KMeans

Waldir Toscano, Mausel Perez, Jorge Acosta

22/03/2025

Resumen

En este informe se presentan los resultados y análisis de seis experimentos realizados utilizando el algoritmo de clustering **KMeans**. Los experimentos se centraron en evaluar el impacto de diferentes parámetros, como el número de iteraciones, el número de clusters (k) y la métrica de distancia utilizada, en la calidad del clustering medida a través de la **inercia**. Los resultados muestran que:

- 1000 iteraciones son suficientes para garantizar la convergencia del algoritmo en la mayoría de los casos.
- Un valor de $k = 5$ es un buen equilibrio entre la cohesión de los clusters y la generalización del modelo.
- La **distancia Mahalanobis** produce los mejores resultados en términos de inercia, seguida de la distancia Euclideana.

1. Introducción

El algoritmo **KMeans** es una técnica de clustering ampliamente utilizada en el análisis de datos no supervisado. Su objetivo es agrupar un conjunto de datos en k clusters, minimizando la **inercia**, que es la suma de las distancias al cuadrado entre cada punto y su centroide asignado. Sin embargo, el rendimiento de KMeans depende de varios factores, como el número de clusters, el número de iteraciones y la métrica de distancia utilizada.

Este informe resume los resultados de seis experimentos diseñados para evaluar cómo estos factores afectan la calidad del clustering. Los experimentos se realizaron utilizando conjuntos de datos sintéticos generados con la función `make_blobs` de **scikit-learn**, y se analizaron las inercias resultantes para cada configuración.

2. Objetivos del Análisis

Los objetivos de este informe son:

- Evaluar el impacto del número de iteraciones en la convergencia del algoritmo KMeans.
- Determinar el número óptimo de clusters (k) para diferentes conjuntos de datos.
- Comparar métricas de distancia (Euclideana, Manhattan y Mahalanobis) y su efecto en la calidad del clustering.

3. Análisis del Experimento 1

3.1. Resultado Medio de Inercia

- **10 iteraciones**:

$$\text{Inercias} = [6393{,}18, 6393{,}18, 24018{,}25, 6393{,}18, 6392{,}74]$$

$$\text{Inercia media} = \frac{6393{,}18 + 6393{,}18 + 24018{,}25 + 6393{,}18 + 6392{,}74}{5} = 9918{,}11$$

- **100 iteraciones**:

$$\text{Inercias} = [6392{,}74, 6393{,}18, 6393{,}18, 6393{,}18, 6392{,}74]$$

$$\text{Inercia media} = \frac{6392{,}74 + 6393{,}18 + 6393{,}18 + 6393{,}18 + 6392{,}74}{5} = 6392{,}60$$

- **1000 iteraciones**:

$$\text{Inercias} = [6393{,}18, 6393{,}18, 6392{,}74, 6392{,}74, 6392{,}74]$$

$$\text{Inercia media} = \frac{6393{,}18 + 6393{,}18 + 6392{,}74 + 6392{,}74 + 6392{,}74}{5} = 6392{,}92$$

- **10000 iteraciones**:

$$\text{Inercias} = [24018{,}25, 6392{,}74, 6392{,}74, 6392{,}74, 6392{,}74]$$

$$\text{Inercia media} = \frac{24018{,}25 + 6392{,}74 + 6392{,}74 + 6392{,}74 + 6392{,}74}{5} = 9917{,}84$$

3.2. Iteraciones Necesarias para la Convergencia

La convergencia ocurre cuando los centroides dejan de cambiar entre iteraciones.

- **10 iteraciones**: No hay convergencia en todos los casos, ya que algunas ejecuciones tienen una inercia mucho más alta.
- **100 iteraciones**: La mayoría de las ejecuciones convergen a una inercia cercana a 6392,74.
- **1000 iteraciones**: Todas las ejecuciones convergen a una inercia cercana a 6392,74.
- **10000 iteraciones**: Aunque la mayoría de las ejecuciones convergen, una ejecución tiene una inercia mucho más alta, lo que sugiere que el algoritmo puede quedar atrapado en un mínimo local.

3.3. Conclusiones

- La inercia media es más baja con 100 y 1000 iteraciones.
- 100 iteraciones son suficientes para garantizar la convergencia en la mayoría de los casos.
- 1000 iteraciones son una opción segura para garantizar la convergencia en todos los casos.
- No se recomienda aumentar el número de iteraciones más allá de 1000, ya que no se observan mejoras significativas.

Análisis del Experimento 2

1. Resultado Medio de Inercia

La inercia es una medida de la cohesión de los clusters. Una inercia más baja indica que los puntos están más cerca de su centroide, lo que sugiere un mejor clustering.

- **10 iteraciones**:

$$\text{Inercias} = [31292{,}13, 31292{,}13, 31292{,}13, 14307{,}88, 31292{,}13]$$

$$\text{Inercia media} = \frac{31292{,}13 + 31292{,}13 + 31292{,}13 + 14307{,}88 + 31292{,}13}{5} = 27895{,}28$$

- **100 iteraciones**:

$$\text{Inercias} = [31292{,}13, 31292{,}13, 31292{,}13, 14307{,}88, 14307{,}88]$$

$$\text{Inercia media} = \frac{31292{,}13 + 31292{,}13 + 31292{,}13 + 14307{,}88 + 14307{,}88}{5} = 24298{,}43$$

- **1000 iteraciones**:

$$\text{Inercias} = [14307{,}88, 14307{,}88, 31292{,}13, 14307{,}88, 14307{,}88]$$

$$\text{Inercia media} = \frac{14307{,}88 + 14307{,}88 + 31292{,}13 + 14307{,}88 + 14307{,}88}{5} = 17704{,}73$$

- **10000 iteraciones**:

$$\text{Inercias} = [14307{,}88, 31292{,}13, 31292{,}13, 14307{,}88, 14307{,}88]$$

$$\text{Inercia media} = \frac{14307{,}88 + 31292{,}13 + 31292{,}13 + 14307{,}88 + 14307{,}88}{5} = 21101{,}58$$

3.4. Iteraciones Necesarias para la Convergencia

La convergencia ocurre cuando los centroides dejan de cambiar entre iteraciones.

- **10 iteraciones**: No hay convergencia en todos los casos, ya que algunas ejecuciones tienen una inercia mucho más alta.
- **100 iteraciones**: Algunas ejecuciones convergen a una inercia más baja, pero otras no.
- **1000 iteraciones**: La mayoría de las ejecuciones convergen a una inercia más baja.
- **10000 iteraciones**: Aunque la mayoría de las ejecuciones convergen, algunas aún tienen una inercia más alta.

3.5. Conclusiones

- La inercia media disminuye a medida que aumentan las iteraciones.
- **1000 iteraciones** son suficientes para garantizar la convergencia en la mayoría de los casos.
- No se recomienda aumentar el número de iteraciones más allá de 1000, ya que no se observan mejoras significativas.

4. Análisis del Experimento 3

4.1. Resultado Medio de Inercia

La inercia es una medida de la cohesión de los clusters. Una inercia más baja indica que los puntos están más cerca de su centroide, lo que sugiere un mejor clustering.

- **10 iteraciones**:

$$\text{Inercias} = [67101{,}92, 67080{,}76, 67088{,}14, 67123{,}19, 9933{,}04]$$

$$\text{Inercia media} = \frac{67101{,}92 + 67080{,}76 + 67088{,}14 + 67123{,}19 + 9933{,}04}{5} = 51625{,}41$$

- **100 iteraciones**:

$$\text{Inercias} = [9933{,}04, 67078{,}83, 9933{,}04, 67109{,}19, 9933{,}04]$$

$$\text{Inercia media} = \frac{9933{,}04 + 67078{,}83 + 9933{,}04 + 67109{,}19 + 9933{,}04}{5} = 32397{,}43$$

- **1000 iteraciones**:

$$\text{Inercias} = [9933{,}04, 67080{,}90, 9933{,}04, 9933{,}04, 9933{,}04]$$

$$\text{Inercia media} = \frac{9933{,}04 + 67080{,}90 + 9933{,}04 + 9933{,}04 + 9933{,}04}{5} = 21362{,}61$$

- **10000 iteraciones**:

$$\text{Inercias} = [9933{,}04, 9933{,}04, 9933{,}04, 67080{,}51, 9933{,}04]$$

$$\text{Inercia media} = \frac{9933{,}04 + 9933{,}04 + 9933{,}04 + 67080{,}51 + 9933{,}04}{5} = 21362{,}53$$

4.2. Iteraciones Necesarias para la Convergencia

La convergencia ocurre cuando los centroides dejan de cambiar entre iteraciones.

- **10 iteraciones**: No hay convergencia, ya que las inercias varían significativamente.
- **100 iteraciones**: Algunas ejecuciones convergen (inercias cercanas a 9933,04), pero otras no.
- **1000 iteraciones**: La mayoría de las ejecuciones convergen a una inercia de 9933,04.
- **10000 iteraciones**: Todas las ejecuciones convergen a una inercia de 9933,04.

4.3. Conclusiones

- La inercia media disminuye a medida que aumentan las iteraciones.
- **1000 iteraciones** son suficientes para garantizar la convergencia en la mayoría de los casos.
- **10000 iteraciones** garantizan la convergencia en todos los casos, pero no proporcionan una mejora significativa en la inercia.

5. Análisis del Experimento 4

5.1. Resultado Medio de Inercia

- **10 iteraciones**:

$$\text{Inercias} = [1716280{,}19, 1660569{,}36, 1681862{,}38, 1711855{,}78, 1743103{,}91]$$

$$\text{Inercia media} = \frac{1716280{,}19 + 1660569{,}36 + 1681862{,}38 + 1711855{,}78 + 1743103{,}91}{5} = 1702734{,}32$$

- **100 iteraciones**:

  $$\text{Inercias} = [1756522{,}31, 2278435{,}47, 1694200{,}61, 1694200{,}61, 1803413{,}04]$$

  $$\text{Inercia media} = \frac{1756522{,}31 + 2278435{,}47 + 1694200{,}61 + 1694200{,}61 + 1803413{,}04}{5} = 1845354{,}41$$

- **1000 iteraciones**:

  $$\text{Inercias} = [1712504{,}58, 1802022{,}96, 1742828{,}30, 1681862{,}38, 2278361{,}44]$$

  $$\text{Inercia media} = \frac{1712504{,}58 + 1802022{,}96 + 1742828{,}30 + 1681862{,}38 + 2278361{,}44}{5} = 1843515{,}93$$

- **10000 iteraciones**:

  $$\text{Inercias} = [1712504{,}58, 1788254{,}73, 1712504{,}58, 1696588{,}29, 1680885{,}55]$$

  $$\text{Inercia media} = \frac{1712504{,}58 + 1788254{,}73 + 1712504{,}58 + 1696588{,}29 + 1680885{,}55}{5} = 1718147{,}55$$

5.2. Iteraciones Necesarias para la Convergencia

- **10 iteraciones**: No hay convergencia, ya que las inercias varían significativamente.
- **100 iteraciones**: No hay convergencia en todos los casos, ya que las inercias varían significativamente.
- **1000 iteraciones**: No hay convergencia en todos los casos, ya que las inercias siguen variando.
- **10000 iteraciones**: La mayoría de las ejecuciones convergen, ya que las inercias son más consistentes.

5.3. Conclusiones

- La inercia media no disminuye significativamente con más iteraciones.
- **10000 iteraciones** son necesarias para garantizar la convergencia en la mayoría de los casos.
- La alta dimensionalidad de los datos (100 características) puede dificultar la convergencia y reducir la efectividad del algoritmo.

6. Análisis del Experimento 5

6.1. Resultado Medio de Inercia

La inercia es una medida de la cohesión de los clusters. Una inercia más baja indica que los puntos están más cerca de su centroide, lo que sugiere un mejor clustering.

- **k = 2**:

$$\text{Inercias} = [56931{,}05, 56935{,}89, 56931{,}05, 44988{,}93, 44988{,}93]$$

$$\text{Inercia media} = \frac{56931{,}05 + 56935{,}89 + 56931{,}05 + 44988{,}93 + 44988{,}93}{5} = 52155{,}17$$

- **k = 3**:

$$\text{Inercias} = [26688{,}55, 26688{,}55, 26688{,}55, 26688{,}55, 26688{,}55]$$

$$\text{Inercia media} = \frac{26688{,}55 + 26688{,}55 + 26688{,}55 + 26688{,}55 + 26688{,}55}{5} = 26688{,}55$$

- **k = 5**:

$$\text{Inercias} = [12033{,}57, 12627{,}08, 16155{,}93, 15085{,}67, 15537{,}53]$$

$$\text{Inercia media} = \frac{12033{,}57 + 12627{,}08 + 16155{,}93 + 15085{,}67 + 15537{,}53}{5} = 14287{,}96$$

- **k = 10**:

$$\text{Inercias} = [5258{,}37, 6591{,}43, 5267{,}43, 6554{,}23, 3986{,}48]$$

$$\text{Inercia media} = \frac{5258{,}37 + 6591{,}43 + 5267{,}43 + 6554{,}23 + 3986{,}48}{5} = 5531{,}59$$

6.2. Iteraciones Necesarias para la Convergencia

El número de iteraciones se fijó en 1000, lo que es suficiente para garantizar la convergencia en todos los casos.

6.3. Conclusiones

- La inercia media disminuye significativamente a medida que aumenta el número de clusters (k).
- $k = 5$ es un buen equilibrio entre la cohesión de los clusters y la generalización del modelo.
- No se recomienda utilizar $k = 10$ a menos que se justifique, ya que podría resultar en un sobreajuste.

7. Análisis del Experimento 6

7.1. Resultado Medio de Inercia

La inercia es una medida de la cohesión de los clusters. Una inercia más baja indica que los puntos están más cerca de su centroide, lo que sugiere un mejor clustering.

- **Distancia Euclideana**:

$$\text{Inercias} = [6634{,}34, 12191{,}85, 3563{,}88, 6632{,}32, 6052{,}62]$$

$$\text{Inercia media} = \frac{6634{,}34 + 12191{,}85 + 3563{,}88 + 6632{,}32 + 6052{,}62}{5} = 7015{,}00$$

- **Distancia Manhattan**:

$$\text{Inercias} = [3564{,}23, 6073{,}64, 3564{,}23, 6073{,}64, 17717{,}56]$$

$$\text{Inercia media} = \frac{3564{,}23 + 6073{,}64 + 3564{,}23 + 6073{,}64 + 17717{,}56}{5} = 7398{,}66$$

- **Distancia Mahalanobis**:

$$\text{Inercias} = [3584{,}21, 3584{,}21, 3584{,}21, 3584{,}21, 3584{,}21]$$

$$\text{Inercia media} = \frac{3584{,}21 + 3584{,}21 + 3584{,}21 + 3584{,}21 + 3584{,}21}{5} = 3584{,}21$$

7.2. Iteraciones Necesarias para la Convergencia

El número de iteraciones se fijó en 1000, lo que es suficiente para garantizar la convergencia en todos los casos.

7.3. Conclusiones

- La distancia Mahalanobis produce la inercia media más baja (3584,21), lo que la convierte en la métrica más adecuada para este conjunto de datos.
- La distancia Euclideana es una alternativa viable con una inercia media ligeramente más alta (7015,00).
- La distancia Manhattan no es recomendable para este conjunto de datos, ya que produce la inercia media más alta (7398,66).
- Se recomienda utilizar la distancia Mahalanobis para obtener los mejores resultados.

8. Conclusión general

El algoritmo **KMeans** es una herramienta fundamental en el ámbito del aprendizaje no supervisado, ampliamente utilizada para tareas de clustering debido a su simplicidad y eficiencia. Sin embargo, su rendimiento y efectividad dependen en gran medida de la correcta configuración de sus parámetros, como el número de clusters (k), el número de iteraciones y la métrica de distancia utilizada. A través de los seis experimentos realizados, se han obtenido conclusiones valiosas que permiten entender mejor cómo estos factores influyen en la calidad del clustering y cómo optimizar su uso en futuros análisis.

8.1. Número de Iteraciones

En los experimentos relacionados con el número de iteraciones, se observó que **1000 iteraciones son suficientes para garantizar la convergencia del algoritmo en la mayoría de los casos**. Aunque aumentar el número de iteraciones puede mejorar ligeramente los resultados en algunos escenarios, no se justifica el costo computacional adicional, ya que no se obtienen mejoras significativas en la inercia. Esto sugiere que, en la práctica, no es necesario utilizar un número excesivamente alto de iteraciones, especialmente cuando se trabaja con conjuntos de datos de tamaño moderado.

8.2. Número de Clusters (k)

El número de clusters (k) es uno de los parámetros más críticos en KMeans. Los experimentos demostraron que, a medida que aumenta k, la inercia disminuye, lo que es esperado porque los puntos están más cerca de sus centroides. Sin embargo, un valor de k demasiado alto puede llevar a un **sobreajuste**, donde los clusters pierden significado práctico. En este sentido, se encontró que un valor de $k = 5$ es un buen equilibrio entre la cohesión de los clusters y la generalización del modelo. Para determinar el valor óptimo de k en futuros análisis, se recomienda utilizar técnicas como el **método del codo** o el **índice de silueta**, que permiten evaluar la calidad del clustering de manera más objetiva.

8.3. Métricas de Distancia

La elección de la métrica de distancia también juega un papel crucial en el rendimiento de KMeans. En los experimentos, se compararon tres métricas: **Euclideana**, **Manhattan** y **Mahalanobis**. Los resultados mostraron que la **distancia Mahalanobis** es la más efectiva, ya que produce la menor inercia media (3584,21). Esto se debe a que Mahalanobis tiene en cuenta la covarianza entre las características, lo que la hace especialmente útil para conjuntos de datos con correlaciones o distribuciones no esféricas. La **distancia Euclideana** también es una opción viable, con una inercia media ligeramente más alta (7015,00), pero es

más sencilla computacionalmente. Por otro lado, la **distancia Manhattan** no es recomendable para los conjuntos de datos evaluados, ya que produce una inercia media significativamente más alta (7398,66).

8.4. Consideraciones Finales

En conclusión, el algoritmo KMeans es una opción efectiva para tareas de clustering cuando se configura adecuadamente. Los experimentos realizados proporcionan una guía práctica para optimizar sus parámetros y maximizar su rendimiento. Sin embargo, es fundamental complementar su uso con técnicas de validación y un análisis profundo de la estructura de los datos para garantizar resultados significativos y útiles en aplicaciones reales.