Maestría en Inteligencia artificial aplicada

TC4034.10 Análisis de grandes volúmenes de datos
Dr. Nestor Velasco Bermeo

Actividad 4.3: Avance de proyecto 1 - Sistema de Recomendación

Equipo 31
Giovanni Andrés Acuña Morales - A01794007
Juan Pablo Noguerón Morales – A01097897
Juan Carlos Soni Contreras – A01610128

19 de Mayo, 2024

Plan de proyecto y cronograma

Durante el desarrollo de este proyecto, vamos a adherirnos a la mayoría de los pasos sugeridos por la metodología OSEM. Dicha metodología tiene definidas ciertas fases típicas por las que un proyecto de Data Science atraviesa. Las fases definidas dentro del framework OSEM son:

- **Obtain** - Obtener y accesar los datos
- **Scrub** - Limpieza y preparación de datos
- **Explore** - Explorar la información usando varias técnicas para intentar entender la historia que ésta te puede contar
- **Model** - Revisar las técnicas de modelado
- **iNterpret** - Interpretación de los resultados y lo que éstos revelan a los stakeholders.

Dicho esto, cabe mencionar que durante el desarrollo del proyecto realizaremos varias iteraciones de las fases mencionadas dependiendo de los hallazgos que encontremos en cada milestone del proyecto. Por lo que será común realizar exploraciones, limpieza de datos y hasta obtención de más información durante fases posteriores del proyecto si así lo consideramos necesario.

Los milestos identificados hasta el momento para este proyecto con su respectiva cronología son:

Sistemas de recomendación - Milestones	Fase OSEM	Start Date	End Date
Investigación sobre sistemas de recomendación		4/22/2024	4/28/2024
1. Primer avance		5/13/2024	5/19/2024
1.1 Selección de industria y conjunto de datos	**Obtain**	5/13/2024	5/19/2024
1.2 Análisis inicial y preprocesamiento de dataset	**Scrub**	5/13/2024	5/19/2024
1.3 Exploración inicial de dataset	**Explore**	5/13/2024	5/19/2024
1.4 Algoritmo de recomendación básico	**Model**	5/13/2024	5/19/2024
2. Segundo avance		5/20/2024	5/26/2024
3. Tercer avance		TBD	TBD

Selección de conjunto de datos y preprocesamiento.

Para este ejemplo se decidió usar el conjunto de datos **MovieLens 100K Dataset**, que contiene clasificaciones de 100.000 películas del repositorio MovieLens, con 100.000 valoraciones de 5 estrellas y 3.600 aplicaciones de etiquetas aplicadas a 9.000 películas por 600 usuarios. Fue recopilado en varias ocasiones por GroupLens, un laboratorio de investigación del Departamento de Informática e Ingeniería de la Universidad de Minnesota. Este conjunto de datos fue publicado en Abril de 1998.

Para el procesamiento del dataset se realiza descargando del repositorio el dataset a preprocesar usando la función ***data = Dataset.load_builtin('ml-100k')***, posteriormente crearemos una instancia con la función ***algo = SVD()***, y por último realizamos una validación cruzada de 5 pliegues usando los pasos anteriores, ***cross_validate(algo, data, measures=['RMSE', 'MAE'], cv=5, verbose=True)***.

Exploración inicial y análisis de conjunto de datos

En este punto se realizará abarcará la industria del entretenimiento, específicamente el área de cine. En este ejercicio se usará el conjunto de datos **MovieLens 100K Dataset**, y se usará para estimar o predecir la calificación de una película en específico, en base al historial que calificaciones de otras películas que el usuario hubiese calificado, tomando en cuenta factores como el género, duración, trama, personajes y múltiples factores que permitirá predecir la posible calificación que el usuario dé a la película. Esto nos permite poder estimar su comportamiento en el estreno y tomar acciones como de reducir su posibilidad de sugerencia a otros usuarios que pueda tener una mayor calificación y evitar que por la mala experiencia termine abandonando la plataforma de video.

El algoritmo creado consta de **7** pasos:

1. **Importar las librerías:** Importaremos las librerías ***surprise*** y ***surprise.model_selection*** que son necesarias para el ejercicio.

```
from surprise import Dataset, Reader, SVD
from surprise.model_selection import cross_validate
```

2. ***Cargar el dataset de MovieLens:*** Con la función ***Dataset.load_builtin***, descargaremos el Dataset ***ml-100k***.
   data = Dataset.load_builtin('ml-100k')
   2.2 Checar Información: data.info() y data.describe() y data.shape()

3. ***Crear una instancia:*** Usando la Función **SVD** se creará una instancia.
   algo = SVD()

4. ***Validación cruzada***: Con la función ***cross_validate***, realizaremos una validación cruzada de 5 pliegues usado la instancia y el dataset creados en los pasos anteriores.
   cross_validate(algo, data, measures=['RMSE', 'MAE'], cv=5, verbose=True)

5. ***Entrenamiento:*** Con la función ***build_full_trainset***, podremos entrenar el dataset
   trainset = data.build_full_trainset()
   algo.fit(trainset)

6. ***Selección de usuario y película***: En este paso elegiremos el usuario y la película para su predicción.
   uid = str(196) # ID del usuario
   iid = str(302) # ID de la película

7. ***Predicción:*** Con la función ***predict*** y con el usuario y la pelicuta generará la predicción de la calificación de la película por parte del usuario.
   pred = algo.predict(uid, iid)
   print ('Redomenda ción:', pred)

Algoritmo de recomendación básico con el conjunto de datos elegido

La ubicación del algoritmo se encuentra en el siguiente nombre y ubicación SistemaReco.ipynb.