UNIVERSIDAD DE GUADALAJARA

Centro Universitario de Ciencias Exactas e Ingenierías

División de Tecnologías para la Integración Ciber-Humana

Análisis de algoritmos

Jennifer Patricia Valencia Ignacio, Código: 223991721

Elizabeth Arroyo Moreno, Código: 221453749

Karla Rebeca Hernández Elizarrarás, Código: 223991977

Ingeniería en computación

Act. 03 Análisis de Clustering con TMAP en base de datos

2 de Octubre de 2025

Introducción

En esta actividad vimos reducción de dimensionalidad y análisis de clusters usando UMAP aplicada al conjunto de datos Fashion-MNIST. La reducción de dimensionalidad sirve para simplificar datos que tienen muchas características, haciendo que sea más fácil visualizarlos y analizarlos sin perder información importante.

Objetivo

El objetivo fue visualizar los datos del dataset Fashion-MNIST reduciendo su dimensionalidad con UMAP para luego identificar y analizar los clusters principales y los subclusters.

Desarrollo

Herramientas Investigadas

Tmap es una forma de hacer pruebas en software, desde que apareció, se ha convertido en una herramienta muy importante porque ofrece una forma clara y ordenada de realizar pruebas, Usar el mismo método ayuda a que el trabajo sea más consistente y eficiente.

La **reducción de dimensionalidad** es una técnica muy útil en el análisis de datos, se trata de transformar datos que tienen muchas características en un espacio con menos dimensiones, pero tratando de conservar la información más importante posible.

Umap (La Aproximación y Proyección de Variedad Uniforme) es una técnica de reducción de dimensionalidad no lineal que sirve para visualizar y explorar conjuntos de datos complejos y de muchas dimensiones, proyectándose en un espacio de dimensiones inferiores

Pandas sirve para el análisis, limpieza, manipulación y exploración de datos tabulares. Permite importar datos de diversas fuentes, como CSV. En este caso, nos sirvió para cargar el archivo fashion-mnist_train.csv, separar las etiquetas y preparar la información antes de usar umap.

Análisis de subclusters:
En el análisis de datos, el término clustering se refiere al proceso de agrupar objetos similares entre sí según un conjunto de características o medidas de distancia. El objetivo principal es descubrir estructuras ocultas dentro de los datos sin necesidad de etiquetas previas.

La existencia de subclusters es un fenómeno común en datos de alta dimensionalidad, donde la complejidad de las relaciones entre variables genera espacios con múltiples regiones de densidad variable.

En el caso del conjunto de datos MNIST, esto significa detectar variaciones dentro de un mismo dígito, como diferentes estilos de escritura del número “3” o trazos más delgados y curvos en algunos “8”, que reflejan la diversidad natural del trazo humano.

Estos subgrupos aparecen porque, aunque las imágenes pertenezcan a una misma clase, las características extraídas de los píxeles pueden presentar diferencias significativas en la forma, orientación o grosor del número.
Por lo tanto, un cluster general como, el correspondiente al dígito “2" puede contener subclusters que representan distintas formas de escribirlo: unos más inclinados, redondeados,más abiertos, etc .

El estudio de subclusters es especialmente importante en datos de alta dimensionalidad como MNIST, donde los vectores generan un espacio complejo y difícil de interpretar.
Para analizar estas variaciones, se emplean técnicas como:

- Clustering jerárquico: organiza los datos en niveles de agrupamiento, permitiendo visualizar relaciones entre clusters y subclusters.
-
- Re-embedding o reducción iterativa de dimensionalidad: se aplica TMAP nuevamente sobre un subconjunto de datos (por ejemplo, solo los dígitos “5”) para observar con mayor detalle su estructura interna.
-
- Análisis multi-resolución: permite examinar los datos a distintas escalas, desde una visión general de todos los números hasta un estudio detallado de un solo dígito.
- Detección de densidad local: identifica zonas más densas dentro de un mismo grupo, revelando estilos o patrones particulares.

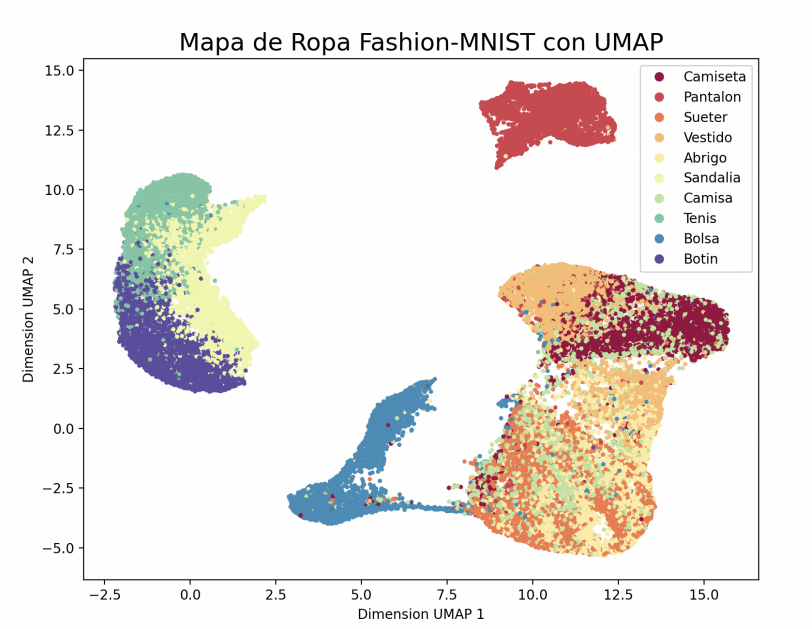
Primero importamos las librerías que se necesitan y con el uso de pandas llama el archivo de fashion-mnist_trani.csv, que contiene miles de imágenes.

```
import pandas as pd # type: ignore
import umap #type: ignore
import matplotlib.pyplot as plt
import numpy as np
from sklearn.cluster import KMeans # type: ignore
import mplcursors # type: ignore
from matplotlib.offsetbox import OffsetImage, AnnotationBbox

df = pd.read_csv('fashion-mnist_train.csv')
```

Separamos las etiquetas (y) de las características(x) y llamamos UMAP para reducir los datos a 2 dimensiones. Esto nos deja visualizar los clusters de ropa en un plano.

```
X = df.drop('label', axis=1)
y = df['label']

label_names = {
    0: 'Camiseta', 1: 'Pantalon', 2: 'Sueter', 3: 'Vestido', 4: 'Abrigo',
    5: 'Sandalia', 6: 'Camisa', 7: 'Tenis', 8: 'Bolsa', 9: 'Botin'
}

reducer = umap.UMAP(n_neighbors=15, min_dist=0.1, n_components=2, random_state=42)
embedding = reducer.fit_transform(X)
```

Después graficamos los resultados, y se puede observar cómo cada tipo de prenda se agrupa en una zona diferente del mapa

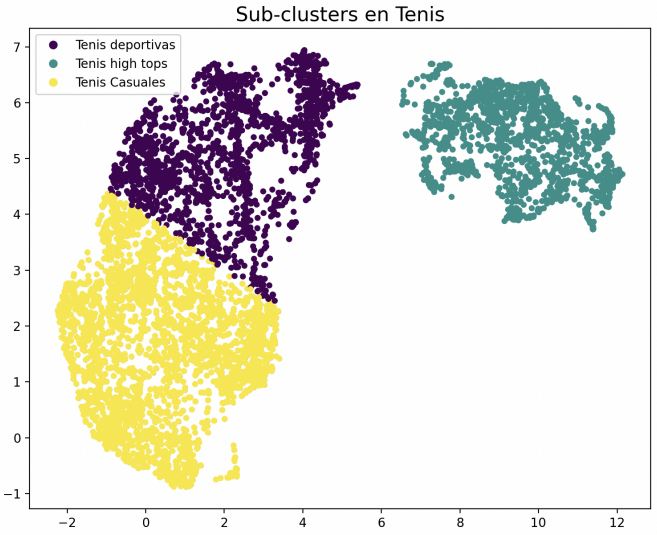


Para la selección de un cluster específico, elegimos el cluster de los tenis (etiqueta 7) y volvimos a aplicar UMAP para tener un mapa más detallado del cluster.

```
id_cluster = 7
nombre_cluster = label_names[id_cluster]
print(f"Vamos a investigar el cluster de: {nombre_cluster}")

X_cluster = X[y == id_cluster]
y_cluster = y[y == id_cluster]

print(f"Encontramos {len(X_cluster)} imágenes de {nombre_cluster}.")

reducer_cluster = umap.UMAP(n_neighbors=10, min_dist=0.05, n_components=2, random_state=42)
embedding_cluster = reducer_cluster.fit_transform(X_cluster)
```

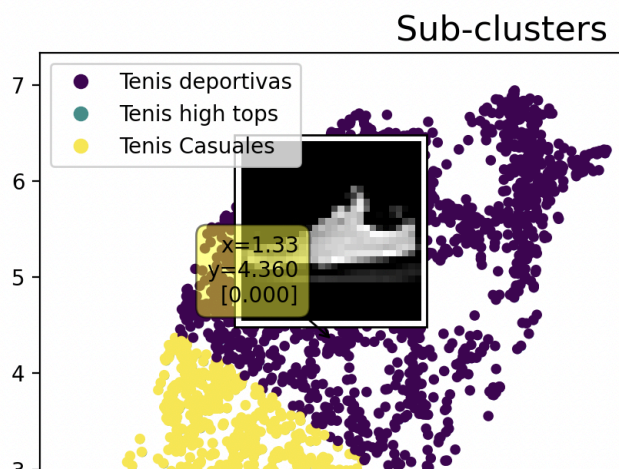
Para crear los sub clusters, usamos el algoritmo de Kmeans para dividir los tenis en subgrupos distintos y después lo graficamos con colores diferentes para identificarlos.

```
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
sub_labels = kmeans.fit_predict(embedding_cluster)

fig, ax = plt.subplots(figsize=(9, 7))
plt.scatter(embedding_cluster[:, 0], embedding_cluster[:, 1], s=10)
scatter_sub = ax.scatter(embedding_cluster[:, 0], embedding_cluster[:, 1], c=sub_labels, cmap='viridis', s=15)
ax.set_title(f'Sub-clusters en {nombre_cluster}',fontsize=16)
ax.legend(handles=scatter_sub.legend_elements()[0], labels=['Tenis deportivas', 'Tenis high tops', 'Tenis Casuales'])

imagenes = X_cluster.to_numpy().reshape(-1, 28, 28)
```

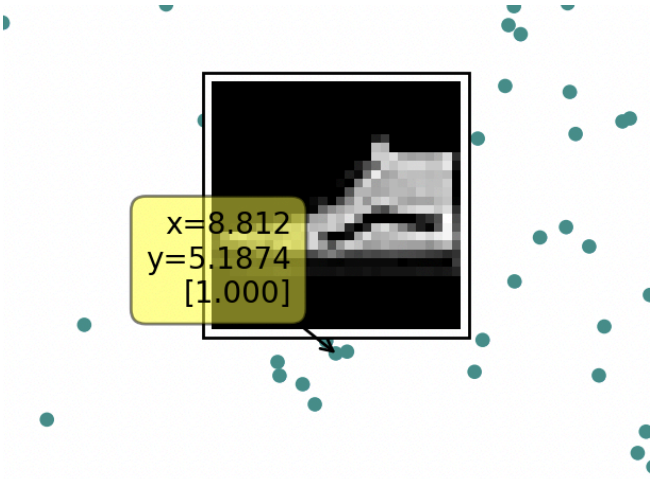


Cada color representa un tipo de tenis diferente según las similitudes encontradas por el modelo. Al analizar las imágenes que forman cada grupo se puede ver un patrón.

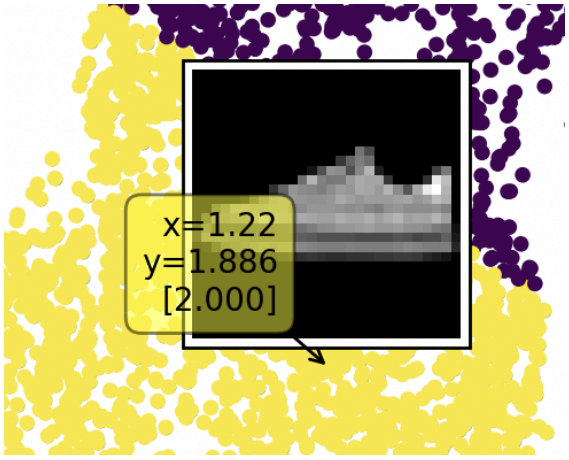
En este grupo, principalmente eran tenis de estilo deportivo. Son más anchos y con suelas más complejas.



En este grupo, los tenis tienen una forma muy distinta porque cubren el tobillo.



Y en este grupo están los tenis “casuales” que son más sencillos en su diseño



Conclusión

Durante esta práctica se logró aplicar de forma efectiva las técnicas de reducción de dimensionalidad y análisis de clustering para explorar el conjunto de datos Fashion-MNIST. El uso de UMAP permitió representar los datos en dos dimensiones, facilitando la visualización de patrones y la identificación de grupos de prendas similares. Posteriormente, con el uso de K-means se pudieron distinguir subgrupos dentro de un mismo tipo de prenda, evidenciando diferencias en el diseño y estilo.

En general, la actividad permitió comprender mejor cómo estas herramientas ayudan a interpretar grandes volúmenes de datos, detectar similitudes y extraer información útil de manera más visual y estructurada.

Referencias

Probst, D. (s. f.). *tmap - Visualize big high-dimensional data*. https://tmap.gdb.tools/

Installing and using UMAP. (s. f.-b). Introduction To Single-cell RNA-seq-ARCHIVED.https://hbctraining-github-io.translate.goog/scRNA-seq/lessons/umap-installation.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5.8 documentation. (s. f.-b). https://umap--learn-readthedocs-io.translate.goog/en/latest/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Alberca, A. S. (2022, 14 junio). *La librería Pandas | Aprende con Alf*. Aprende Con Alf. https://aprendeconalf.es/docencia/python/manual/pandas/

Análisis de clúster: Definición, tipos y ejemplos. (2024, marzo 4). Análisis de clúster: Definición, tipos y ejemplos; forms.app. https://forms.app/es/blog/analisis-de-conglomerados