Universidad del Valle De Guatemala

Facultad de Ingeniería

Deep Learning

Javier Fong

Laboratorio 6: Sistemas de recomendaciones

Javier Mombiela 20067
Roberto Ríos 20979

Guatemala, 29 de septiembre 2023

Análisis Exploratorio de Datos (EDA) del Dataset

Antes de entrar en los detalles de los dos modelos de sistemas de recomendación, es importante realizar un análisis exploratorio de datos para comprender mejor la información con la que se están trabajando.

Dataset para el Modelo Basado en Contenido
El conjunto de datos utilizado para el modelo basado en contenido parece ser un archivo CSV llamado 'joined.csv'. Este conjunto de datos contiene información sobre libros y usuarios. Se observan las siguientes características:

- Se han realizado transformaciones en las columnas 'Publisher', 'Book-Title' y 'Book-Author' mediante la codificación de etiquetas.
- Las características utilizadas en el modelo incluyen 'User-ID', 'Book-Title', 'Book-Author', 'Year-Of-Publication' y 'Publisher'.
- Los datos se han dividido en conjuntos de entrenamiento y prueba con una división del 80% para entrenamiento y 20% para prueba.

Dataset para el Modelo Basado en Filtros Colaborativos
El conjunto de datos utilizado para el modelo basado en filtros colaborativos se compone de tres archivos CSV: 'Books.csv', 'Users.csv' y 'Ratings.csv'. Se observan las siguientes características:

- Los conjuntos de datos contienen información sobre libros, usuarios y calificaciones de libros.
- Se han realizado transformaciones en las columnas 'User-ID' e 'ISBN' mediante la codificación de etiquetas.
- Los datos se dividen en conjuntos de entrenamiento y prueba con una división del 90% para entrenamiento y 10% para prueba.

Estructura de las Redes y Funcionamiento

Modelo Basado en Contenido
El modelo basado en contenido utiliza una red neuronal con las siguientes capas:

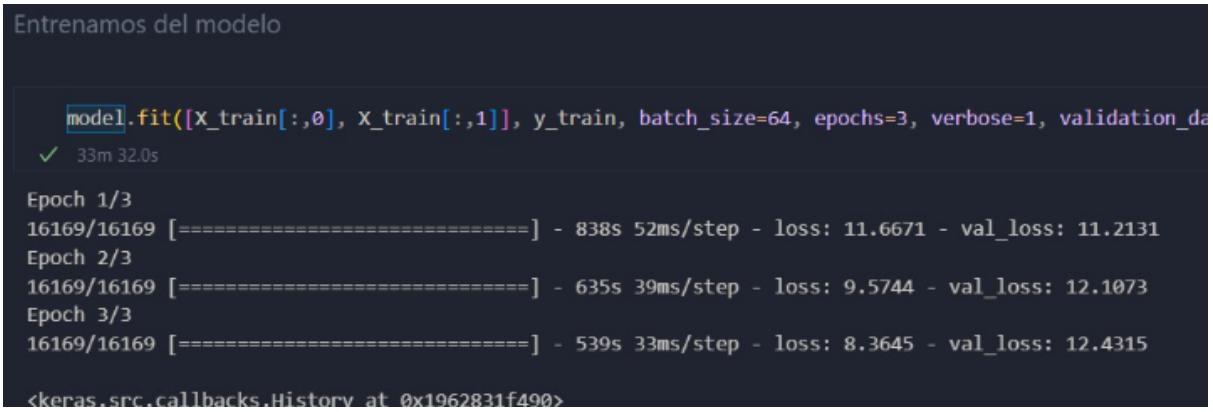
- Capa de entrada con las características relevantes del usuario y el libro.
- Capa oculta con 64 neuronas y función de activación ReLU.
- Capa oculta con 32 neuronas y función de activación ReLU.
- Capa de salida con una sola neurona y función de activación lineal.
- El modelo se compila utilizando la función de pérdida de error cuadrático medio (MSE) y el optimizador Adam. Posteriormente, se entrena el modelo durante 5 épocas.

Modelo Basado en Filtros Colaborativos
El modelo basado en filtros colaborativos utiliza una arquitectura de redes neuronales más compleja que involucra la incorporación de usuarios y libros. Las capas del modelo incluyen:

- Capa de entrada para el usuario.
- Capa de entrada para el libro.
- Capa de incorporación (embedding) para usuarios y libros.
- Capas densas para procesar las incorporaciones y hacer predicciones.
- El modelo se compila utilizando la función de pérdida de error cuadrático medio (MSE) y el optimizador Adam. Luego, se entrena durante 3 épocas.

Comparación de Modelos y Resultados

Modelo Basado en Contenido
Los resultados del modelo basado en contenido se presentan con las siguientes métricas:




```
32198/32198 [==============================] - 53s 2ms/step
        Book-Title  Book-Author  Predicted-Rating
766365        8558           54         14.516417
865366        7648          545         11.881720
865365        7648          545         11.881720
865364        7648          545         11.881720
899768        4842          212          9.658482
899769        4842          212          9.658482
770337        1956          592          9.635492
770336        1956          592          9.635492
437299        3813          682          9.164554
437301        3813          682          9.164554
```

Error Cuadrático Medio (MSE): 14.8576
Raíz del Error Cuadrático Medio (RMSE): 3.8546

Coeficiente de Determinación (R^2): -1.9383e-05

Las recomendaciones generadas por el modelo incluyen predicciones de calificación para varios libros. Como se puede observar en la segunda imagen, algunas de las predicciones sobrepasan el rating de 10, lo cual puede ser un problema, ya que las calificaciones generalmente están en el rango de 1 a 10.

Modelo Basado en Filtros Colaborativos

Los resultados obtenidos con este modelo son los siguientes:

```
Entrenamos del modelo

model.fit([X_train[:,0], X_train[:,1]], y_train, batch_size=64, epochs=3, verbose=1, validation_data=([X_test[:,0], X_test[:,1]], y_test))
✓ 33m 32.0s

Epoch 1/3
16169/16169 [==============================] - 838s 52ms/step - loss: 11.6671 - val_loss: 11.2131
Epoch 2/3
16169/16169 [==============================] - 635s 39ms/step - loss: 9.5744 - val_loss: 12.1073
Epoch 3/3
16169/16169 [==============================] - 539s 33ms/step - loss: 8.3645 - val_loss: 12.4315

<keras.src.callbacks.History at 0x1962831f490>
```

```
8480/8480 [==============================] - 9s 1ms/step
                                                Book-Title  \
78867                          The Shrinking of Treehorn
184411  Michelin THE GREEN GUIDE Quebec, 4e (THE GREEN...
79431   The Blue Day Book: A Lesson in Cheering Yourse...
31331                              A Kiss for Little Bear
38292                                       The Lorax
3028                                             Free
16190                                       Falling Up
238677                         Fiction Writer's Handbook
66613                                M.Y.T.H. Inc. Link
53754           A Baby...Maybe -- How To Hunt a Husband

                      Book-Author  Predicted-Rating
78867          Florence Parry Heide         8.933759
184411  Michelin Travel Publications         8.719502
79431         Bradley Trevor Greive         8.674919
31331        Else Holmelund Minarik         8.596144
38292                      Dr. Seuss         8.582002
3028                    Paul Vincent         8.418205
16190                Shel Silverstein         8.393199
238677                 Hallie Burnett         8.323505
66613                   Robert Asprin         8.290651
53754                   Bonnie Tucker         8.270459
```

Como se puede observar en los resultados, este modelo tuvo un mejor, desempeñó que el primero, podemos ver por las épocas que este modelo tuvo un poco de overfitting, por lo cual este modelo aún pudo haber mejorado más. Aun así, vemos que tiene buenas recomendaciones y que los ratings son mejores que los del primer modelo.

¿Qué Modelo Funciona Mejor y por Qué?

El modelo basado en filtros colaborativos se destaca como la mejor opción en esta comparación. Esto se debe a su capacidad para aprender a partir de las interacciones pasadas de los usuarios con los libros. Al considerar cómo los usuarios han calificado y se han relacionado con los libros previamente, este modelo puede identificar patrones de preferencia de una manera efectiva. Además, su uso de representaciones vectoriales para usuarios y libros permite personalizar las recomendaciones, teniendo en cuenta las afinidades específicas de un usuario por categorías de libros o autores.

En contraste, el sistema de recomendaciones basado en contenido presenta desventajas notables. En este caso, se observa que los resultados tienen un alto Error Cuadrático Medio (MSE) y un bajo Coeficiente de Determinación (R^2), lo que sugiere un rendimiento deficiente. Esto podría deberse a que se basa principalmente en las características de los libros y los usuarios, lo cual puede ser problemático si la información detallada sobre el contenido no está disponible o no es lo suficientemente descriptiva. Además, las predicciones que sobrepasan el rating de 10 plantean preocupaciones sobre la precisión del modelo.

En conclusión, el modelo basado en filtros colaborativos sobresale por su capacidad para aprender de las interacciones pasadas y personalizar recomendaciones. En contraste, el sistema de recomendaciones basado en contenido muestra desventajas notables, incluyendo un bajo rendimiento en este contexto y la limitación de depender en gran medida de las características del contenido.