Tópicos avanzados en analítica

Proyecto No 2

Autores:
Ian Nicolas Rincón, Rosemary Ríos Pulido,
Carlos Felipe Mora, John Sebastián Martínez
Andrés Parra Rodríguez [a]

Docente:
Sergio Mora

[a] Estudiantes de Maestría en Analítica para la Inteligencia de Negocios
Departamento de Ingeniería Industrial
Pontificia Universidad Javeriana
Bogotá, Colombia

Descripción General

Una empresa de comercio electrónico (Amazon) desea mejorar la clasificación de productos para incrementar las ventas y mejorar la experiencia del usuario. Actualmente, el sistema de clasificación se basa únicamente en productos comprados juntos, pero desean implementar un sistema más avanzado que también tenga en cuenta las características de los productos y las relaciones entre ellos.

Objetivo General

Implementar un sistema de clasificación basado en grafos utilizando el dataset Amazon de PyTorch Geometric. Este sistema utilizará tanto la estructura de los gráficos (relaciones entre productos) como los atributos de los nodos (características de los productos) para proporcionar recomendaciones más precisas y relevantes.

Objetivo Específicos

- Mejorar la precisión de las clasificaciones de los 10 productos; al utilizar un modelo de aprendizaje sobre grafos, se puede capturar mejor la estructura compleja de las relaciones entre productos y las características de estos.
- Incrementar las ventas; recomendaciones más precisas y relevantes pueden llevar a un aumento en las ventas al ofrecer productos que los usuarios están más inclinados a comprar.
- Mejorar la experiencia del usuario; usuarios satisfechos con las recomendaciones son más propensos a regresar y utilizar la plataforma de nuevo, mejorando la retención de clientes.

Relevancia en su Contexto

Amazon, como líder en comercio electrónico, busca constantemente optimizar su sistema de clasificación para mantenerse competitivo y mejorar la experiencia del usuario. El contexto de esta mejora se centra en varios aspectos clave:

- **Competitividad en el Mercado:** En un entorno altamente competitivo, ofrecer recomendaciones precisas puede ser un diferenciador crucial. Los usuarios esperan recibir sugerencias que sean relevantes y útiles, lo que puede influir en su decisión de compra y lealtad hacia la plataforma.

- **Volumen y Diversidad de Productos:** Amazon tiene un vasto catálogo de productos, lo que hace que la clasificación precisa sea un desafío significativo. Un sistema basado únicamente en productos comprados juntos es limitado y no captura toda la riqueza de las interacciones posibles.

- **Comportamiento del Usuario:** La satisfacción del usuario se correlaciona con la personalización de sus experiencias de compra. Un sistema de clasificación mejorado puede aumentar la retención de usuarios y fomentar compras repetidas.

Utilidad de un Modelo de Grafos

Un modelo de grafos es particularmente útil para el problema de clasificación de productos por varias razones:

- Los grafos permiten modelar las relaciones complejas entre los productos de una manera que los métodos tradicionales no pueden. Por ejemplo, no solo se puede ver qué productos se compran juntos, sino también cómo están relacionados a través de características compartidas o patrones de compra de diferentes usuarios.

- Cada producto puede ser representado como un nodo en el grafo con atributos (características del producto). Esto permite al sistema considerar tanto las relaciones (aristas) entre productos como las propiedades individuales de cada producto, proporcionando una visión más holística y precisa para las recomendaciones.

- Los modelos de grafos son escalables y pueden adaptarse fácilmente a cambios en el catálogo de productos. Nuevos productos y relaciones pueden ser integrados en el sistema sin necesidad de rediseñar el modelo desde cero.

- Al utilizar un enfoque de aprendizaje sobre grafos, se puede aprovechar el poder del aprendizaje profundo para extraer patrones y representaciones latentes de alta calidad. Esto mejora significativamente la precisión de las recomendaciones en comparación con métodos más simples.

- Los grafos permiten una personalización más avanzada. Por ejemplo, si un usuario ha mostrado interés en varios productos que están estrechamente relacionados en el grafo, el sistema puede recomendar productos que estén un poco más alejados, pero aún relacionados, ampliando el alcance de las sugerencias relevantes.

Conjunto de Datos Para Emplear

El conjunto de datos utilizado proviene del dataset Amazon proporcionado por PyTorch Geometric, disponible en el siguiente enlace " https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Amazon.html?highlight=amazon" . Este dataset contiene información sobre productos, incluyendo

las reseñas de los usuarios que son utilizadas como características de los nodos en el modelo.

PyTorch Geometric es una extensión de PyTorch diseñada para el aprendizaje profundo en datos que tienen una estructura de grafo. Proporciona herramientas y modelos predefinidos para facilitar la implementación y experimentación con técnicas avanzadas de análisis de grafos, como Graph Neural Networks (GNNs), Graph Convolutional Networks (GCNs), y Graph Attention Networks (GATs).

El dataset de Amazon disponible en PyTorch Geometric es un conjunto de datos utilizado para tareas de análisis de grafos y aprendizaje profundo. Este dataset incluye información detallada sobre productos, sus reseñas, y las relaciones de compra conjunta entre productos.

Contenido del Dataset

- Productos como Nodos: Cada producto en Amazon es representado como un nodo en el grafo.
- Relaciones de Compra Conjunta como Bordes: Los bordes entre nodos indican que dos productos se compran frecuentemente juntos.
- Reseñas como Atributos de Nodos: Las características de los nodos se derivan de las reseñas de los productos, las cuales son procesadas en una representación de bolsa de palabras.

ANÁLISIS DESCRIPTIVO

Grafica de grafo

Cada nodo en el grafo representa un accesorio para computadores, y una arista entre dos nodos indica que los productos son frecuentemente comprados juntos. Los atributos de los nodos representan diversas características de los productos.

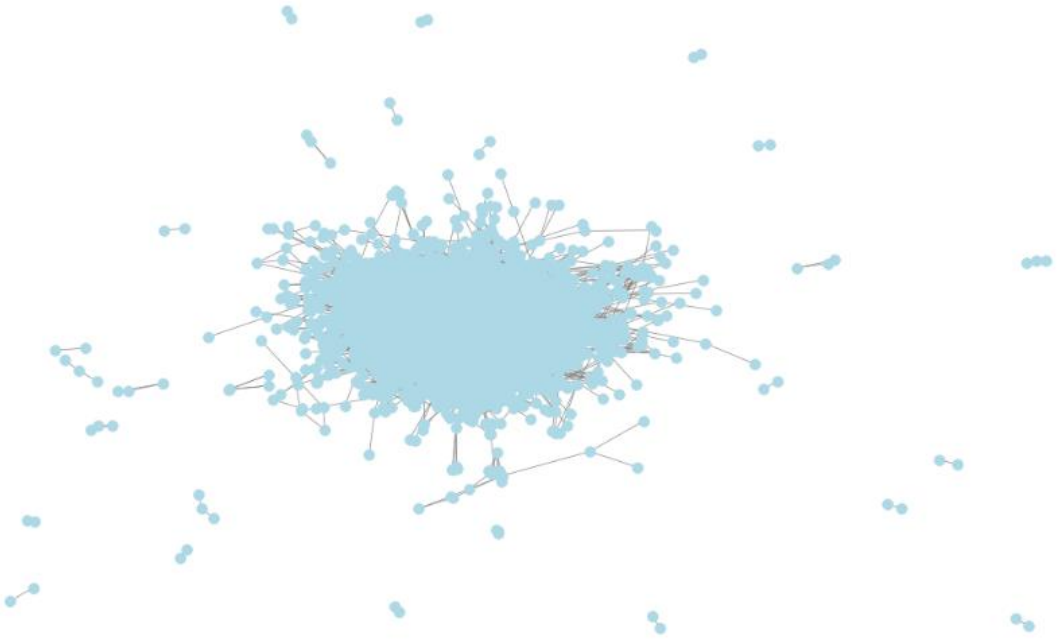

Ilustración 1. Grafica del grafo

Se presentan 491.722 posibles combinaciones de productos relacionados a los computadores donde se presenta 2 columnas que indica un producto en cada columna y el registro muestra la relación entre estos productos para ver si son complementarios o sustitutos.

Hay 13.752 nodos (productos de computadores) en el grafo y cada nodo tiene 767 características, lo que indica que cada producto tiene un vector de 767 características que puede incluir información como descripción del producto, especificaciones técnicas, precios, reseñas, etc. Hay 491,722 aristas en el grafo, esto quiere decir el número de veces en que fueron comprados dos productos juntos. Cada accesorio de computador está clasificado en una de las 10 clases diferentes. Las etiquetas de los nodos están en el rango de 0 a 9. Cada nodo tiene un vector de características de longitud 767. No hay características asociadas a las aristas.

Para el análisis descriptivo se analizó en 2 etapas:

Etapa inicial con el 97,9% de los nodos en donde se presenta lo siguiente:

- 245.861 relaciones entre los accesorios de computador (representa el 50% de compras de 2 productos al tiempo).
- En promedio los productos de computadores aparecen comprados con otro producto complementario en 36 ocasiones.
- El camino más corto del producto cero al producto 13471 se necesita en promedio 3 relaciones del producto cero con otros productos para que llegue al producto 13471.

- Se necesita en promedio 3 relaciones del producto cero con otros productos para que se elija al mismo tiempo el producto intermedio (6736).
- El camino más corto con mayor frecuencia en el grafo es una compra de 3 productos al tiempo ya que representa más del 40% de los nodos de acuerdo con el siguiente histograma:

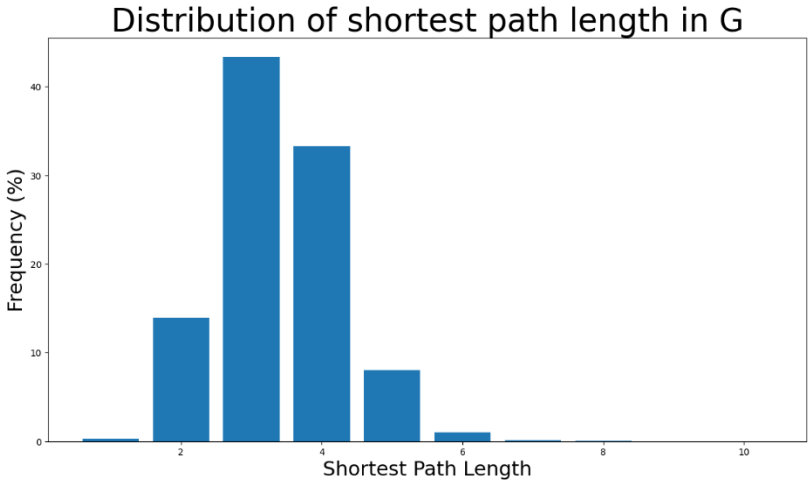

Ilustración 2. Histograma del camino más corto

- El diámetro del grafo, considerando solo las componentes conectadas, es 10. Esto significa que el componente conectado más extensa del grafo, la mayor distancia entre cualquier par de nodos es 10 aristas, evidenciando que las máximas relaciones que se pueden dar entre productos es de 10 ocasiones con el objetivo de que se conecte un producto a otro
- En el grado de densidad existe aproximadamente el 0.27% de todas las posibles conexiones entre productos. Esto quiere decir que los productos no están altamente interconectados en el grafo por razones como que los clientes no suelen comprar grandes cantidades de productos juntos o que existen diferentes grupos de productos que no tienen muchas conexiones entre sí.
- El grafo se divide en 33 grupos de productos interconectados, pero que no tienen conexiones directas entre ellos.
- Los productos en Amazon forman diferentes grupos o comunidades que están conectados internamente pero no tienen conexiones directas con otros grupos.
- Una densidad baja y un número relativamente alto de componentes conectadas sugieren una estructura dispersa con grupos de productos interconectados internamente pero no directamente entre sí.

Top de productos	Grado de centralidad	Frecuencia de compra con otros productos
12.888	22,2%	2.992
8.210	18,6%	2.508
8.140	18,5%	2.495
4.528	13,7%	1.850
1.524	11,6%	1.565

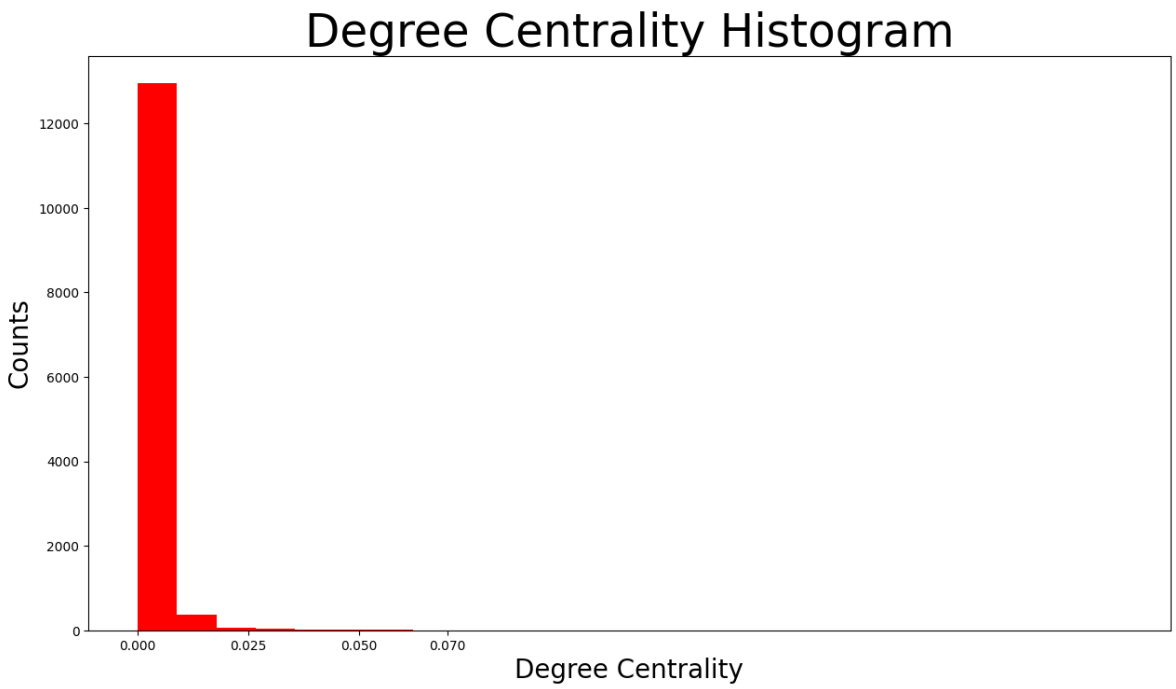

Ilustración 3. Histograma de grado de centralidad

- En el gráfico anterior se evidencia que hay pocas conexiones directas para 13 mil productos, mientras que menos de 100 productos tienen una conexión directa del 2,5%.

Segunda etapa se corre con una muestra aleatoria de 1000 nodos para las medidas de centralidad intermedia, centralidad de cercanía y centralidad del vector propio

Centralidad intermedia	Centralidad de cercanía	Centralidad del vector propio
8.140	8.140	8.140
8.872	7.939	4.135
904	11.262	13.160
7.797	4.135	3.067
11.262	3.697	5.715

- El producto 8140 es crucial en la red de compras de Amazon para accesorios de computadores. Es un producto central que conecta muchos otros productos (alta intermedia), es accesible desde muchos otros productos (alta cercanía), y está asociado con productos influyentes (alta vector propio). Es probable que este producto sea un accesorio muy popular y esencial que los compradores suelen agregar a sus carritos junto con muchos otros artículos.
- El producto 11262 actúa como un conector esencial (alta intermedia) y es fácilmente accesible desde muchos otros productos (alta cercanía). Sin embargo, a diferencia del producto 8140, este producto puede no estar conectado a otros productos influyentes, sino que su importancia radica más en su rol de facilitador de conexiones en la red.
- El producto 4135 es muy accesible desde muchos otros productos (alta cercanía) y está asociado con productos influyentes y populares (alto vector propio). Aunque no siempre actúa como un conector clave en la red (como en el caso de alta intermedia), es un producto popular bien conectado y quizá sea un accesorio de alta demanda comprado con otros productos importantes.
- En promedio Un 35,1% de los productos están relativamente bien conectados con sus vecinos. Significa que existe una tendencia a que los productos se agrupen o formen comunidades dentro de la red de compras.
- Mas de 700 casos tienen una probabilidad de agruparse cerca de un 38%.
- Se presenta 4.582.407 combinaciones de 3 productos que se compran al mismo tiempo.
- En promedio hay 340 canastas de compras donde se compran los mismos 3 productos a la vez.
- Existen 342 puentes en el grafo de compra de Amazon, lo que quiere decir que hay muchas conexiones críticas en el grafo de productos de Amazon.
- Existen 6406 puentes locales en el grafo de productos de Amazon indicando la presencia de muchas conexiones críticas para la conectividad de nodos vecinos.
- Se presenta una tendencia de disasortativity, lo que significa que los productos populares (con alto grado) tienden a estar conectados con productos menos populares (con bajo grado), y viceversa.
- Se identificaron 136 comunidades en el grafo.

Ilustración 4. Grafo segmentado por nodos

Para el grafo anterior se quitó los nodos sin conexiones para ver cómo se agrupan en comunidades las compras de Amazon y se evidencia que las comunidades más grandes son la rosada y verde oscuro las que tienen un comportamiento de compras similares.

Atributos topológicos básicos

- En promedio, un nodo está conectado a casi 36 nodos más, esto quiere decir que en promedio los productos de computadores aparecen comprados con otro producto complementario en 36 ocasiones.
- Se calcula el camino más corto del producto cero al producto 13471 en donde se interpreta que, para pasar de un producto a otro de forma más rápida, se necesita en promedio 3 relaciones del producto cero con otros productos para que llegue al producto 13471.
- Diámetro del grafo (ignorando nodos sin conexiones): 10
- Densidad: Existe aproximadamente el 0.27% de todas las posibles conexiones entre productos. Esto quiere decir que los productos no están altamente interconectados en el grafo por razones como que los clientes no suelen comprar grandes cantidades de productos juntos o que existen diferentes grupos de productos que no tienen muchas conexiones entre sí.
- El grafo se divide en 33 grupos de productos interconectados, pero que no tienen conexiones directas entre ellos.

MODELO DE GRAFOS

Modelo	Preprocesamiento de los datos	Definición del modelo	Entrenamiento del modelo	Accuracy
Red neuronal (GNN sin early stop)	En el proceso de construcción de un grafo utilizando las bibliotecas pandas y NetworkX, uno de los pasos fundamentales es la creación de un DataFrame que almacene la información de las aristas del grafo Se definen las máscaras de entrenamiento, validación y prueba para dividir el conjunto de datos en tres partes distintas. Los nodos del grafo se dividen en conjuntos de entrenamiento (60%), validación (30%) y prueba (10%). Se aplica cada máscara al conjunto de datos para seleccionar los nodos correspondientes para cada conjunto.	Se calcula la matriz de adyacencia densa a partir de los bordes del grafo usando la función to_dense_adj de Torch Geometric. Se crea una clase SparseLayer que define una capa lineal que opera en datos dispersos y que se utiliza para construir el modelo de GNN. Se define una clase GNN que representa un Grafo Neural Network (GNN) básico. El modelo de GNN consiste en dos capas de SparseLayer con activación ReLU entre ellas. La salida final se pasa a una función de activación softmax para obtener probabilidades de clase.	Se utiliza la pérdida de entropía cruzada como función de pérdida y el optimizador Adam con un coeficiente de decaimiento de peso de 5e-4. El modelo se entrena iterando a través de un número específico de épocas. En cada época, se calcula la pérdida y se realiza la retropropagación para actualizar los pesos del modelo. Se monitorea la precisión del entrenamiento y la pérdida de validación después de cierto número de épocas. Si no hay mejora en la precisión de la validación durante 5 épocas consecutivas, se detiene el entrenamiento temprano.	3,39%
Red neuronal 2 (GNN con early stop)	En el proceso de construcción de un grafo utilizando las bibliotecas pandas y NetworkX, uno de los pasos fundamentales es la creación de un DataFrame que almacene la información de las aristas del grafo Se definen las máscaras de entrenamiento, validación y prueba para dividir el conjunto de datos en tres partes distintas. Se	Se define un modelo de Red Neuronal de Grafos (GNN) utilizando la arquitectura de capas lineales definida en la clase SparseLayer. El modelo GNN consta de dos capas de SparseLayer con activación ReLU entre ellas. La salida final se pasa a una función de activación softmax	Se entrena el modelo GNN utilizando el conjunto de entrenamiento y se valida en el conjunto de validación. La pérdida se calcula utilizando la función de pérdida de entropía cruzada. Durante el entrenamiento, se monitorea la precisión del entrenamiento y la pérdida de validación después de cada 20	45,4%

	asigna aproximadamente el 80% de los datos al entrenamiento, el 10% a la validación y el 10% restante a la prueba. Se aplica cada máscara al conjunto de datos para seleccionar los nodos correspondientes para cada conjunto.	para obtener probabilidades de clase. Se utilizan los siguientes hiperparámetros: Dimensión de entrada: igual al número de características de los nodos en el conjunto de datos. Dimensión oculta: 16. Dimensión de salida: igual al número de clases en el conjunto de datos. Tasa de aprendizaje del optimizador Adam: 0.01. Coeficiente de decaimiento de peso: 5e-4.	épocas. Se utiliza un criterio de detención temprana basado en la precisión de validación para evitar el sobreajuste. Si la precisión de validación no mejora durante 5 épocas consecutivas, se detiene el entrenamiento anticipado.	
Red neuronal convolucional 1	En el proceso de construcción de un grafo utilizando las bibliotecas pandas y NetworkX, uno de los pasos fundamentales es la creación de un DataFrame que almacene la información de las aristas del grafo Se definen las máscaras de entrenamiento, validación y prueba para dividir el conjunto de datos en tres partes distintas. Se asigna aproximadamente el 80% de los datos al entrenamiento, el 10% a la validación y el 10% restante a la prueba. Se aplica cada máscara al conjunto de datos para seleccionar los nodos	Se define un modelo de Red Neuronal Convolucional para Grafos (GCN). El modelo consta de dos capas de GCN (Graph Convolutional Network), donde se aplica la convolución a los nodos del grafo. Cada capa GCN se complementa con una activación ReLU. La salida final se procesa mediante una función softmax para obtener las probabilidades de clase.	Se entrena el modelo GCN utilizando el conjunto de entrenamiento y se valida en el conjunto de validación. La pérdida se calcula utilizando la función de pérdida de entropía cruzada. Durante el entrenamiento, se monitorea la precisión del entrenamiento y la pérdida de validación después de cada 20 épocas.	72%
Red neuronal convolucional 2		Se define un modelo de Red Neuronal Convolucional para	Durante el entrenamiento, se utiliza una función de pérdida de entropía cruzada.	83,3%

	correspondientes para cada conjunto.	Grafos (GCN) con 2 capas ocultas. El modelo consta de tres capas GCN, donde la segunda capa oculta es añadida para mejorar la capacidad del modelo. Se utiliza la función de activación Leaky ReLU para ambas capas ocultas. La salida final se procesa mediante una función softmax para obtener las probabilidades de clase.	Se aplica un esquema de parada anticipada (early stopping) para detener el entrenamiento si no hay mejora en la pérdida de validación después de cierto número de épocas consecutivas (patience). Se monitorea la pérdida y la precisión en el conjunto de validación después de cada 20 épocas.	
Red convolucional 5 capas con regularización	Los nodos del grafo se dividen en conjuntos de entrenamiento (60%), validación (30%) y prueba (10%).	El modelo definido es una red neuronal convolucional de grafos (GCN) con cinco capas convolucionales. Se utiliza la función de activación tanh después de cada capa GCN. La última capa aplica log_softmax para obtener las probabilidades de clase.	Se utiliza la pérdida de entropía cruzada como función de pérdida y el optimizador Adam con un coeficiente de decaimiento de peso de 5e-4 como regularización. El modelo se entrena iterando a través de un número específico de épocas. Se utiliza para la parada temprana si no hay mejora en la pérdida de validación durante 20 épocas consecutivas.	63%
Modelo de atención con múltiples		Se definen múltiples capas de convolución atencional GATv2, incluyendo la capa de entrada,	Durante el entrenamiento, se utiliza una función de pérdida de entropía cruzada.	81,6%

capas de convolución		capas intermedias y la capa de salida. Dropout: Se aplica dropout con una probabilidad de 0.6 para regularización. Forward: En el método forward, se aplica dropout seguido de convoluciones atencionales y activación ELU en las capas intermedias. La última capa utiliza log_softmax para la salida de clasificación.	Se realiza el entrenamiento por el número de épocas especificado (100 en este caso) y se imprime el rendimiento cada 20 épocas, con una tasa de aprendizaje del 1% junto a una regularización que aplica una penalización de decaimiento de peso con un factor de 0.01 a los pesos durante el proceso de optimización.	
Modelo de atención con dos capas de convolución (GAT 2 capas)		Se define una red neuronal gráfica (GAT) con dos capas de convolución atencional (GATv2Conv). Se aplica dropout, convolución atencional, activación ELU y finalmente log_softmax para la clasificación.		86,9%

Matriz de confusión del modelo ganador

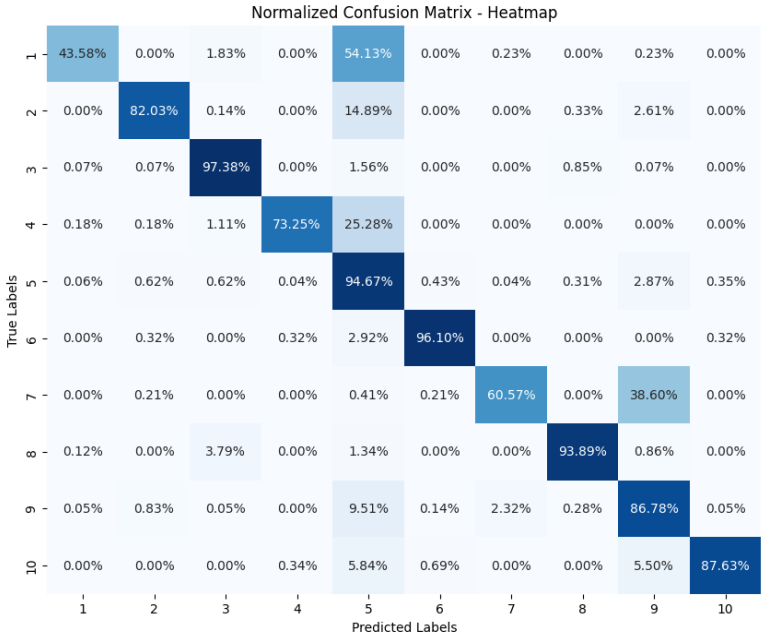

Ilustración 5. Matriz de confusión del mejor modelo (Modelo de atención con dos capas de convolución)

Con el modelo de atención de dos capas de convolución, logramos clasificar los 10 mejores productos para computadoras de Amazon con una precisión notable, evidenciada por un accuracy más alto del 86.9%.

Al evaluar los 10 productos con los datos reales de prueba, se nota que, en 8 de ellos, el modelo muestra una menor incertidumbre en su clasificación, además de predecir más del 70% de los casos correctamente (estos productos incluyen el 2, 3, 4, 5, 6, 8, 9 y 10). Sin embargo, se encuentra que el modelo enfrenta dificultades al predecir las clases de los productos 1 y 7. En particular, tiende a confundir el producto 1 con el producto 5, y el producto 7 con el producto 9.

REPORTE DE TRABAJO EN EQUIPO

Rosemary Ríos Pulido: Código para el análisis descriptivo, parte escrita de la descripción del problema, diapositivas de la parte descriptiva del modelo y construcción del modelo red neuronal 1 (GNN sin early stop).

John Sebastián Martínez: Descripción del problema, comentarios, apoyo del código y documento descriptivo, generación del modelo Red neuronal 2 (GNN con early stop), diapositivas de la parte descriptiva del modelo.

Carlos Felipe Mora: Construcción de los modelos de red convolucional 1 y red convolucional 2, generación de diapositivas y análisis del resultado de los modelos.

Ian Nicolas Rincon: Construcción de los modelos de red convolucional 1 y red convolucional 2, generación de diapositivas y análisis del resultado de los modelos.

Andrés Parra: Identificación de los datos de Amazon, algoritmos de Graph Atention Convolutional (GAT), Organización de códigos.

Fuentes

- PyTorch Geometric. (n.d.). Amazon dataset. PyTorch Geometric Documentation. Retrieved May 19, 2024, from https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Amazon.html?highlight=amazon