CAIM Lab, session 9: Recommender systems

17 de diciembre del 2024

Alex Meca Moñino, Alejandro Ruiz Patón

Introducción

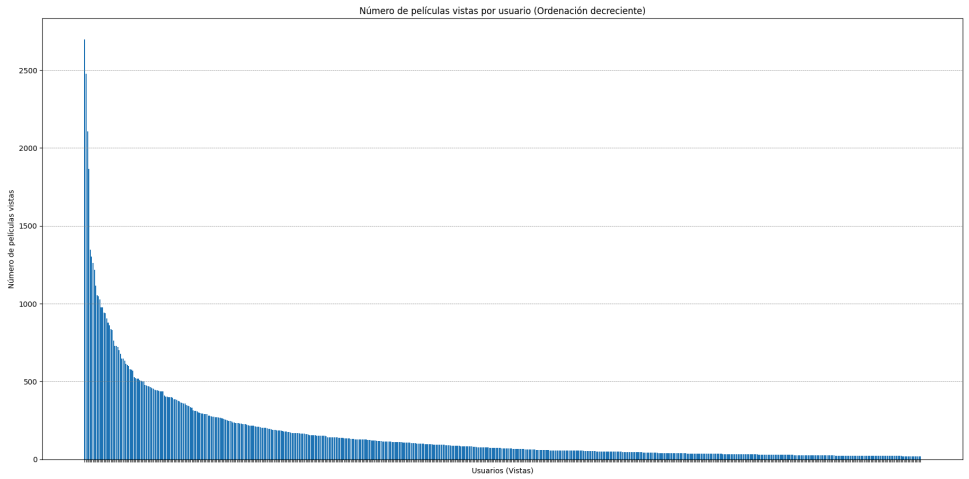
El objetivo de esta práctica es procesar y analizar un dataset que contiene información sobre películas y las puntuaciones otorgadas por distintos usuarios. A partir de los datos disponibles, se buscará extraer información relevante y desarrollar dos sistemas de recomendación distintos, evaluando y comparando su desempeño.

Análisis estadístico de los datos

El primer paso en este análisis es explorar los datos proporcionados para extraer información estadística relevante que no está explícitamente disponible, pero que resulta clave para comprender mejor el comportamiento de los sistemas de recomendación. Para ello, hemos analizado dos archivos principales del dataset: “ratings.csv” y “movies.csv”. A continuación, se presentan los hallazgos más significativos:

- **Cantidad de usuarios únicos**
  El dataset contiene un total de **610 usuarios únicos**, identificados con una ID que varía desde 1 hasta 610.

- **Distribución del número de calificaciones por usuario**
  Consideramos interesante analizar si todos los usuarios califican un número similar de películas o si existe una proporción significativa de usuarios más activos. A continuación, se presenta un análisis de esta distribución.



En la gráfica se observa que existe una variabilidad considerable en el número de calificaciones proporcionadas por cada usuario. Esto indica que una minoría de usuarios es responsable de la mayoría de las interacciones, mientras que la mayoría de los usuarios califica de manera más limitada.
Curiosamente, el gráfico resultante tiene una forma exponencial, donde pocos usuarios destacan por su alta actividad en calificaciones.

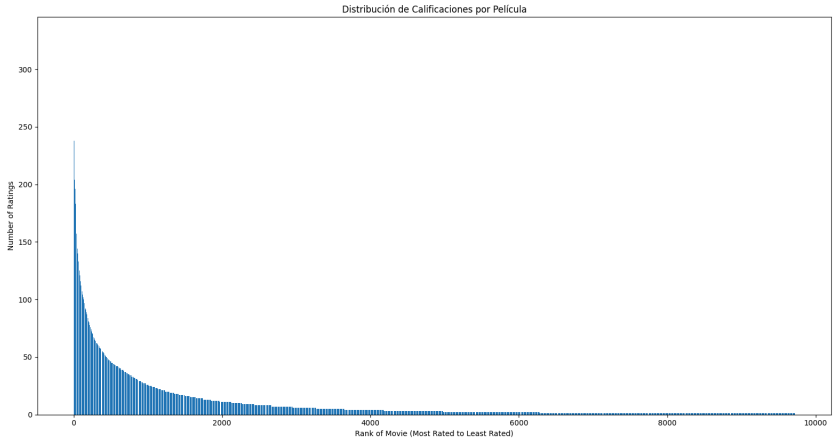
- **Porcentaje de las calificaciones que dan los usuarios**
  Otro análisis interesante que hemos extraído de los datos es la proporción de puntuaciones que los usuarios asignan a las películas. Este análisis permite conocer si los usuarios del dataset tienden a calificar las películas con valores bajos, intermedios o altos.

rating	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
proportion (%)	1.35	2.79	1.78	7.49	5.50	19.88	13.03	26.60	8.48	13.10

  Estos datos reflejan una tendencia hacia calificaciones más altas, especialmente en el rango de **3.0 a 4.0**, lo que podría sugerir que este conjunto de usuarios tiende a valorar generosamente las películas que ven. Además, al analizar la selección de películas en movies.csv observamos que contiene títulos que son bien valorados, lo que también podría influir en este patrón.

- **Cantidad de películas únicas**
  El dataset incluye un total de 9742 películas, lo que proporciona una gran variedad de opciones para los sistemas de recomendación.

- **Distribución de visualizaciones**
  En el gráfico siguiente se muestra cómo se distribuyen las visualizaciones o calificaciones que tiene cada película.



  No todas las películas han recibido la misma cantidad de calificaciones. Mientras que algunas destacan con un número elevado de valoraciones, la mayoría tienen pocas interacciones. Por curiosidad, hemos consultado cuáles son las 5 películas más calificadas dentro del dataset. A continuación se encuentran en una tabla ordenadas de más calificada a menos.

Top	movieId	title
1	356	Forrest Gump (1994)
2	318	Shawshank Redemption, The (1994)
3	296	Pulp Fiction (1994)
4	593	Silence of the Lambs, The (1991)
5	2571	Matrix, The (1999)

Observaremos después si alguna de las películas más calificadas se encuentra dentro del conjunto de películas mejor valoradas.

- **Análisis de los géneros mejor valorados**

Top	Género	Calificación media
1	Film-Noir	3.920115
2	War	3.808294
3	Documentary	3.797785
4	Crime	3.658294
5	Drama	3.656184

Estos resultados sugieren que los géneros más de nicho como Film-Noir y Documentary tienden a recibir puntuaciones más altas, probablemente debido a que atraen a un público más especializado.

Construcción de un sistema de recomendación por calificación

El primer sistema de recomendación que hemos desarrollado en esta sesión es el sistema naive. Este sistema hace recomendaciones basándose únicamente en las películas con las mejores calificaciones. Las películas con los puntajes más altos serán las primeras en ser recomendadas a los usuarios. A continuación, se presentan en una tabla las 10 películas mejor valoradas:

Top	Title	Calificación
1	Heidi Fleiss: Hollywood Madam (1995)	5.0
2	Che: Part Two (2008)	5.0
3	Vampire in Venice (Nosferatu a Venezia) (Nosferatu in Venice) (1986)	5.0
4	Idiots and Angels (2008)	5.0
5	Louis Theroux: Law & Disorder (2008)	5.0
6	True Stories (1986)	5.0
7	Red Sorghum (Hong gao liang) (1987)	5.0
8	Man and a Woman, A (Un homme et une femme) (1966)	5.0
9	Wings, Legs and Tails (1986)	5.0
10	My Life as McDull (Mak dau goo si) (2001)	5.0

El principal problema de este sistema es que al no considerar el número total de calificaciones recibidas por cada película, es común que las recomendaciones estén dominadas por películas que tienen solo unas pocas críticas extremadamente positivas. En el caso del Top 1 (Heidi Fleiss), por ejemplo, observamos que solo hay 2 calificaciones en total y estas tienen una puntuación de 5.0. Esto

demuestra cómo películas con pocas evaluaciones pueden distorsionar el ranking, dándole un peso excesivo a las calificaciones más altas, aunque sean pocas. Este sistema se podría mejorar un poco poniendo un filtro de número de calificaciones. Es decir, solo tener en cuenta aquellas películas que tengan como mínimo k calificaciones.

Además, este enfoque hace que se pierda la posibilidad de recomendar películas menos conocidas que podrían ser interesantes para ciertos usuarios. Al filtrar solo las películas mejor valoradas, se dejan de lado aquellas que podrían ser únicas o con gran potencial, pero que aún no han sido vistas por una gran cantidad de personas.

Finalmente, como este sistema se basa únicamente en las películas más populares o valoradas por la mayoría, se asume erróneamente que lo que gusta a la mayoría será igualmente relevante para cada usuario, lo que puede llevar a recomendaciones menos personalizadas y diversas.

Tiempo de complejidad del sistema
Este sistema se compone de varios pasos:

- coste de groupby = O(n) donde n es el número de ratings diferentes, ya que por cada rating se asigna a un grupo y por cada grupo se calcula la media de sus ratings.
- coste de merge = O(n*m) donde m es el número de películas únicas, ya que se combinan ambas tablas.
- coste de sort = O(nlogn).
- coste de filtrado = O(k) donde k es el número de películas que aparecen en el top.

En conclusión, el coste final del sistema naive varía dependiendo de la relación entre el número de películas únicas m y el número total de calificaciones n. Si m es mucho mayor que n, el coste será aproximado a O(n*m), mientras que si m es considerablemente menor, el coste se aproxima a O(nlogn).

Construcción de un sistema de recomendación user-to-user

En este apartado implementamos el sistema de recomendación basado en la similitud entre usuarios. Este enfoque permite predecir las preferencias de un usuario al recomendarle películas basándose en los patrones de otros usuarios que son similares.

Dentro del proceso de recomendación de películas a un usuario (x) se pide seleccionar un conjunto de usuarios diferentes a x (U) para basarnos en las calificaciones de estos. Para escoger los usuarios que pertenecerán a “U” hemos implementado el cálculo de similitudes con Pearson entre “x” y el resto de usuarios. Finalmente, seleccionamos los “n” usuarios más similares a “x” para añadirlos a “U”. El parámetro “n” es modificable en el código ya que se encuentra como una variable global.

¿Por qué utilizamos Pearson?
El cálculo de similitudes utilizando Pearson es útil porque mide las desviaciones entre las calificaciones esperadas y reales para calcular la similitud entre dos usuarios. Esto permite representar si dos usuarios son completamente opuestos o muy similares. Sin embargo, este modelo no está diseñado para comparar vectores donde todos los valores son iguales, ya que en tales casos siempre devuelve un valor de semejanza de 0.0. Por ejemplo, para dos vectores como [4, 4, 4] o [1, 1, 1], el resultado sería idéntico debido a la ausencia de desviación.

¿Por qué calculamos de esta forma el conjunto U?

- Por precisión: Usar los usuarios más similares permite que las predicciones sean basadas en comportamientos cercanos al usuario objetivo.
- Por eficiencia: Al poder escoger el número “n” de usuarios dentro del conjunto se reduce el costo computacional.

Además de este método, tuvimos otras ideas para la selección de U:

- Selección de otros usuarios similares que han valorado películas que aún no ha visto el usuario actual.
- Selección aleatoria de usuarios.
- Selección de usuarios que han calificado muchas películas

Comparación de los sistemas recomendadores

En esta sección, comparamos ambos sistemas de recomendación basándonos en la precisión de la respuesta y la complejidad del cálculo. Para ello, utilizamos “k”, el número de películas que se recomiendan a un usuario. La evaluación de cada sistema se realiza utilizando las películas del conjunto de validación, las cuales se han reservado previamente y no fueron consideradas durante el proceso de entrenamiento del sistema.

Suponemos que una recomendación es más precisa si la frecuencia de temas en las películas recomendadas refleja una proporción similar a la del conjunto de validación. En términos técnicos, hemos calculado el vector de proporciones de cada tema para cada sistema de recomendación y lo hemos comparado utilizando la similitud coseno con el vector del conjunto de validación.

A continuación, se presentan los resultados del análisis:

Comparación de los sistemas recomendadores (Rec1 vs Rec2)

En función de k

- Con k = 5:

userId	Rec1 Similarity	Rec2 Similatity
123	0.696	0.352
467	0.808	0.555
607	0.850	0.283

- Con k = 10:

userId	Rec1 Similarity	Rec2 Similatity
123	0.639	0.423
467	0.765	0.759
607	0.795	0.582

- Con k = 20:

userId	Rec1 Similarity	Rec2 Similatity
123	0.647	0.452

467	0.717	0.644
607	0.828	0.721

Podemos observar cómo, al aumentar el valor de k, las recomendaciones del sistema user-to-user mantienen un alto nivel de similitud. Esto indica que el sistema ajusta las recomendaciones de acuerdo con las preferencias individuales del usuario, proporcionando una mayor personalización.

Por otro lado, la similitud del sistema naive comienza con valores bajos, ya que estos usuarios pueden tener preferencias distintas a las que predominan en el conjunto de películas más populares. A medida que se aumenta k, el rango de películas consideradas se amplía, permitiendo que el sistema naive incremente su similitud al incluir una diversidad más amplia de temáticas y géneros, lo que resulta en una mejora en su precisión.

De esta forma, observamos que el sistema user-to-user ofrece recomendaciones mucho más precisas y personalizadas.

Variación del tiempo de ejecución

userId	User-To-User Time (s)	Naive Time (s)
123	6.0378	0.007
467	8.219	0.007
607	3.985	0.007

El sistema naive al ser más sencillo e independiente de los gustos del usuario objetivo mantiene un tiempo de ejecución constante y muy inferior al user-to-user. Este depende directamente del número de películas que ha calificado, de la selección del conjunto U y de otros factores que incrementan el tiempo de cómputo en función del usuario.

Análisis con un número mayor de usuarios

userId	Rec1 Similarity	Rec2 Similarity
30	0.346	0.131
59	0.676	0.439
123	0.696	0.352
467	0.808	0.555
502	0.400	0.431
607	0.850	0.283

En esta tabla analizamos el comportamiento de ambos sistemas utilizando un mayor número de usuarios. Se observa que, en general, el sistema user-to-user mantiene una similitud superior al sistema naive.

Sin embargo, encontramos casos curiosos. Por ejemplo, el usuario 30, aunque presenta una mejor similitud en el sistema user-to-user, muestra un valor muy bajo. Esto puede deberse a que el número de calificaciones de este usuario es reducido, lo que limita la información utilizada para entrenar el sistema, afectando la precisión en las recomendaciones.

Otro caso interesante es el del usuario 502, cuya representación es bastante similar en ambos sistemas. Esto podría ser explicado por el hecho de que este usuario tiende a disfrutar de películas con altas calificaciones, lo que hace que tanto user-to-user como naive generen recomendaciones similares.

Conclusiones y dificultades

En general, ambos sistemas de recomendación ofrecen distintos enfoques útiles, cada uno adecuado para situaciones específicas. User-to-user proporciona una experiencia más personalizada con una precisión elevada, aunque a costa de un mayor tiempo de cómputo. Por otro lado, naive logra resultados más rápidos, mejorando considerablemente el tiempo de ejecución, aunque su precisión es moderada y no siempre tan alta, sin llegar a ofrecer valores de similitud extremadamente bajos.

En cuanto a las dificultades enfrentadas durante la práctica, no se han presentado problemas significativos. A pesar de que el análisis realizado puede haber resultado un tanto extenso, buscamos cubrir cada aspecto de forma detallada para garantizar un entendimiento claro de los resultados obtenidos.