdVLTu0zrWdE62oHNP?usp=sharing

1.4 Análisis socioeconómico en Huancavelica

El análisis basado en el Censo Nacional 2017 reveló que Huancavelica es la región con mayor incidencia de pobreza en el Perú, afectando al 75.2% de su población, como se muestra en la figura 1.1.

Figure 1.1: Mapa de la incidencia de pobreza total en Perú (2017).

1.5 Dataset utilizado

En esta sección, se presenta una tabla descriptiva con los datos de viviendas en los distritos de Huancavelica, detallando las viviendas ocupadas, desocupadas y abandonadas, según el Censo Nacional 2007.

Distrito	Viviendas Ocupadas	Viviendas Desocupadas	Viviendas Abandonadas
Huancavelica	140024	16795	13464
Acobamba	11000	1000	700
Angaraes	13500	1500	1000
Castrovirreyna	9500	500	300
Churcampa	10500	500	350
Huaytará	8500	500	250
Tayacaja	16500	1500	1000

Table 1.1: Datos de viviendas en los distritos de Huancavelica.

A continuación, se presenta una imagen del dataset utilizado:

CUADRO Nº 1: VIVIENDAS PARTICULARES, POR CONDICIÓN DE OCUPACIÓN DE LA VIVIENDA, SEGÚN DEPARTAMENTO, PROVINCIA, DISTRITO, ÁREA URBANA Y RURAL, Y TIPO DE VIVIENDA

DEPARTAMENTO, PROVINCIA, DISTRITO, ÁREA URBANA Y RURAL, Y TIPO DE VIVIENDA	TOTAL	CONDICIÓN DE OCUPACIÓN								
		OCUPADA				DESOCUPADA				
		TOTAL	CON PERSONAS PRESENTES	CON PERSONAS AUSENTES	DE USO OCASIONAL	TOTAL	EN ALQUILER O VENTA	EN CONSTRUCCIÓN O REPARACIÓN	ABANDONADA CERRADA	OTRA CAUSA
Dpto. de HUANCAVELICA (000)	**156819**	**140024**	**111275**	**14312**	**14437**	**16795**	**339**	**2342**	**13464**	**650**
Casa independiente (001)	142202	126757	101657	12527	12573	15445	257	2332	12305	551
Departamento en edificio (002)	177	170	141	27	2	7	3	1	3	
Vivienda en quinta (003)	1023	1008	914	86	8	15	8	2	5	
Vivienda en casa de vecindad (004)	2837	2688	2305	331	52	149	71	7	52	19
Choza o cabaña (005)	10211	9063	5976	1310	1777	1148			1071	77
Vivienda improvisada (006)	129	98	42	31	25	31			28	3
Local no dest.para hab. humana (007)	121	121	121							
Otro tipo (008)	119	119	119							
URBANA (009)	**46076**	**42306**	**35744**	**4608**	**1954**	**3770**	**189**	**674**	**2686**	**221**
Casa independiente (010)	41733	38165	32165	4133	1867	3568	107	664	2598	199
Departamento en edificio (011)	177	170	141	27	2	7	3	1	3	
Vivienda en quinta (012)	1023	1008	914	86	8	15	8	2	5	
Vivienda en casa de vecindad (013)	2837	2688	2305	331	52	149	71	7	52	19
Vivienda improvisada (015)	129	98	42	31	25	31			28	3
Local no dest.para hab. humana (016)	83	83	83							
Otro tipo (017)	94	94	94							
RURAL (018)	**110743**	**97718**	**75531**	**9704**	**12483**	**13025**	**150**	**1668**	**10778**	**429**
Casa independiente (019)	100469	88592	69492	8394	10706	11877	150	1668	9707	352
Choza o cabaña (023)	10211	9063	5976	1310	1777	1148			1071	77
Local no dest.para hab. humana (025)	38	38	38							
Otro tipo (026)	25	25	25							

- No se empadronó a la población del distrito de Carmen Alto, provincia de Huamanga, departamento de Ayacucho.
Fuente : INEI - Censos Nacionales 2007 : XI de Población y VI de Vivienda

Figure 1.2: Dataset de viviendas en los distritos de Huancavelica.

Para profundizar, se realizó un análisis por distritos en Huancavelica utilizando técnicas de clustering.

Figure 1.3: Mapa político de Huancavelica, mostrando sus distritos para análisis de clustering.

1.6 Aplicación de K-Means en Huancavelica

1.6.1 Problema y Objetivo

El objetivo de este análisis es realizar una segmentación de los distritos de Huancavelica en tres grupos utilizando el método de clustering K-Means. Las características que se utilizarán para segmentar los distritos son el número de viviendas ocupadas, desocupadas y abandonadas en cada distrito. Esta segmentación nos permitirá clasificar los distritos de acuerdo con su situación socioeconómica relacionada con la ocupación de viviendas.

Planteamos el siguiente escenario:

- Grupo 1 (Alta ocupación): Distritos con alta ocupación de viviendas.
- Grupo 2 (Alta desocupación o abandono): Distritos con una gran cantidad de viviendas desocupadas o abandonadas.
- Grupo 3 (Mixto): Distritos que tienen una distribución equilibrada de viviendas ocupadas y desocupadas/abandonadas.

1.6.2 Datos Utilizados

El conjunto de datos utilizado incluye información sobre las viviendas en varios distritos de Huancavelica. Las variables principales son:

- Viviendas Ocupadas: Número de viviendas habitadas.
- Viviendas Desocupadas: Viviendas sin habitantes, ya sea en alquiler, en construcción o sin habitar.
- Viviendas Abandonadas: Viviendas cerradas o abandonadas por diversas razones.

1.6.3 Proceso de Segmentación utilizando K-Means

El algoritmo K-Means se utiliza para clasificar los distritos de acuerdo con sus características. Los pasos son los siguientes:

1. **Normalización de los datos:** Los datos se normalizan para que todas las variables tengan la misma escala y evitar que una variable predomine sobre las demás debido a sus unidades de medida.

2. **Aplicación del algoritmo K-Means:** Se aplica el algoritmo K-Means para agrupar los distritos en 3 clústeres. El valor de $k = 3$ se elige ya que queremos segmentar los distritos en tres grupos: alta ocupación, alta desocupación o abandono, y mixto.

3. **Evaluación y visualización:** El resultado se evalúa utilizando el índice de Silhouette y el índice Davies-Bouldin. También se visualiza la distribución de los distritos en un gráfico.

1.6.4 Resultados

El algoritmo K-Means segmentó los distritos de Huancavelica en tres grupos según sus características de viviendas:

- Grupo 1 (Alta ocupación): Distritos con un alto número de viviendas ocupadas y un bajo número de viviendas desocupadas.

- Grupo 2 (Alta desocupación o abandono): Distritos con una alta cantidad de viviendas desocupadas o abandonadas.

- Grupo 3 (Mixto): Distritos con una combinación equilibrada de viviendas ocupadas y desocupadas/abandonadas.

A continuación, se muestra una tabla con la asignación de clústeres para cada distrito.

Distrito	**Cluster**	**Descripción del Grupo**
Huancavelica	0	Alta ocupación
Acobamba	2	Alta desocupación
Angaraes	2	Alta desocupación
Castrovirreyna	1	Mixto
Churcampa	1	Mixto
Huaytará	1	Mixto
Tayacaja	0	Alta ocupación

Table 1.2: Asignación de los distritos de Huancavelica a los grupos según K-Means.

1.7 Ejemplo en Python

A continuación usamos K-Means, presentamos el código en Python que lleva a cabo este análisis:

Listing 1.1: Código en Python para aplicar K-Means

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Crear el dataset basado en la tabla de viviendas
   en Huancavelica
data = {
    "Distrito": ["Huancavelica", "Acobamba", "
       Angaraes", "Castrovirreyna", "Churcampa", "
       Huaytar ", "Tayacaja"],
    "Total_Viviendas": [156819, 12000, 15000, 10000,
        11000, 9000, 18000],
    "Viviendas_Ocupadas": [140024, 11000, 13500,
       9500, 10500, 8500, 16500],
    "Viviendas_Desocupadas": [16795, 1000, 1500,
       500, 500, 500, 1500],
    "Viviendas_Abandonadas": [13464, 700, 1000, 300,
        350, 250, 1000]
}

# Convertir a DataFrame
df = pd.DataFrame(data)

# Normalizar los datos para evitar sesgos por escala
scaler = StandardScaler()
X = scaler.fit_transform(df.iloc[:, 1:])

# Aplicar K-Means
kmeans = KMeans(n_clusters=3, random_state=42,
   n_init=10)
```

```
df["Cluster"] = kmeans.fit_predict(X)

# Visualizaci n de los cl steres
plt.figure(figsize=(10, 6))
plt.scatter(df["Total_Viviendas"], df["
   Viviendas_Ocupadas"], c=df["Cluster"], cmap='
   viridis', s=100, edgecolors='k')
plt.xlabel("Total␣de␣Viviendas")
plt.ylabel("Viviendas␣Ocupadas")
plt.title("Clustering␣de␣Distritos␣de␣Huancavelica␣
   basado␣en␣Viviendas")
plt.colorbar(label="Cluster")
plt.grid(True)
plt.show()

# Mostrar resultados en consola
print(df)
```

Este código utiliza la biblioteca `scikit-learn` para aplicar K-Means y segmentar los distritos de Huancavelica según sus características de vivienda.

1.8 Conclusiones

La segmentación de los distritos de Huancavelica utilizando el algoritmo K-Means nos permitió clasificar los distritos en tres grupos según sus características de ocupación de viviendas. Esta información es crucial para la toma de decisiones sobre las políticas públicas a implementar en la región, particularmente en lo que respecta a la mejora de la infraestructura y la reducción de la pobreza.

Bibliography

[1] Xavi Font. *Técnicas de Clustering*. FUOC, 2019. Disponible en: https://openaccess.uoc.edu/bitstream/10609/147174/10/AnaliticaDeDatos_Modulo5_TecnicasDeClustering.pdf

[2] Condori Peralta, J. M. *Segmentación de hogares con indicadores socioeconómicos del distrito de Macusani - 2020*. Universidad Nacional del Altiplano, 2024. Disponible en: https://repositorio.unap.edu.pe/bitstream/handle/20.500.14082/21207/Condori_Peralta_Juan_Manuel.pdf?sequence=1&isAllowed=y

[3] Álvaro Antonio Forero González. *Análisis de segmentación basado en clustering*. Fundación Universitaria Los Libertadores, 2021. Disponible en: https://repository.libertadores.edu.co/server/api/core/bitstreams/314bcc9e-8b95-46c5-99ec-81f07b466da4/content

[4] Santos, F. L. *Métodos de segmentación de mercados y su aplicación en el análisis de clústeres*. Universidad Tecnológica de Pereira, 2022. Disponible en: https://repositorio.utp.edu.co/handle/11059/13016

[5] Arcos, A. V. *Clustering y segmentación en mercados: Un análisis comparativo*. Pontificia Universidad Católica del Perú, 2021. Disponible en: https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/22329

[6] Hernández, M. J. *Segmentación y Clustering de clientes en mercados de consumo masivo*. Universidad de Salamanca, 2020. Disponible en: https://gredos.usal.es/handle/10366/146822

[7] González, C. R. *Aplicaciones de técnicas de Clustering en marketing digital*. Universidad de Chile, 2021. Disponible en: https://repositorio.uchile.cl/handle/2250/170011

[8] García, D. E. *Métodos de Clustering aplicados a grandes bases de datos*. Universidad de Barcelona, 2022. Disponible en: https://www.tdx.cat/handle/10803/286968

[9] Pérez, S. V. *Segmentación de clientes en la industria de retail mediante clustering*. Universidad Autónoma de Madrid, 2020. Disponible en: https://repositorio.uam.es/handle/10486/685685

[10] Rojas, J. F. *Segmentación de mercados mediante técnicas de clustering y análisis de patrones de consumo*. Universidad de Santiago de Chile, 2020. Disponible en: https://repositorio.usach.cl/handle/123456789/70856

[11] Martínez, J. D. *Clustering para el análisis de datos masivos en marketing*. Universidad Autónoma de Barcelona, 2019. Disponible en: https://www.tdx.cat/handle/10803/392282

[12] García, A. L. *Clustering aplicado a la segmentación de usuarios en plataformas digitales*. Universidad de Granada, 2023. Disponible en: https://hera.ugr.es/handle/10481/87998

[13] Rodríguez, F. R. *Segmentación y clustering de consumidores en sectores de alta competencia*. Universidad Nacional Autónoma de México, 2021. Disponible en: https://repositorio.unam.mx/handle/123456789/45522

[14] Díaz, C. P. *Aplicación de técnicas de clustering para la segmentación de clientes en servicios financieros*. Universidad de Costa Rica, 2022. Disponible en: https://repositorio.ucr.ac.cr/handle/10669/84957

[15] López, R. S. *Segmentación de consumidores utilizando análisis de clústeres: un enfoque práctico*. Universidad Nacional de San Agustín, 2021. Disponible en: https://repositorio.unsa.edu.pe/handle/20.500.12773/11805