Universidad de Sevilla

Máster Ciencia de Datos y Big Data

Sistemas de Recomendación

Carlos Pérez Manzano

Introducción

Los Sistemas de Recomendación son empleados en diferentes industrias. Damos a continuación algunos ejemplos de su uso:

1. **Obtener un mayor número de ventas**: comercios como Amazon tienen algortimos de recomendación con el objetivo de llamar la atención del usuario, aumentando así la probabilidad de compra de productos de este.

2. **Mejorar la experiencia del usuario**: redes sociales como Instagram obtienen la mayoría de sus beneficios de pubilicidad. Es por ello, que se asocia a los usuarios cierta publicidad que no perjudique la experiencia del usuario.

3. **Aumentar la interacción entre usuarios**: recomendar usuarios que pueden resultar conocidos es una aplicación de la mayoría de redes sociales. También en plataformas de búsqueda de empleo se recomiendan empresas a usuarios en función de las ofertas de empleo en las que este se ha postulado.

En general el objetivo de un Sistema de Recomendación va a ser el mismo, dar a un usuario una lista de recomendaciones ordenada de mayor a menor preferencia. Existen distintas técnicas para abordar este fin. El filtrado colaborativo se basa en los usuarios más similares al usuario a recomendar. El filtrado basado en contenido, que se fundamenta en las características de los productos a recomendar. También existen modelos que combinan ambas visiones.

En nuestro caso, vamos a realizarlo con un enfoque basado en filtrado colaborativo e Inteligencia Colectiva. Es decir, nos basaremos en las similitudes entre usuarios, pero tendremos en cuenta las valoraciones de todos los usuarios, tanto aquellos con unos gustos muy parecidos como aquellos que no coincidan prácticamente en sus preferencias.

Para la ejemplificación de este modelo se han tomado los datasets de la plataforma **MovieLens**, que se encuentran en el siguiente enlace:

https://grouplens.org/datasets/movielens/latest/

Contiene datos de valoraciones de usuarios a películas desde el 29 de marzo de 1996 al 24 de septiembre de 2018. Se seleccionaron aleatoriamente valoraciones de usuarios que hubiesen realizado al menos 20 opiniones. En [1] se puede encontrar el código en Python.

Creación del modelo y posibles mejoras

Debido al enfoque que vamos a dar al Sistema de Recomendación, el aspecto clave que va a determinar en mayor medida la precisión del modelo es la elección de la medida de similaridad entre usuarios. Existen muchas medidas de similaridad: coeficiente de correlación de Pearson, similaridad euclidea, Manhattan, coeficiente de Jaccard, etc. Para el modelo vamos a escoger la similaridad del coseno.

Definición. Dados dos vectores $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$, se define la similardiad del coseno como:

$$s(a, b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

La similaridad del coseno mide el ángulo entre dos vectores multidimensionales, siendo el valor 1 si tienen la misma dirección y sentido, 0 si son ortogonales y -1 si tienen la misma dirección y sentido contrario. La principal ventaja que tiene el uso de esta medida es la independencia de la magnitud de los vectores. Esto es realmente interesante para nuestro caso pues dos usuarios pueden tener gustos muy similares y sin embargo tener una dureza distinta a la hora de valorar las películas.

Consideramos U el conjunto de usuarios y P el conjunto de películas potencialmente recomendables. Si se quiere recomendar películas a un usuario $u_j \in U$ el modelo se construirá de la siguiente manera:

1. Contruir la matriz X de pares película-usuario donde se recogen las valoraciones. Consideraremos por fijar ideas las películas en filas y los usuarios en columnas. Por tanto la dimensión de X es $|P| \times |U|$.

2. Identificar las películas no vistas por el usuario, es decir, $\{p_i \mid x_{ij} = NaN,\ i \in P\}$.

3. Rellenar los valores nulos de X.

4. Calcular la matriz S de similitudes entre usuarios.

5. Calcular la valoración estimada de esas películas para el usuario u_i. La puntuación $\hat{r}(i,j)$ de una película i para el usuario j se calcula de la siguiente forma:

$$\hat{r}(i,j) = \frac{\sum_{k=1}^{|U|} x_{ik}\, s_{kj}}{\sum_{k=1}^{|U|} |s_{kj}|}$$

   Es decir, ponderamos las valoraciones del resto de usuarios en función de la similitud.

6. Seleccionamos las películas con una mayor puntuación.

7. Mostramos las recomendaciones.

Cabe destacar que el algoritmo que se ha indicado se podría realizar con cualquier similaridad distinta a la del coseno, pues solo influye en la creación de la matriz de similaridades. También es importante notar que si consideramos los usuarios fijos no es necesario calcular esta matriz de similaridades siempre que se quiera dar una recomendación, pues es invariante y puede ser almacenada previamente.

Prodecemos ahora a discutir el punto 2) del algortimo, ya que son prácticamente incontables los métodos disponibles para la imputación de los valores nulos. Destaquemos algunas de las opciones:

- Rellenar un valor nulo x_{ij} por la media de las valoraciones de la película para el resto de usuarios, es decir $\frac{\sum_{k=1}^{|U|} x_{ik}}{|U|}$. Esta técnica es aceptable en caso de que las valoraciones de las películas tengan poca variabilidad respecto de su media. Sin embargo, no tiene en cuenta la posible variabilidad en la dureza de las opiniones de los usuarios. Además este método va a sobreestimar las puntuaciones estimadas, pues la valoración para una película no vista por dos usuarios va a ser la misma.

- Imputar el valor nulo x_{ij} por la media de las valoraciones de las películas puntuadas por el usuario u_j $\frac{\sum_{k=1}^{|P|} x_{kj}}{|P|}$. Esta opción trata mejor la diferencia de la dureza al dar opiniciones de los usuarios, y soluciona el problema de la sobreestimación de las puntuaciones estimadas. Sin embargo, se comportará peor con usuarios con gran variabilidad en las valoraciones.

- Una opción similar a la anterior pero algo más fina es sustituir por la media de las valoraciones del usuario para las películas que tengan unas características similares. Por ejemplo, en el dataset de ejemplo tenemos la información del género, por tanto podríamos sustituir por la media de las películas valoradas por el usuario del mismo género en caso de que existan y por la media global de las películas valoradas en caso contrario.

- También podríamos hacer uso de imputadores más sofisticados como KNN, tomando como valoración de la película la media de los usuarios más cercanos que han valorado esa película.

Bibliografía

[1] https://github.com/carlosperman/Sistema-de-Recomendacion

[2] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. https://doi.org/10.1145/2827872