Análisis de Convergencia de K-Means en Datos Sintéticos

Alan Miranda. Marlon Caviedes, Cristian Orozco
Infotep

Convergencia en Baja y Media Dimensión (Experimentos 1–4)

Objetivo y Metodología:

Se aplicó el algoritmo K-Means a 1000 muestras sintéticas para evaluar el número de iteraciones hasta la convergencia en distintas configuraciones de dimensión y número de clusters. Se realizaron 5 repeticiones para cada experimento, promediando las iteraciones obtenidas.

Configuraciones:

- **Experimento 1 (2D, $k = 3$, Euclidiana):**
  Ejemplo de iteraciones: R1=4, R2=4, R3=4, R4=8, R5=3 (Promedio $\approx$ 4.6).
- **Experimento 2 (3D, $k = 3$, Euclidiana):**
  Ejemplo de iteraciones: R1=4, R2=14, R3=4, R4=5, R5=4 (Promedio $\approx$ 6.2).
- **Experimento 3 (10D, $k = 3$, Euclidiana):**
  Se generó una gráfica de barras con los promedios de iteraciones obtenidos para distintos valores de `max_iter`: 2.80, 6.80, 4.40, 11.60.
- **Experimento 4 (100D, $k = 3$, Euclidiana):**
  Resultados: `max_iter`=10: 3.00, 100: 3.20, 1000: 4.40, 10000: 3.20 iteraciones promedio.

Conclusiones

- En 2D y 3D la convergencia es rápida (entre 3 y 7 iteraciones, en promedio).
- En 10D y 100D se observa mayor variabilidad en el número de iteraciones, lo que se atribuye a la complejidad del espacio y la aleatoriedad en la inicialización.
- Aumentar el valor de `max_iter` no necesariamente reduce el número de iteraciones, ya que la convergencia ocurre antes y depende de la posición inicial de los centroides.

Influencia de k y de la Métrica de Distancia (Experimentos 5 y 6)

Experimento 5: Variación de k en 2D

Se evaluó el efecto de variar el número de clusters en datos 2D con 1000 muestras, usando $k = 2$, $k = 3$, $k = 5$ y $k = 10$. Cada configuración se repitió 5 veces.

- $k = 2$: Iteraciones: R1=3, R2=6, R3=6, R4=4, R5=3 (Promedio ≈ 4.4).
- $k = 3$: Iteraciones: R1=13, R2=5, R3=4, R4=5, R5=7 (Promedio ≈ 6.8).
- $k = 5$: Iteraciones: R1=39, R2=9, R3=18, R4=14, R5=17 (Promedio ≈ 19.4).
- $k = 10$: Iteraciones: R1=4, R2=24, R3=19, R4=19, R5=21 (Promedio ≈ 17.4).

Experimento 6: Comparación de Métricas de Distancia en 2D ($k = 5$)

Se compararon tres métricas de distancia en 2D para $k = 5$ (5 repeticiones):

- **Euclidiana**: Iteraciones: R1=9, R2=14, R3=36, R4=27, R5=11 (Promedio ≈ 19.4).
- **Manhattan**: Iteraciones: R1=18, R2=6, R3=10, R4=13, R5=9 (Promedio ≈ 11.2).
- **Mahalanobis**: Iteraciones: R1=17, R2=8, R3=22, R4=10, R5=18 (Promedio ≈ 15).

Conclusiones

- La variación de k influye en la complejidad: al aumentar k, en general se requiere un mayor número de iteraciones, aunque existe variabilidad entre repeticiones.
- En el Experimento 6, la métrica Manhattan mostró convergencia más rápida (menor promedio de iteraciones) que Euclidiana y Mahalanobis.
- La elección de la métrica y el número de clusters tiene un impacto significativo en la convergencia y en la formación final de clusters.