Ciencia de la Computación

Estructuras de Datos Avanzadas

Docente Rosa Yuliana Paccotacya Yanque

Lab.1 Maldición de la Dimensionalidad - Distancia Euclidiana

Entregado el 25/03/2025

Briceño Quiroz Anthony Angel

Semestre VI

2025-1

Introducción

En este informe, se presentan los histogramas de las distancias euclidianas obtenidos para conjuntos de puntos generados en espacios de diferente dimensionalidad: 10, 50, 100, 500, 1000, 2000 y 5000 dimensiones. El objetivo es analizar cómo cambia la distribución de las distancias a medida que aumenta la dimensionalidad del espacio y evidenciar los efectos de la "maldición de la dimensionalidad".

Este fenómeno, descrito por Bellman (1961), se refiere a cómo en espacios de alta dimensión, las distancias entre puntos tienden a ser similares, lo que afecta el rendimiento de algoritmos basados en medidas de distancia. Esto tiene implicaciones en áreas como el aprendizaje automático, el clustering y la búsqueda de vecinos más cercanos (Aggarwal, 2001).

¿Como se generó los datos?

Para la generación de datos, se consideraron 100 puntos aleatorios en un espacio de dimensión variable, con coordenadas distribuidas uniformemente en el intervalo [0,1]. Se calcularon las distancias euclidianas entre todos los pares de puntos y se representaron mediante histogramas.

Se utilizaron los siguientes parámetros:

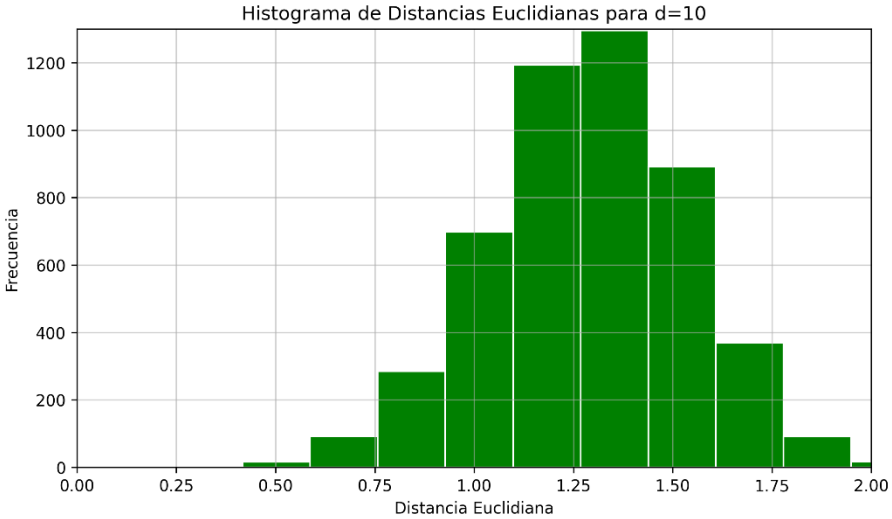
- Número de puntos: 100
- Dimensiones analizadas: 10, 50, 100, 500, 1000, 2000, 5000
- Distribución de coordenadas: Uniforme entre 0 y 1
- Métrica de distancia: Euclidiana

Los histogramas muestran cómo cambia la distribución de las distancias a medida que la dimensionalidad aumenta, lo que permite identificar patrones en la concentración de medidas y la dispersión de las distancias.

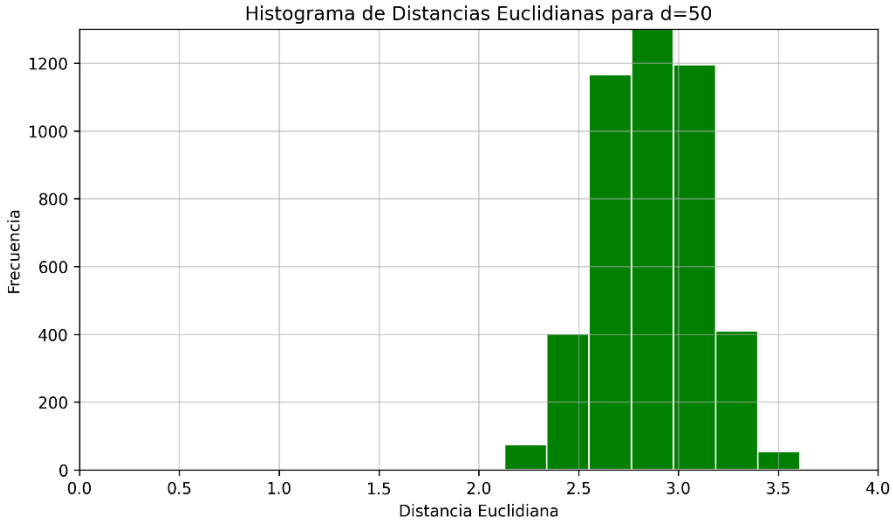
Resultados y Análisis

A continuación, se presentan los histogramas obtenidos para cada caso, junto con un análisis detallado de la distribución de las distancias:

- Dimensión 10
  - Eje X: Distancia euclidiana en el rango de 0 a 2.
  - Eje Y: Frecuencia de ocurrencia (hasta 1300).
  - En esta dimensión baja, se observa que las distancias están más dispersas y la variabilidad entre valores es alta. Se pueden encontrar distancias pequeñas y grandes con mayor frecuencia.

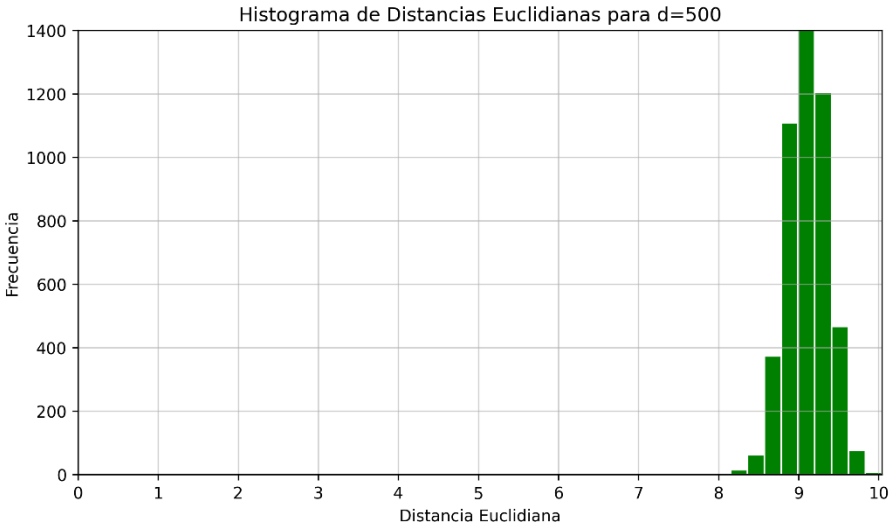


- Dimensión 50
  - Eje X: Rango de distancias hasta 4.
  - Eje Y: Frecuencia hasta 1300.
  - Se observa un desplazamiento de la distribución hacia la derecha, lo que indica que las distancias promedio han aumentado. Sin embargo, aún se mantiene una distribución relativamente dispersa.

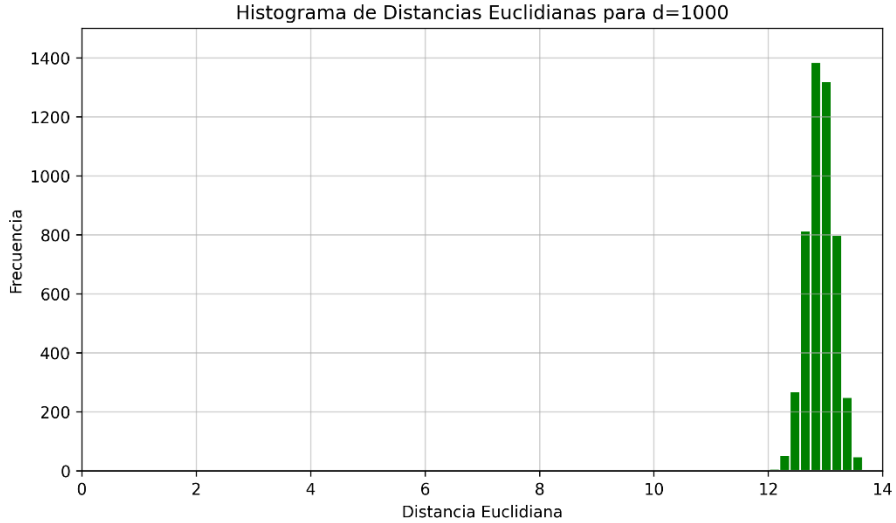


- Dimensión 100
  - Eje X: Rango de distancias hasta 5.
  - Eje Y: Frecuencia hasta 1400.
  - A medida que la dimensión crece, las distancias siguen aumentando. Ahora, la mayoría de los puntos se encuentran en un rango más alto y las distancias pequeñas son menos frecuentes.

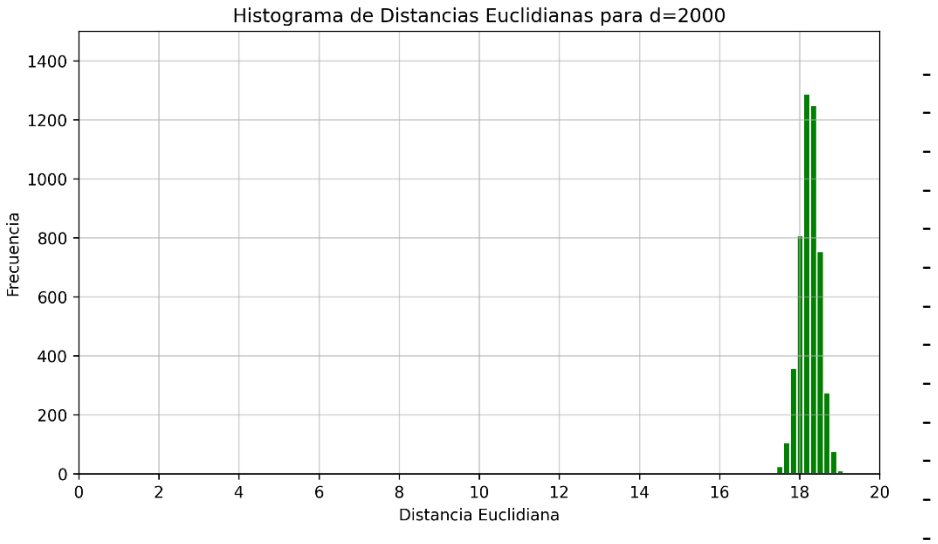
- Dimensión 500
  - Eje X: Rango de distancias hasta 10.
  - Eje Y: Frecuencia hasta 1400.
  - Se refuerza la tendencia observada, con la mayoría de las distancias en valores más altos y menor concentración en la parte baja del histograma. Esto muestra que, en dimensiones altas, los puntos están mucho más alejados entre sí.



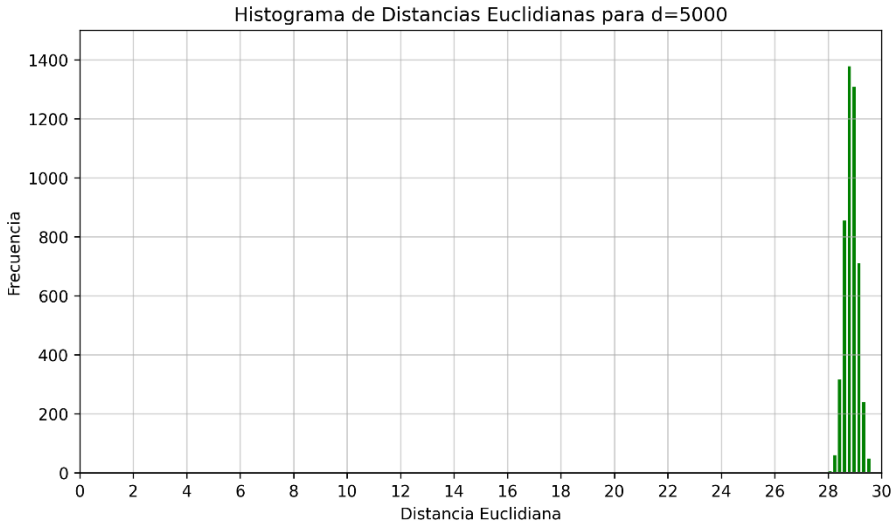
- Dimensión 1000
  - Eje X: Rango de distancias hasta 14.
  - Eje Y: Frecuencia hasta 1400.
  - La distribución sigue desplazándose a la derecha, con un pico definido y menor dispersión relativa. En esta dimensión, prácticamente todas las distancias están en un rango específico, lo que sugiere la aparición de la "concentración de la medida".



- Dimensión 2000
  - Eje X: Rango de distancias hasta 20.
  - Eje Y: Frecuencia hasta 1400.
  - Análisis: La mayoría de las distancias están concentradas en un rango estrecho, lo que indica que prácticamente todas las distancias entre puntos son similares. Este es un efecto clave de la maldición de la dimensionalidad.



- Dimensión 5000
  - Eje X: Rango de distancias hasta 30.
  - Eje Y: Frecuencia hasta 1400.
  - Análisis: En esta dimensión extremadamente alta, casi todas las distancias tienen valores similares. En aplicaciones prácticas, esto significa que la distancia euclidiana deja de ser útil para distinguir entre puntos.



Conclusiones

El análisis de los histogramas muestra claramente el efecto de la maldición de la dimensionalidad:

- Mientras más aumenta las dimensiones, las distancias promedio entre puntos también aumentan, lo que indica que los puntos están más separados. Esto se en el lado izquierdo ya que la concentración de puntos (distancias euclidianas) se va yendo a la derecha.
- En dimensiones altas, la mayoría de las distancias son similares, lo que dificulta la discriminación de puntos basada en la distancia euclidiana. Este punto en especifico me costo entenderlo un poco, pero lo entendi. ¿Como sucede esto? SI lo vemos del lado geométrico, obviamente los puntos van a estar mas dispersos a medida que haya más dimensiones, pero aca pasa algo inverso con las distancias, ya que estas, sean mínimas o máximas se vuelven casi iguales
- En espacios de alta dimensión, ocurre el fenómeno concentración de la medida, que hace que las distancias sean similares, esto tiene que ver mucho, por no decir del todo con la varianza relativa, que disminuye.
- En bajas dimensiones, la dispersión de las distancias es mayor, pero en dimensiones altas es menor
- Este fenómeno tiene implicaciones directas en áreas como el aprendizaje automático, el análisis de datos y la recuperación de información, donde muchas veces se requiere encontrar vecinos cercanos.

Referencias

- https://albertolumbreras.net/posts/maldicion-dimensionalidad.html
- https://www.datacamp.com/es/tutorial/euclidean-distance
- https://github.com/numpy/numpy
- https://numpy.org/doc/stable/user/basics.html
- https://en.cppreference.com/w/cpp/numeric/random/random_device
- https://en.cppreference.com/w/cpp/numeric/random/generate_canonical
- https://en.cppreference.com/w/cpp/numeric/random/uniform_real_distribution