PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

IIC3633 Sistemas Recomendadores (2024-1)

Tarea 1

Indicaciones

- Fecha de entrega: **Viernes 5 de abril de 2024, 20:00 horas.**
- La tarea debe realizarse **individualmente o en grupos de máximo dos personas**. La copia será sancionada con una nota 1,0 en la tarea, además de las sanciones disciplinarias correspondientes.
- Entrega a través de pestaña grupos en la plataforma CANVAS.
- Cada hora o fracción de atraso descuenta 0,5 puntos de la nota obtenida, llegando a 1,0 en 6 horas. Se considera como entrega el último archivo subido por alguno de los miembros del grupo. No se revisarán tareas que hayan sido subidas con anterioridad a la última.
- Se sugiere hacer la tarea en Google colab o en jupyter notebooks para facilitar la revisión. Deberán entregar estos notebooks ejecutados como parte de su código.
- Los datos entregados contienen más información de la estrictamente necesaria para el desarrollo de las actividades. Está permitido utilizar esta información extra si se desean mejorar los métodos descritos en las actividades. Sin embargo, debe existir una justificación para el uso de los mismos y además, los métodos de las actividades no pueden perder su estructura.

OBJETIVO

En esta tarea tendrán la oportunidad de poner en práctica sus conocimientos sobre Sistemas Recomendadores. En particular, exprimentarán con recomendación no personalizada, basada en feedback implícito y basada en contenido.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

En esta tarea utilizarán un subset del dataset HM Personalized Fashion Recommendations de Kaggle que contiene información del historial de compras de clientes a lo largo del tiempo en la tienda HM. Contiene información del cliente así como del mismo artículo.

El dataset con el que trabajarán consiste en:

- Dataset de train (*transactions_train.csv*): **1,318,501** registros que contienen **45,000** usuarios distintos. Cada fila contiene el id del usuario, el id del artículo, la fecha de la transacción, el precio y el canal de venta. Descargar aquí.
- Dataset de test (*users_test.csv*): **44,917** ID de usuarios distintos. Estos son los usuarios a los que deben proporcionar recomendaciones en la Actividad 5. Descargar aquí.
- Dataset de validación (*transactions_val.csv*): **140,312** registros con **42,513** usuarios distintos. Cada fila contiene el id del usuario, el id del artículo, la fecha de la transacción, el precio y el canal de venta. Descargar aquí.
- Dataset de usuarios (*customers.csv*): **45,000** registros de usuarios diferentes. Cada fila contiene el id del usuario, el estado de su membresía, edad y código postal. Descargar aquí.
- Dataset de artículos (*articles.csv*): **77,650** registros de items distintos. Cada fila tiene el id del artículo, el código, nombre, número del tipo de artículo, nombre del tipo de artículo, nombre del grupo del artículo, el número y el nombre de la apariencia gráfica, el código y nombre del grupo de color, número y nombre del departamento, número y nombre de la sección, número y nombre de la prenda y finalmente una descripción del artículo. Descargar aquí.
- Dataset de imágenes (*images_rescaled.zip*): **77,398** imagenes de distintos items. Donde el nombre de las imágenes corresponde al id del artículo que representan. Descargar aquí.

LIBRERÍAS

Pueden utilizar cualquier librería en python implementadas para recomendación. Las más utilizadas son pyreclab, surprise e implicit, pero esto queda a su criterio.

Para recomendación basada en contenido pueden utilizar funciones de similaridad y de reducción de dimensionalidad ya implementadas en librerías como scikit-learn.

MÉTRICAS DE EVALUACIÓN MODELOS

En esta tarea las métricas que se les pide para evaluar el desempeño de todos los modelos de recomendación son **recall@10, recall@20, MAP@10, MAP@20, ndcg@10 y ndcg@20**.

Importante para la evaluación de todos los modelos de recomendación

En esta tarea se considerarán como relevantes los productos que el usuario compró y no relevantes en otro caso.

ACTIVIDAD 1: ANÁLISIS EXPLORATORIO (15 %)

Para comprender de mejor manera el conjunto de datos e identificar las relaciones entre las variables es que en esta actividad se le pide hacer el siguiente análisis exploratorio sobre los datos de training:

- Grafique la distribución del número de compras por usuario, identifique los ids de los 10 usuarios más activos en el dataset. Comente la forma de la distribución obtenida y qué porcentaje de las interacciones han sido hechas por estos 10 usuarios.

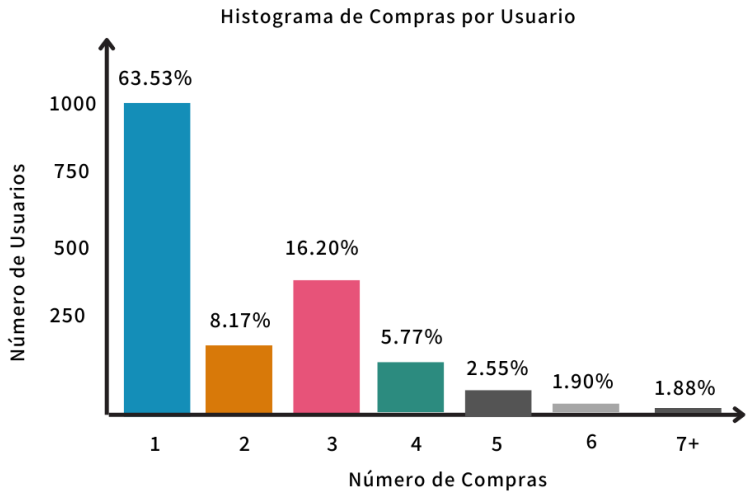


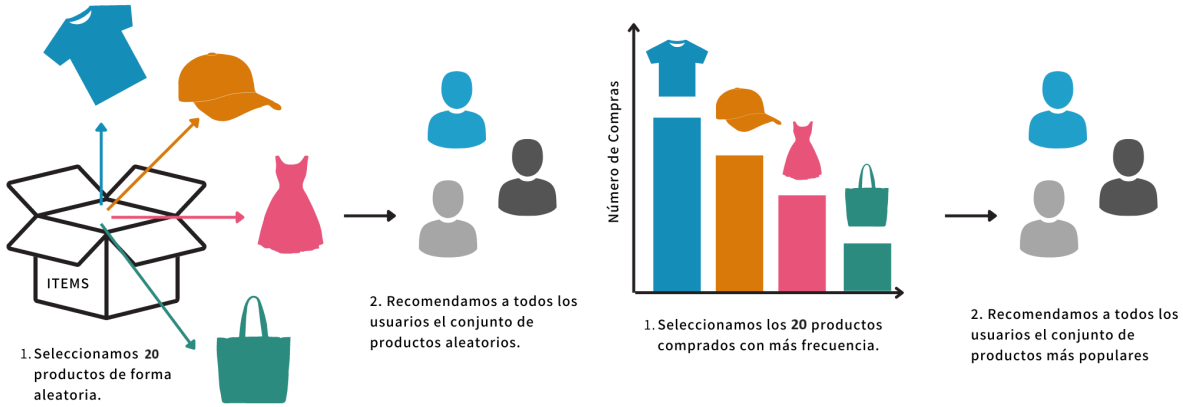
Figura 1: Ejemplo de gráfico de distribución, en este caso de compras por usuario. Haga algo similar para la cantidad de interacciones en el dataset de training de la tarea.

- Grafique la distribución de interacciones por producto. Identifique los nombres y ids de los 10 productos que han sido más comprados. Comente la forma de la distribución y qué porcentaje de las interacciones han sido sobre estos 10 productos.
- Genere una tabla resumen con el número de usuarios distintos, el número de productos distintos, promedio y desviación estandar de productos por usuario, promedio y desviación estándar de usuarios por productos y densidad del conjunto de datos (o *sparsity*) en cuanto a compras.

ACTIVIDAD 2: RECOMENDACIÓN NO PERSONALIZADA (10 %)

En esta actividad el objetivo es realizar dos recomendaciones. Primero se pide recomendar los 20 productos más populares (*most popular*) y luego realizar una recomendación de 20 productos escogidos de manera aleatoria (*random*). Por lo general estos métodos no personalizados se utilizan como baseline para comprobar que los métodos utilizados funcionan bien y tienen un buen rendimiento.

Se les pide calcular métricas de evaluación en el dataset de validación para ambos métodos no personalizados: *random* y *most popular*, seguido de una comparación y análisis de los resultados obtenidos.



(a) Recomendación aleatoria **(b)** Recomendación de items más populares.

Figura 2: Ejemplos de recomendación de *baseline*

ACTIVIDAD 3: RECOMENDACIÓN BASADA EN FEEDBACK IMPLÍCITO (30 %)

En esta actividad el objetivo es recomendar basándose en los productos con los que ha interactuado el usuario.

En este caso utilizarán dos modelos: Factorización Matricial optimizada con Alternate Least Squares (ALS) y Factorización Matricial optimizada con Bayesian Personalized Ranking (BPR).

Se le pide:

- Mostrar un análisis de sensibilidad de resultados en el dataset de validación para métricas recall@10, MAP@10 y nDCG@10 modificando dimensión de factores latentes (50,100,200,500,1000) y el algoritmo de optimización utilizado (ALS o BPR). Grafique cada uno, compare los resultados y comente los resultados obtenidos.
- Reportar y analizar tiempos de entrenamiento en cada uno de los casos, así como comentar sobre los recursos utilizados para ejecutar los modelos, especificando si se emplearon recursos de CPU, GPU u otros.

ACTIVIDAD 4: RECOMENDACIÓN BASADA EN CONTENIDO (30 %)

En esta actividad el objetivo es recomendar basándose en el contenido de los elementos que el usuario ha comprado.

Para ello les entregamos una serie de datos de los productos, como lo son el color, tipo de producto, apariencia gráfica, descripción del producto, entre otros.

Para llegar a la recomendación, en esta actividad les pedimos:

1. Calcular un embedding del texto para cada item utilizando la información disponible de la columna *detail_desc* de articles.csv[1]. Para esto, les pedimos que utilicen *Contrastive Language-Image Pre-Training* link o *Universal Sentence Encoding* link, ambos disponibilizados abiertamente en internet. Puede usar otros encoders de texto, pero debe especificar los detalles del modelo, dimensionalidad, URL de origen, etc. (BERT, ELECTRA, etc.)

2. Reducir dimensionalidad de los vectores o embeddings de productos utilizando PCA u otro algoritmo de reducción, el objetivo de este paso es para que el cálculo de la similaridad sea lo más eficiente posible al momento de hacer la recomendación.

3. Calcular una representación vectorial de cada usuario como el promedio de vectores de productos con los que ha interactuado el usuario en el dataset de entrenamiento.

4. Hacer la recomendación calculando alguna métrica de similaridad entre el vector del usuario y los vectores de cada producto.

   En la figura 3 se resumen los pasos a seguir para recomendar.

Finalmente, se les pide:

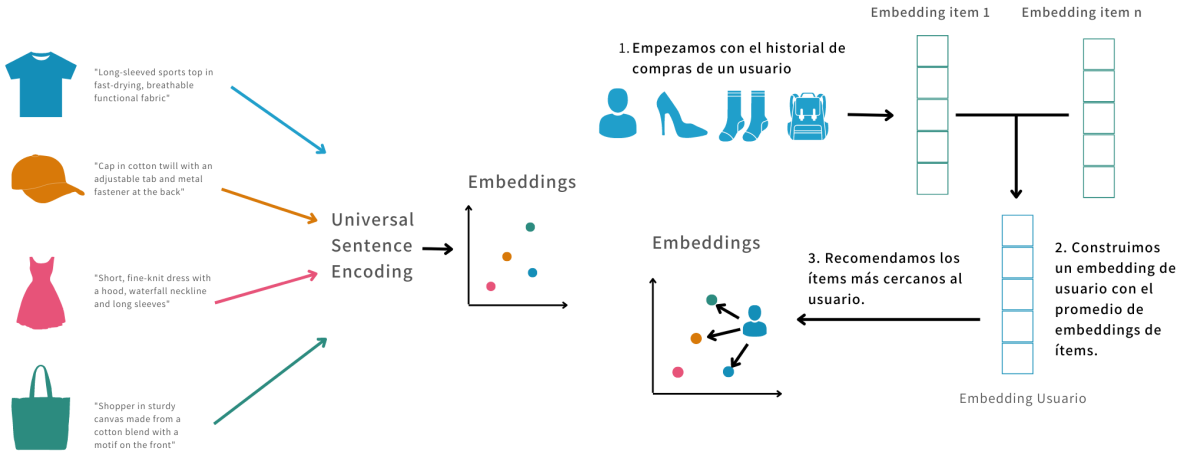
- Hacer un análisis de sensibilidad de resultados en el dataset de validación para métricas recall@10, MAP@10 y nDCG@10 modificando dimensión de los vectores luego de reducir dimensionalidad (10,50,100) y métricas de similaridad (coseno, euclideana y manhattan). Grafique cada uno, compare y comente los resultados obtenidos.

ACTIVIDAD 5: COMPARACIÓN DE MÉTODOS (15 %)

En esta actividad se le pide:

- Hacer una tabla comparativa de los resultados de las métricas solicitadas en el dataset de validación para el mejor modelo de recomendación (con mejor combinación de hiperparámetros) de cada uno de los métodos vistos, es decir recomendación no personalizada (Random y Most Popular), basada en interacciones (Matrix Factorization ALS y Matrix Factorization BPR) y basada en contenido (solo el que obtuvo los mejores resultados). Recuerde mostrar en la tabla la mejor combinación de hiperparámetros de cada modelo que los llevaron a obtener dichos resultados.

1 Si desea, puede usar información de la metadata disponible en otras columnas de items.csv para generar embeddings más informativo



(a) Generación de *Embedding* para cada producto

(b) Utilización *embeddings* de productos para generar representación de usuario

Figura 3: Pasos a seguir para generar recomendación basada en contenido.

- Hacer un análisis y discusión de los resultados que expliquen posibles razones que puedan estar incidiendo en los resultados obtenidos.
- Seleccione e indique el mejor de todos los métodos (a su parecer) y con este, genere 10 recomendaciones por cada usuario del dataset de testing. La lista de recomendaciones **debe** ser entregada en este formato:

```
—
"user_id1": [item_id1, item_id2192, ...],
"user_id2": [item_id121, item_id234, ...],
"user_id3": [item_id191, item_id223, ...],
...
"
```

En base a estas predicciones calcularemos las métricas finales de resultado para cada pareja (las mejores 5 obtienen una bonificación a la nota final).

ACTIVIDAD **BONUS**: MÉTODOS DE ENSAMBLAJE Y UTILIZACIÓN DE TEXTO + IMÁGENES (2 PTOS.)

Como tarea opcional, debe generar un modelo de ensamblaje (ensamble learning) que combine uno de los métodos de recomendación basada en feedback implícito (ALS) o (BPR) con recomendación basada en contenido, aprovechando la información textual de la actividad 4, y visual proporcionada por las imágenes de los artículos disponibles en el archivo *images_rescaled.zip*.

- Un método de ensamblaje es un algoritmo que combina dos o más algoritmos de aprendizaje para mejorar su estabilidad o predictibilidad. Existen distintas formas de realizar un modelo de ensamblaje (bagging, random forests, AdaBoost, etc.) Ustedes pueden elegir la forma que les parezca más conveniente para la realización de este ejercicio. En el siguiente link pueden encontrar un tutorial de Simplilearn que les puede servir para entender el concepto y las distintas formas de realizar ensemble learning
- Para utilizar la información de las imágenes, deben generar embeddings de las imágenes utilizando una red neuronal (por ejemplo, CLIP o ResNet50), y, con estos embeddings, repetir el procedimiento de la Actividad 4.
- Incluir un análisis de ablation del modelo, especialmente en sus componentes generales con interacciones, el contenido textual y las imágenes. Permitiendo una comprensión más profunda del funcionamiento y la contribución de cada aspecto del modelo en el problema.
- Pueden utilizar el modelo que obtengan en esta actividad para generar las recomendaciones del último paso de la Actividad 5 (lo que les dará mejores chances de quedar entre las mejores 5)

ENTREGABLES

La tarea deberá ser entregada a través de la plataforma CANVAS, se les solicita enviar los siguientes archivos:

Se deberá ENTREGAR un informe en formato PDF, así como código en uno o varios **Jupyter Notebook con todas las celdas ejecutadas**, es decir, no se debe borrar el resultado de las celdas antes de entregar. Si las celdas se encuentran vacías, se asumirá que la celda no fue ejecutada. Es importante que toda la información solicitada de parámetros y análisis tengan una explicación, es decir, no basta con el *output* de una celda para responder una pregunta, se debe explicar qué se está respondiendo.

Informe. Junto con el código debe estar un informe en formato PDF que contenga el desarrollo de cada una de las actividades solicitadas. Es importante que el informe sea autocontenido, y que todos los resultados y figuras incluídos en el informe estén respaldadas por el código (es decir, el informe y el código deben ser consistentes entre sí).

Código. Por cada uno de los métodos solicitados debe entregar el código que permita replicar los resultados obtenidos. Se solicita entregar uno o varios jupyter notebooks que permitan replicar experimentos.

Es obligatorio agregar un archivo README.md que permita entender la estructura de archivos y detalles necesarios para replicar los experimentos realizados.