Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
IIC3633 - Sistemas Recomendadores

Tarea 1

12 de septiembre de 2022

Sebastián Contreras - Alonso Venegas

Índice

1. Análisis exploratorio

1.1. Distribución de interacciones por usuario

Se obtuvo que los 10 usuarios con más actividad en el dataset de entrenamiento realizaron el 0,6 % del total de las compras presentes en este, reportando más de 600 compras cada uno.

Usuario	Nº de compras	Porcentaje que representa en el dataset (%)
Top 1	1346	0.102086
Top 2	950	0.072052
Top 3	910	0.069018
Top 4	875	0.066363
Top 5	789	0.059841
Top 6	740	0.056124
Top 7	724	0.054911
Top 8	670	0.050815
Top 9	664	0.050360
Top 10	651	0.049374

Tabla 1: Número de compras realizadas por los 10 usuarios más activos del dataset de entrenamiento.

Los IDs de los 10 usuarios con más actividad en el dataset de entrenamiento fueron los siguientes:

1. a65f77281a528bf5c1e9f270141d601d116e1df33bf9df512f495ee06647a9cc
2. 84c34f4f564db1f437943c77af41f83bf6fd7c01701cbb050070369176905712
3. e55d5ddb3a0c3fb1b4df8edbf526ba12989ab2a852c72774e3f3338cbbb3335e
4. 2df54f0d0653811fe06479c93905f3e6ecc6d07edf39d8b56e5b66c86182bedf
5. 9f12a01e2982f70a820b5dd61528bf769b94c5c5e43b23704f1f654784bcda58
6. bbebb44478948f5052c3f4c5dc04f08653e7938886a85685917fd22b92f22cd0
7. d3b5f70ec21ad1718cf4951445e97007de0d4e85c39ea9fd2fedaf1966280943
8. ad3090d52d11671ffb43bfaa85e3620eff669e8c92c9114ba7755876254cbba8
9. 0152d53f51444891ea07013fd1fb8325415bb09bb6798a59359b21a8326d801b
10. 01a4717d38b651e46dda7f1ab8d1494af2682a847fa9a52a8f9ab1a09acd0294

Sin embargo, el resto de los usuarios realizó una cantidad mucho menor de compras, concentrándose la mayor parte de ellos en el rango de 2 a 6 compras realizadas. Pasadas las 6 compras, la cantidad de usuarios disminuye hasta una meseta que se mantiene entre las 7 y 20 compras. Posterior a dicha cantidad de compras, se observa una disminución drástica en el número de usuarios, aunque es de notar que, de forma agregada, los usuarios con 100 o más compras constituyen un grupo considerable dentro del dataset.

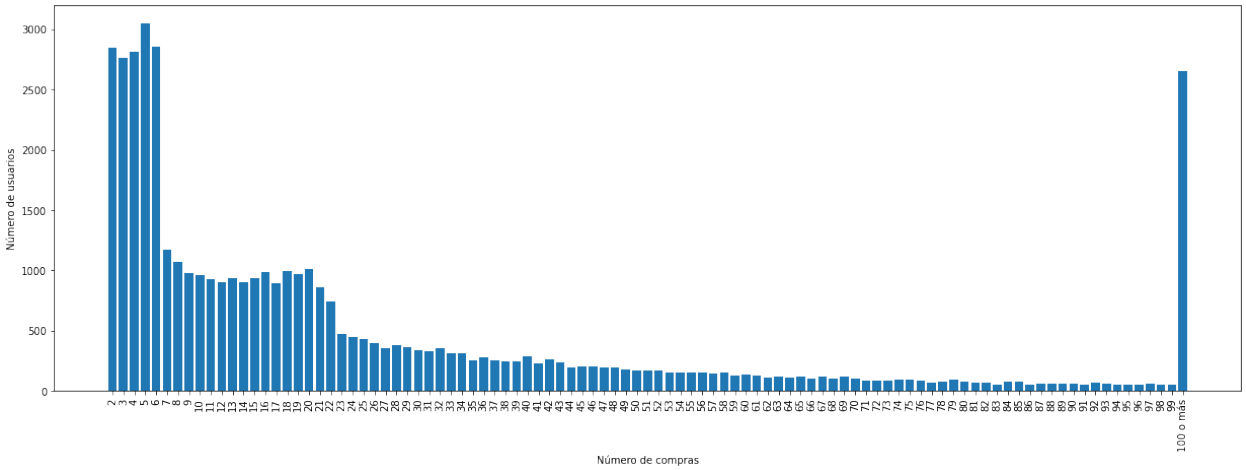

Figura 1: Cantidad de usuarios que ha realizado un determinado número de compras en el dataset de entrenamiento.

1.2. Distribución de interacciones por producto

Los 10 productos más comprados representan el 0,8 % del total de interacciones en el dataset.

Producto	Nº de compras	Porcentaje que representa en el dataset (%)
Jade HW Skinny Denim TRS	1978	0.150019
Jade HW Skinny Denim TRS	1357	0.102920
Tilly (1)	1302	0.098749
7p Basic Shaftless	1225	0.092909
Tilda tank	1010	0.076602
Curvy Jeggings HW Ankle	991	0.075161
Greta Thong Mynta Low 3p	920	0.069776
Luna skinny RW	899	0.068183
Luna skinny RW	849	0.064391
7p Basic Shaftless	849	0.064391

Tabla 2: Número de compras realizadas a los 10 productos más comprados del dataset de entrenamiento.

Se observa que la mayor parte de los productos fueron comprados 1 sola vez, mientras que el número de productos que fueron comprados una mayor cantidad de veces disminuye a medida que aumentamos el número de compras, asemejándose la curva a un decrecimiento exponencial.

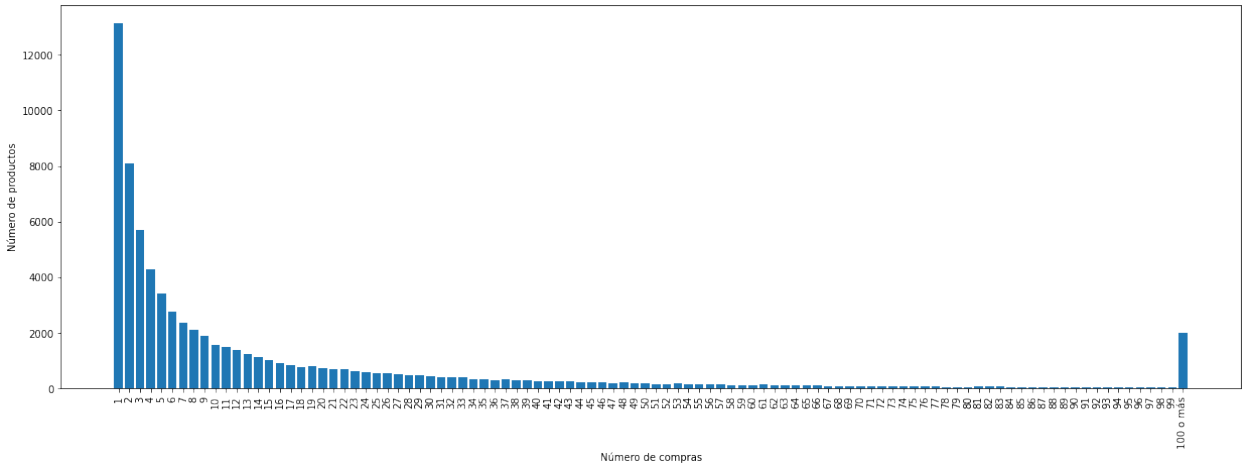

Figura 2: Cantidad de productos que fueron comprados una cierta cantidad de veces en el dataset de entrenamiento.

1.3. Tabla resumen

Variable	Valor
Número de usuarios distintos	45000
Número de ítems distintos	73080
Promedio de productos por usuario	29.300022
Desviación estándar de productos por usuario	44.939461
Promedio de usuarios por producto	18.041886
Desviación estándar de usuarios por producto	38.01747
Densidad del conjunto de datos en cuanto a compras	0.000342

Tabla 3: Vista general de la distribución de interacciones en el dataset de entrenamiento.

2. Recomendación no personalizada

2.1. *Most Popular*

Se seleccionaron los 30 productos más populares del set de entrenamiento y se les recomendó dichos productos a todos los usuarios. Luego, se contrastaron los ítems recomendados con los ítems que compró cada usuario en el set de validación para determinar la relevancia o no relevancia de cada ítem recomendado, obteniéndose las siguientes métricas:

@	NDCG	MAP
10	0.0030	0.0092
20	0.0026	0.0102
30	0.0023	0.0106

Tabla 4: Métricas obtenidas para la recomendación *Most Popular*.

2.2. *Random*

Se seleccionó una muestra de 30 productos del set de entrenamiento, al azar y sin reposición, y se les recomendó dichos productos a todos los usuarios. Las métricas obtenidas para dicha recomendación fueron las siguientes:

@	NDCG	MAP
10	$2.1572 \cdot 10^{-5}$	$4.1556 \cdot 10^{-5}$
20	$2.3379 \cdot 10^{-5}$	$6.0385 \cdot 10^{-5}$
30	$2.1758 \cdot 10^{-5}$	$6.6804 \cdot 10^{-5}$

Tabla 5: Métricas obtenidas para la recomendación *Most Popular*.

3. Recomendación basada en feedback implícito

Para esta sección se entrenaron 10 modelos diferentes: 5 modelos de factorización de matriz optimizada con ALS y 5 modelos de factorización de matriz optimizada con BPR. En ambos casos, los 5 modelos difieren en el número de factores latentes indicado para cada uno: 50, 100, 200, 500 y 1000.

Con respecto a los tiempos de entrenamiento de cada modelo, notamos que aquellos optimizados con BPR tuvieron un tiempo de entrenamiento menor que sus contrapartes optimizadas con ALS. En el caso de BPR, la variación del tiempo de entrenamiento en función de los factores latentes fue relativamente lineal, mientras que en ALS dicha variación sufrió un aumento brusco entre los 500 y los 1000 factores latentes.

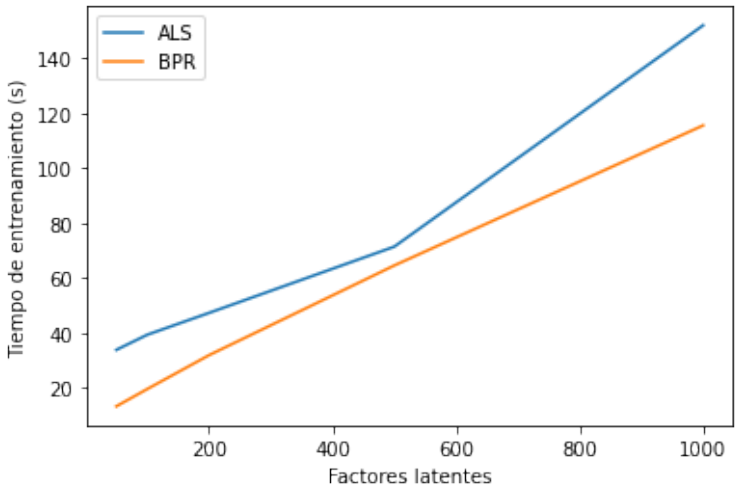

Figura 3: Tiempo de entrenamiento en función de la cantidad de factores latentes seleccionada para cada modelo.

Con respecto a las métricas de evaluación, se observa un mejor desempeño por parte de ALS para toda cantidad de factores latentes en comparación con los modelos entrenados con BPR. Para ALS se observa un aumento de NDCG@10 a medida que se aumenta la cantidad de factores latentes, mientras que BPR alcanzó su máximo NDCG@10 con 200 factores latentes, disminuyendo su desempeño desde ese punto en adelante. Para las métricas MAP@10 se presentó el mismo comportamiento que las métricas NDCG@10 para ambos tipos de modelos, solo que en ambos casos se obtuvieron valores mayores que sus contrapartes en NDCG.

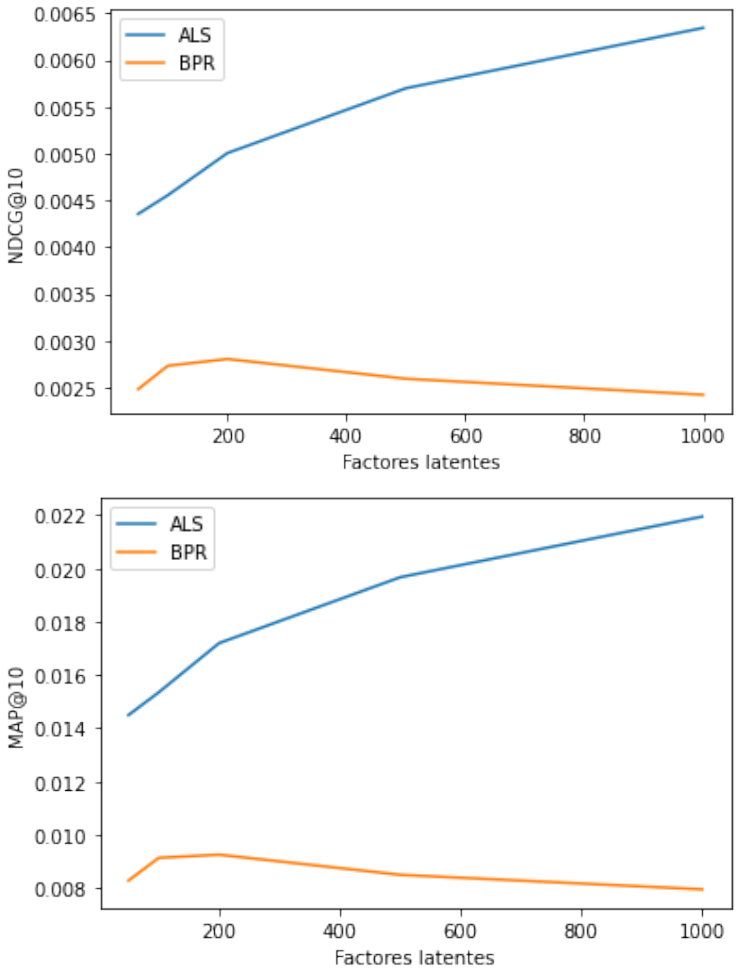

Figura 4: Métricas NDCG@10 y MAP@10 en función de la cantidad de factores latentes seleccionada para cada modelo.

4. Recomendación basada en contenido

1. Se realiza el embeding de la descripción de cada articulo de la base de datos de entrenamiento. Esto se realiza con universal sentence encoding, que vectoriza el lenguaje natural en vectores de 512 componentes.

2. Luego se realiza la reducción de componentes usando PCA, lo que reduce la dimensionalidad de cada vector de 512 a 20 componentes. Esto implica que se pierde información, pero la ganancia en terminos de tiempo y esfuerzo computacional vale la pena.

3. Luego, necesitamos calcular una representación vectorial del usuario, para calcular distancia con la representación vectorial de los articulos y así poder recomendar. Para esto nos piden hacerlo con el promedio de los articulos con los que el usuario ya iteractuo (osea compró). Para se usan las funciones de los dataframes, primero se agregan los embeding como una nueva columna del df de los articulos, luego se hace un join con el train_set, que contiene las transacciones, y luego se agrupa por id de usuario, calculando el promedio de los embedings, esto nos deja una base de dato donde para cada id de usuario está su representación vectorial.

4. Se calcula las recomenadciones para cada usuario en el set de validación calculando la distancia euclideana entre su representación vectorial y la respresentación vectorial de todos los items. Esto último lo intentamos hacer pero el código estuvo corriendo por mas de una hora y no terminó. Finalemnte la comparacion de métricas no pudo ser realizada por el problema mencionado anteriormente

5. Discusión de resultados

Entre las posibles razones que pueden estar incidiendo sobre los resultados obtenidos se encuentra el supuesto de que los datos contenidos en el dataset de validación son los únicos en los que estará interesado el usuario, siendo que podría perfectamente no ser así.

Además, para la mayor parte de los modelos se calcularon métricas "@10", cuando tal vez una métrica "@30" hubiera presentado mejores resultados, ya que aumentan las probabilidades de ver un producto en el que el usuario sí estaba interesado.