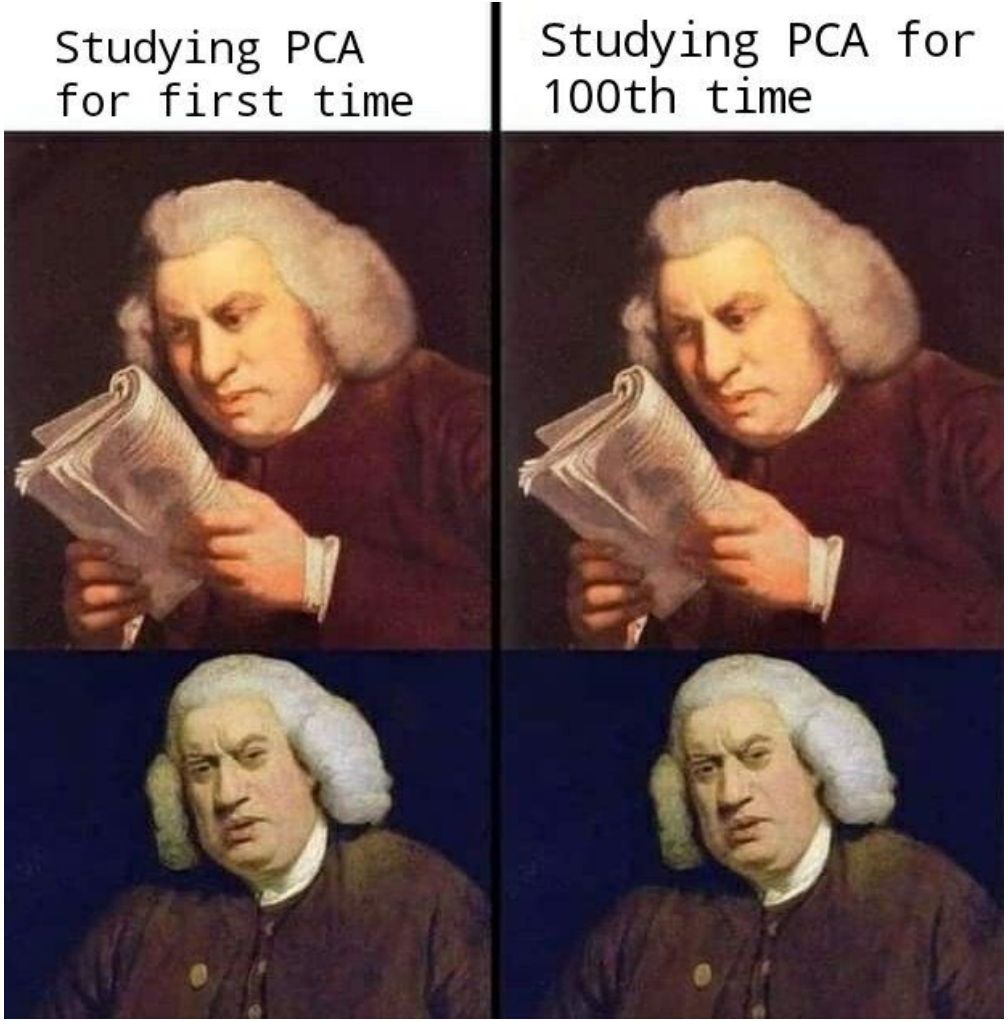
Introducción a la Ciencia de Datos

Maestría en Ciencias de la Computación

Dr. Irvin Hussein López Nava


Studying PCA
for first time
Studying PCA for
100th time

2.4 Reducción de dimensionalidad


¿Qué es
la reducción de
Dimensionalidad?

Definiciones

ChatGPT

Es un proceso que consiste en disminuir el número de características en un conjunto de datos para simplificar su representación y mejorar la eficiencia del análisis o del modelo de aprendizaje automático.

https://chat.openai.com/

Wikipedia

Es la transformación de datos de un espacio de alta dimensión a un espacio de baja dimensión, de forma que la representación de baja dimensión conserve algunas propiedades significativas de los datos originales.

https://en.wikipedia.org/wiki/Dimensionality_reduction

Gemini

Es el proceso de reducir el número de variables, manteniendo la mayor cantidad de información posible. Razones: (i) mejorar el rendimiento, (ii) reducir la complejidad del modelo, (iii) facilitar la visualización de los datos.

https://gemini.google.com/

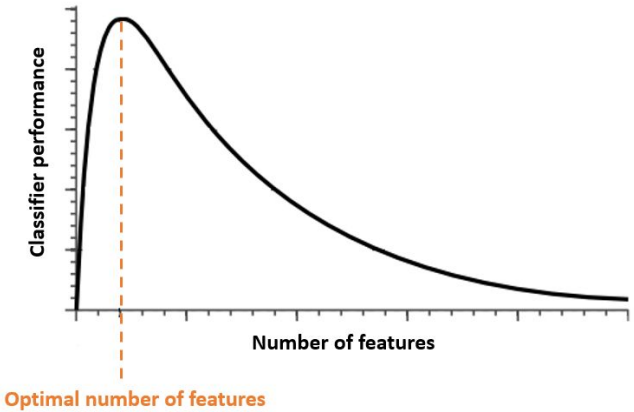
¿Qué es?

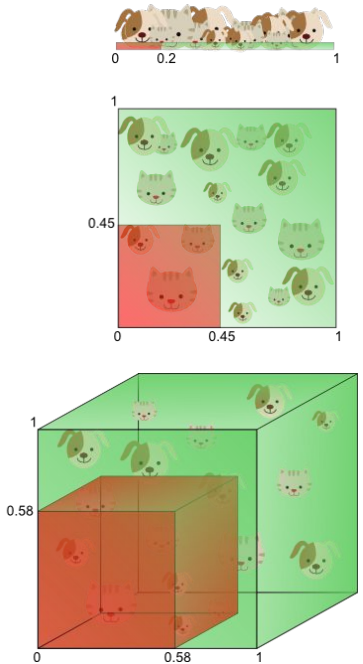
Muchos problemas en Aprendizaje Automático pueden llegar a tener miles de características, esto conlleva muchos problemas, e.g.:

- La fase de entrenamiento puede ser extremadamente lenta.
- Dificulta encontrar una buena solución.

Esto se conoce como la **maldición de la dimensionalidad** y la reducción de dimensionalidad es el proceso de reducir el número de características a las más relevantes en términos simples.

The curse of dimensionality





¿Para qué sirve?

La mayoría de las aplicaciones de reducción de la dimensionalidad se utilizan para:

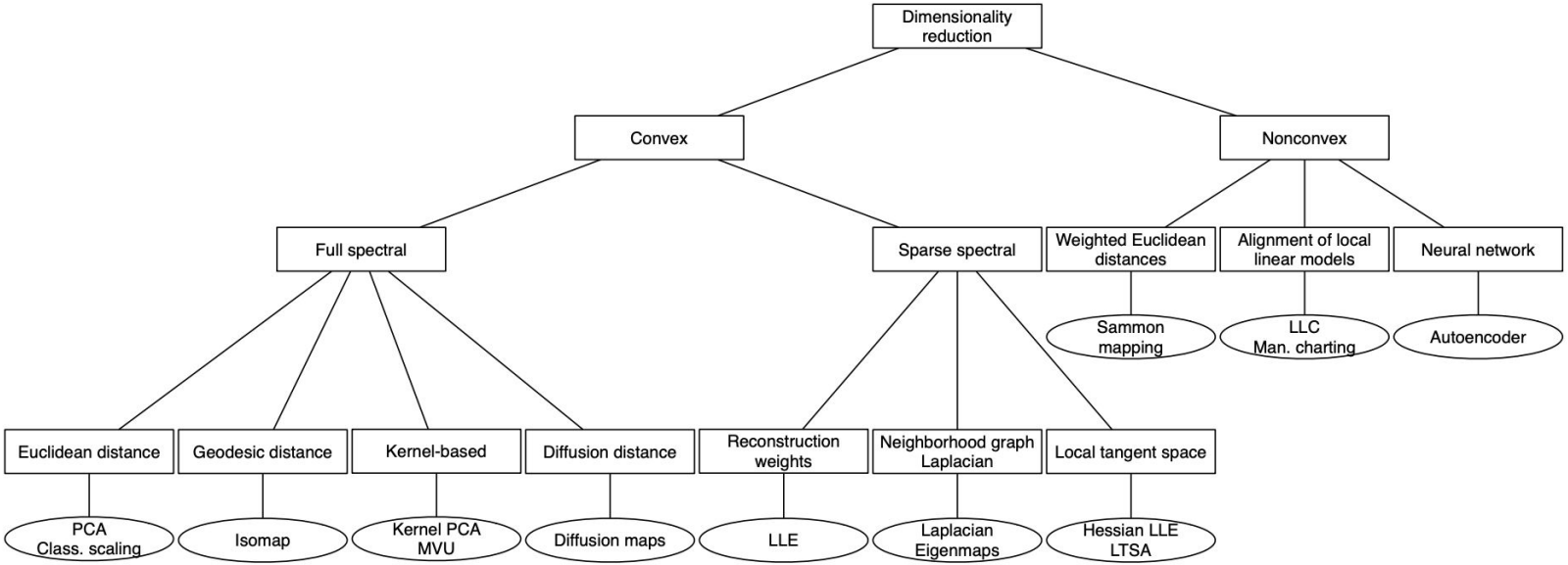
- Compresión de datos
- Reducción del ruido
- Clasificación de datos
- Visualización de datos

Al reducir la dimensionalidad a dos o tres dimensiones, es posible visualizar los datos en un gráfico 2D o 3D, lo que permite obtener información importante analizando estos patrones en términos de clústeres.

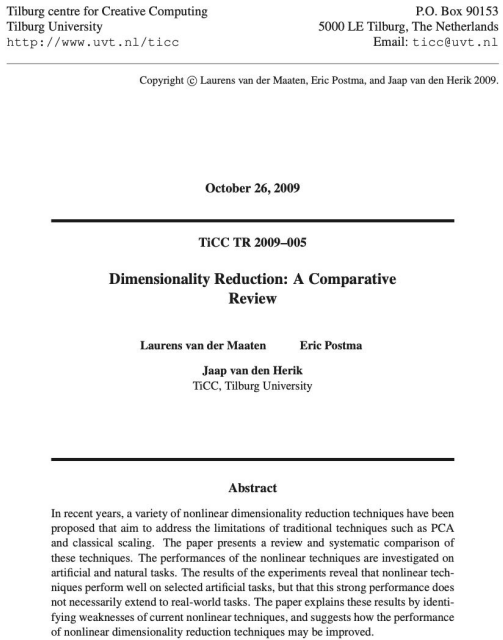
Enfoques

- **Proyección:** esta técnica consiste en proyectar cada punto de datos de alta dimensión en un subespacio de dimensión inferior, de forma que se preserven, ~aproximadamente, las distancias entre los puntos.
- **Aprendizaje múltiple:** se basa en la hipótesis o suposición múltiple (*manifold*), que sostiene que la mayoría de los conjuntos de datos de alta dimensión del mundo real se encuentran cerca de múltiples dimensiones mucho más bajas; en la mayoría de los casos, esta suposición se basa en la observación o la experiencia más que en la teoría o la lógica pura.

Taxonomía de técnicas de reducción de dimensionalidad



Van Der Maaten, L., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Journal of Machine Learning Research, 10(66-71), 13.



Tilburg centre for Creative Computing
Tilburg University
http://www.uvt.nl/ticc

P.O. Box 90153
5000 LE Tilburg, The Netherlands
Email: ticc@uvt.nl

October 26, 2009

TiCC TR 2009–005

Dimensionality Reduction: A Comparative Review

Laurens van der Maaten Eric Postma

Jaap van den Herik
TiCC, Tilburg University

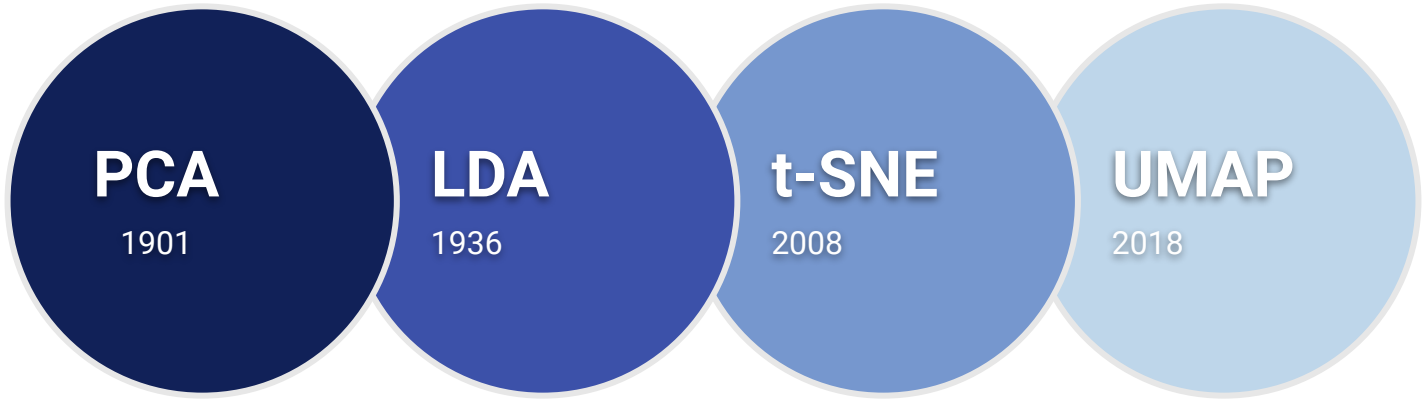
Abstract

In recent years, a variety of nonlinear dimensionality reduction techniques have been proposed that aim to address the limitations of traditional techniques such as PCA and classical scaling. The paper presents a review and systematic comparison of these techniques. The performances of the nonlinear techniques are investigated on artificial and natural tasks. The results of the experiments reveal that nonlinear techniques perform well on selected artificial tasks, but that this strong performance does not necessarily extend to real-world tasks. The paper explains these results by identifying weaknesses of current nonlinear techniques, and suggests how the performance of nonlinear dimensionality reduction techniques may be improved.

Lectura 2

Van Der Maaten, L., Postma, E. O., & van den Herik, H. J. (2009). **Dimensionality reduction: A comparative review.** Journal of Machine Learning Research, 10(66-71), 13.

Algunas de las técnicas más utilizadas



Pearson, K. (1901). Principal components analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 6(2), 559.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.
Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).
McInnes, L., Healy, J., & Melville, J. (2018). **Umap**: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Análisis de Componentes Principales

- Uno de los algoritmos de reducción de la dimensionalidad "no supervisados" más conocidos es PCA (*Principal Component Analysis*).
- El proceso de PCA implica encontrar un conjunto de componentes principales que son combinaciones lineales de las características originales.
    - Estos componentes están ordenados de manera que el primero captura la mayor varianza en los datos, el segundo captura la segunda mayor varianza, y así sucesivamente.
    - Al seleccionar un número menor de componentes que representen la mayoría de la varianza en los datos, se puede reducir significativamente la dimensionalidad del conjunto de datos sin perder información esencial.

Componentes principales

- También conocidos como vectores propios o eigenvectores, son las direcciones en las cuales los datos tienen la mayor variabilidad.
  - Estos componentes son el resultado de una transformación lineal de los datos originales para reducir la dimensionalidad mientras se conserva la información más importante.
- La interpretación de estos componentes es desafiante, ya que generalmente no corresponden directamente a las características originales.

Formalización

- Para explicar formalmente los métodos de reducción de dimensionalidad, es conveniente definir algunos elementos comunes:
  - $X \in R^{n \times d}$ es la matriz de datos, donde n es el número de muestras y d es la dimensionalidad original.
  - $y \in R^{n}$ es el vector de etiquetas para los datos (en los casos supervisados, como LDA).
- El objetivo de cada técnica es encontrar una transformación $T : R^{d} \rightarrow R^{k}$, donde $k < d$, que preserve la información relevante de los datos en un espacio de menor dimensionalidad.

PCA

- PCA busca encontrar una transformación lineal $T \in R^{d \times k}$ que proyecte los datos X a un subespacio de menor dimensionalidad, maximizando la varianza en ese subespacio.
- **Paso 1**: Se centran los datos restando la media de cada dimensión:

$$\mathbf{X}_{\text{centrada}} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{X}$$

- **Paso 2**: Se calcula la matriz de covarianza $C \in R^{d \times d}$:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}_{\text{centrada}}^{\top}\mathbf{X}_{\text{centrada}}$$

PCA

- **Paso 3**: Se realiza la descomposición en valores propios de C:

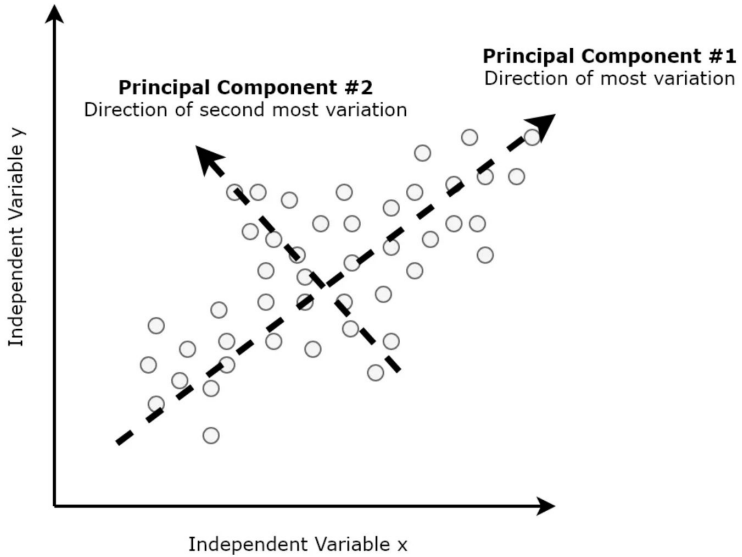
$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

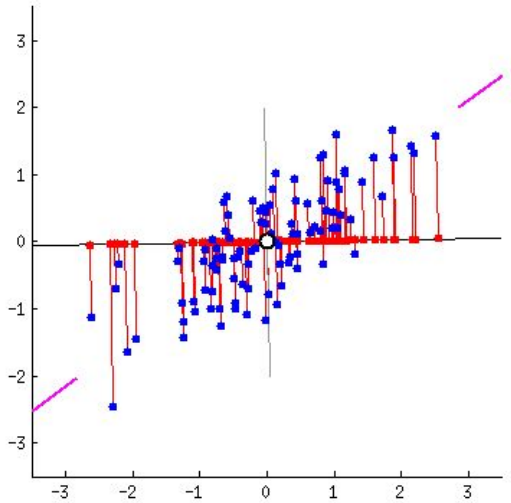
  donde v_i son los vectores propios y λ_i son los valores propios ordenados en orden decreciente. Se seleccionan los k vectores propios correspondientes a los k valores propios más grandes.
- **Paso 4**: La nueva representación de los datos es:

$$\mathbf{X}' = \mathbf{X}_{\text{centrada}} \mathbf{V}_k$$

  donde $V_k \in R^{d \times k}$ contiene los k vectores propios.

PCA





https://medium.com/analytics-vidhya/dimensionality-reduction-principal-component-analysis-d1402b58feb1

Análisis Discriminante Lineal

- A diferencia del PCA, que es una técnica de reducción de dimensionalidad no supervisada que se utiliza para reducir la complejidad de los datos, LDA es una técnica **supervisada** que se utiliza para la clasificación y la maximización de la separación entre clases en un conjunto de datos.
- En LDA, el objetivo es encontrar una transformación lineal que maximice la distancia entre las clases y minimice la dispersión dentro de cada clase.
    - Esto significa que LDA busca proyectar los datos en un espacio de menor dimensión de tal manera que las muestras de diferentes clases estén más separadas entre sí, lo que ayuda para la clasificación.

LDA

- LDA busca maximizar la separabilidad entre clases al proyectar los datos en un subespacio de menor dimensionalidad. Se basa en maximizar la razón entre la dispersión entre clases y la dispersión dentro de las clases.
- **Paso 1**: Se define la matriz de covarianza entre clases (S_B) y la matriz de covarianza dentro de las clases (S_W):

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top$$

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{x \in C_i} (x - \boldsymbol{\mu}_i)(x - \boldsymbol{\mu}_i)^\top$$

donde c es el número de clases, μ_i es la media de la clase i, y μ es la media global.

LDA

- **Paso 2**: Se maximiza la razón de dispersión resolviendo el problema generalizado de valores propios:

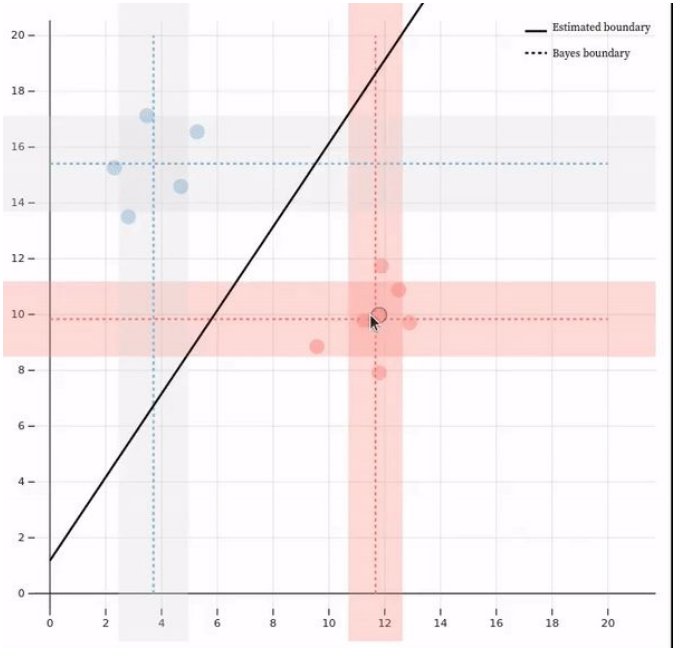
$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

- **Paso 3**: Se seleccionan los k vectores propios correspondientes a los k valores propios más grandes y se proyectan los datos:

$$\mathbf{X}' = \mathbf{X}\mathbf{V}_k$$

donde $V_k \in R^{d \times k}$.

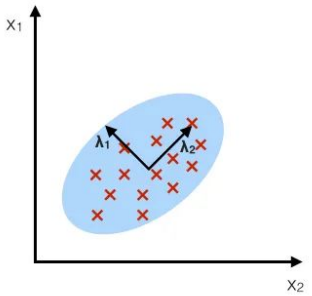
LDA



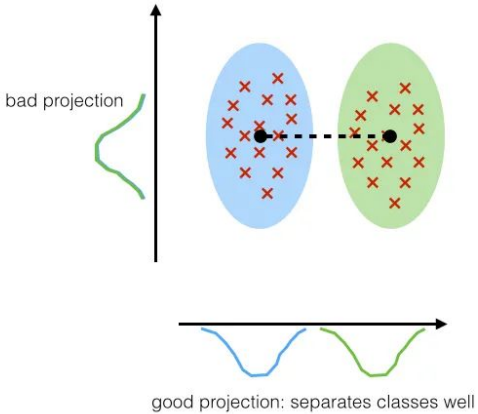
https://towardsdatascience.com/linear-discriminant-analysis-lda-can-be-so-easy-b3f46e32f982

PCA vs. LDA

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



t-distributed Stochastic Neighbor Embedding

- *t-SNE* es una técnica no lineal utilizada para el análisis y la visualización de datos de alta dimensionalidad.
- A diferencia de PCA, t-SNE se especializa en capturar relaciones no lineales y estructuras ‘embebidas’ en los datos.
- t-SNE funciona creando distribuciones de probabilidad conjunta tanto en el espacio de alta dimensión como en el espacio de baja dimensión.
    - Luego, ajusta estas distribuciones para minimizar la divergencia entre ellas, lo que reduce la dimensión al tiempo que mantiene la información sobre las relaciones entre las muestras.

t-SNE

- t-SNE busca preservar las relaciones de vecindad local de los datos, transformando las distancias euclidianas en probabilidades de similitud.
- **Paso 1**: Se define la probabilidad condicional $p_{j|i}$ de que el punto x_j sea el vecino de x_i en el espacio original:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

donde σ_i es el ancho de la vecindad de x_i.

t-SNE

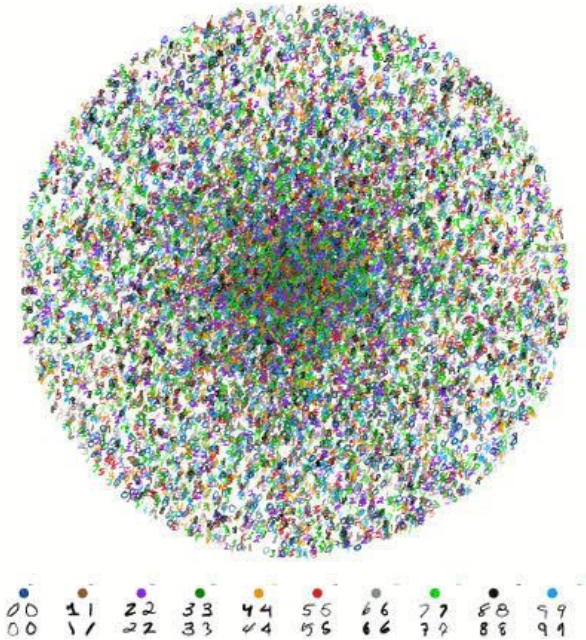
- **Paso 2**: En el espacio reducido, se calculan las probabilidades q_{ij} usando una distribución t de Student con un grado de libertad:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

- **Paso 3**: Minimización de la divergencia de Kullback–Leibler entre las distribuciones de p_{ij} y q_{ij}:

$$\mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Realtime tSNE Visualizations with TensorFlow.js



https://blog.research.google/2018/06/realtime-tsne-visualizations-with.html

Uniform Manifold Approximation and Projection

- UMAP es una técnica no lineal utilizada en el análisis de datos y el aprendizaje automático.
- Al igual que t-SNE, UMAP se especializa en la visualización de datos de alta dimensión en un espacio de menor dimensión, con un enfoque en la preservación de las estructuras complejas y las relaciones entre las muestras.
- Preserva estructuras globales y locales en los datos al mismo tiempo. Se basa en la topología y utiliza conceptos de aprendizaje múltiple para lograrlo.
    - A diferencia de t-SNE, que puede tener dificultades para manejar grandes conjuntos de datos, UMAP tiende a escalar mejor y es más eficiente computacionalmente.

UMAP

- UMAP es también un método no lineal que busca preservar tanto la estructura global como local de los datos.
- **Paso 1**: Se construye un grafo ponderado de vecindades locales, donde la distancia entre puntos está basada en una métrica $d(x_i, x_j)$.
- **Paso 2**: Se define una función de similitud en el espacio de alta dimensionalidad:

$$p_{ij} = \exp(-d(x_i, x_j) - \rho_i)$$

donde ϱ_i es un parámetro que ajusta la densidad local.

UMAP

- **Paso 3**: En el espacio de baja dimensionalidad, se define una función de similitud utilizando una curva de probabilidad suavizada:

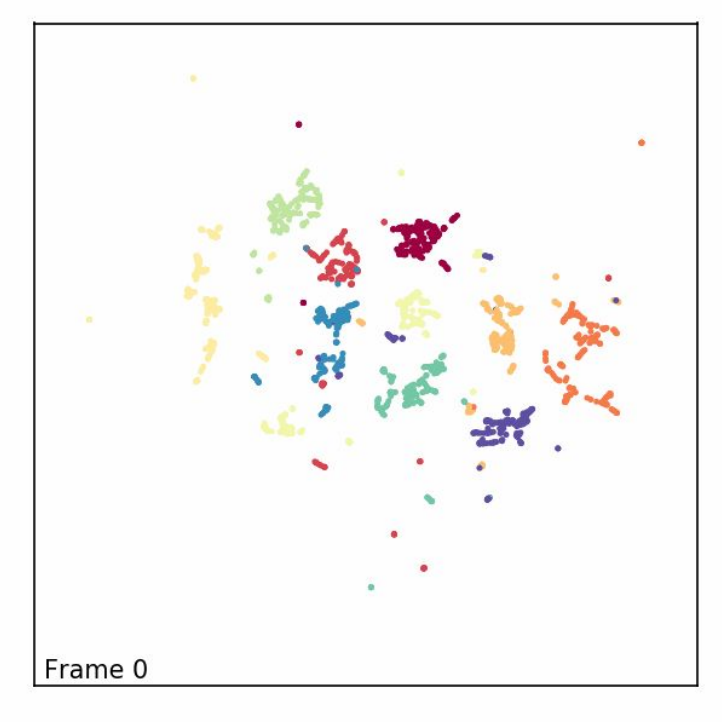
$$q_{ij} = (1 + a\|y_i - y_j\|^{2b})^{-1}$$

donde a y b son parámetros ajustados para controlar la forma de la curva.

- **Paso 4**: El objetivo es minimizar la diferencia entre las distribuciones de similitud en los dos espacios usando una función de entropía cruzada:

$$\mathcal{L} = \sum_{(i,j)} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}}$$

UMAP



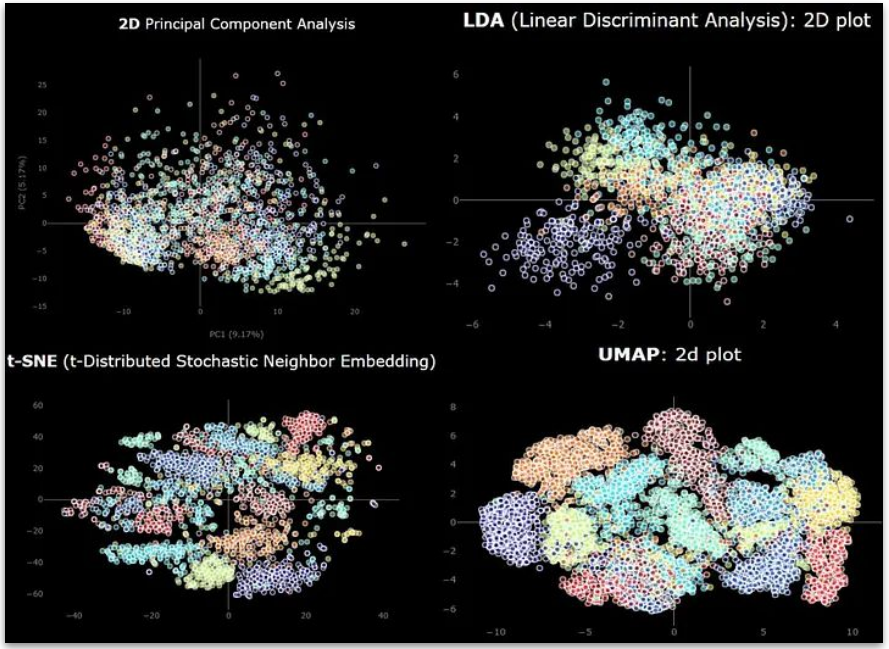
https://umap-learn.readthedocs.io/en/latest/aligned_umap_basic_usage.html

Resumen de las técnicas

- Todos los métodos buscan encontrar una transformación T que proyecte los datos a un espacio de menor dimensionalidad.
- PCA y LDA son métodos lineales, mientras que t-SNE y UMAP son métodos no lineales.
- LDA y UMAP requieren supervisión, ya que utilizan etiquetas de clase, mientras que PCA y t-SNE son métodos no supervisados.

Comparativa

The sample of KMNIST

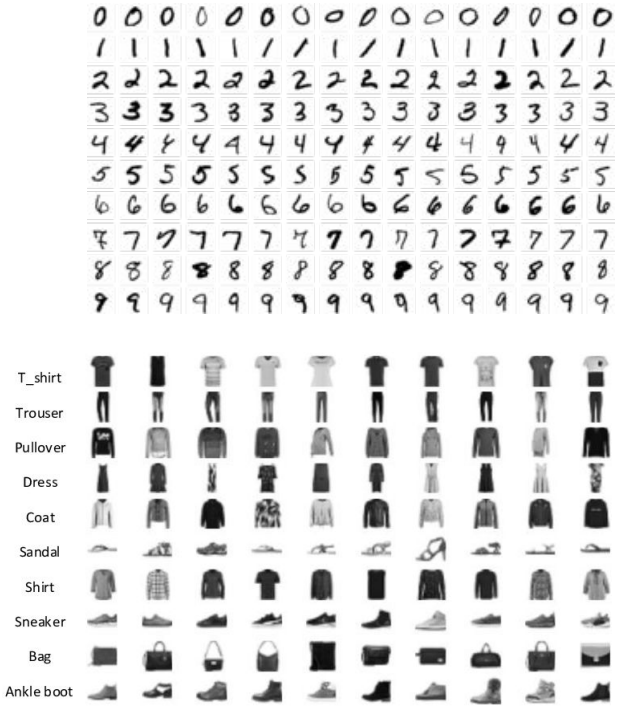


https://medium.com/@net_satsawat/dimension-reduction-tutorial-on-kuzushiji-mnist-93419101474e

Caso de estudio

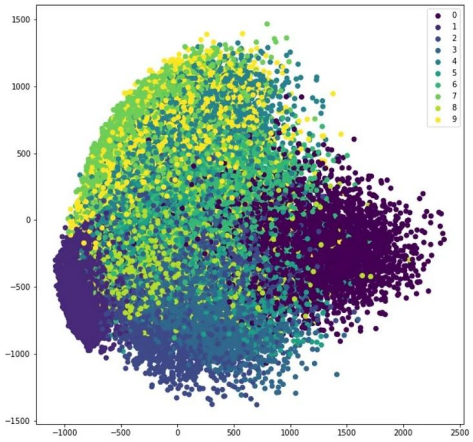
MNIST

El conjunto original de datos de imágenes **MNIST** de dígitos manuscritos es una referencia popular para los métodos de aprendizaje automático basados en imágenes.
Posteriormente se desarrolló **Fashion-MNIST**.

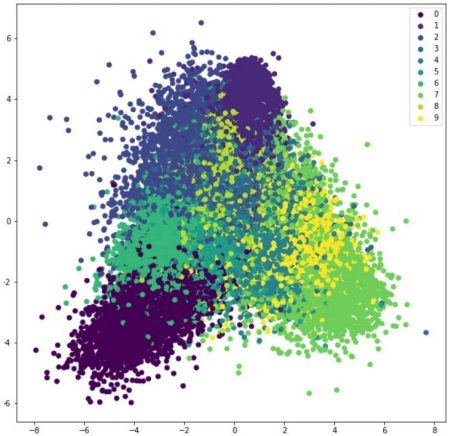


https://www.tensorflow.org/datasets/catalog/mnist
https://www.kaggle.com/datasets/zalando-research/fashionmnist

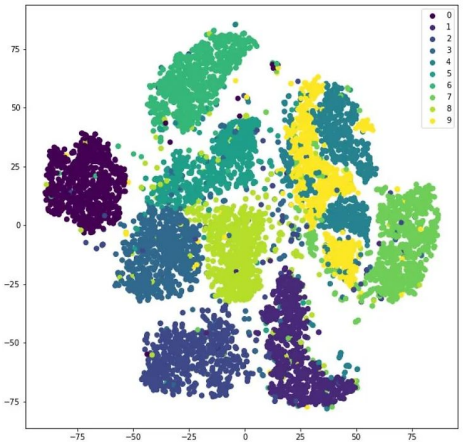
Comparativa para MNIST



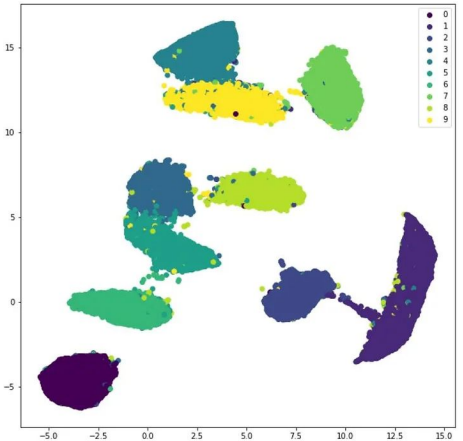
PCA



LDA



t-SNE



UMAP

https://www.hackersrealm.net/post/dimensionality-reduction-machine-learning-python

Sign Language MNIST

La base de datos de letras de la Lengua de Señas Americana (ASL) de gestos con las manos representa un problema multiclase con **24 clases** de letras (excluidas la J y la Z, que requieren movimiento).

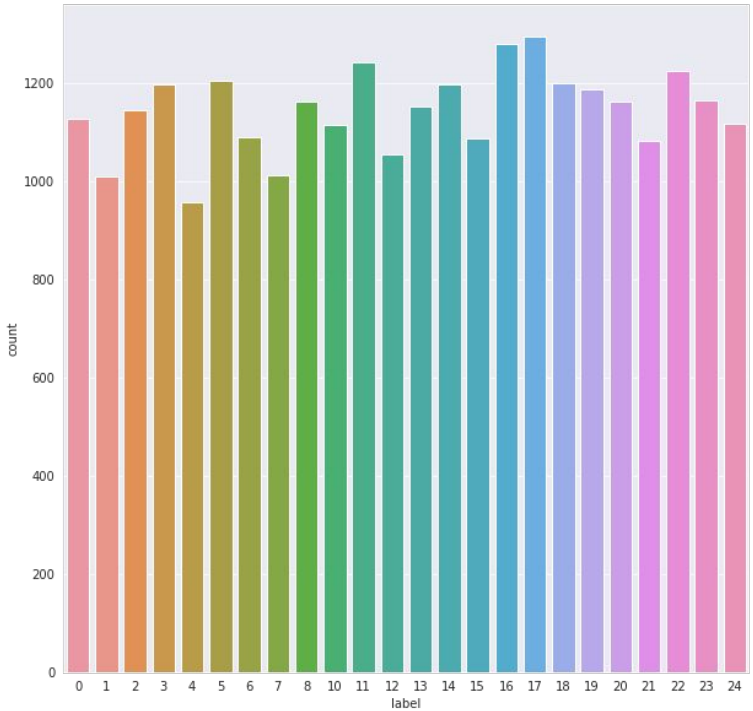
https://www.kaggle.com/datasets/datamunge/sign-language-mnist

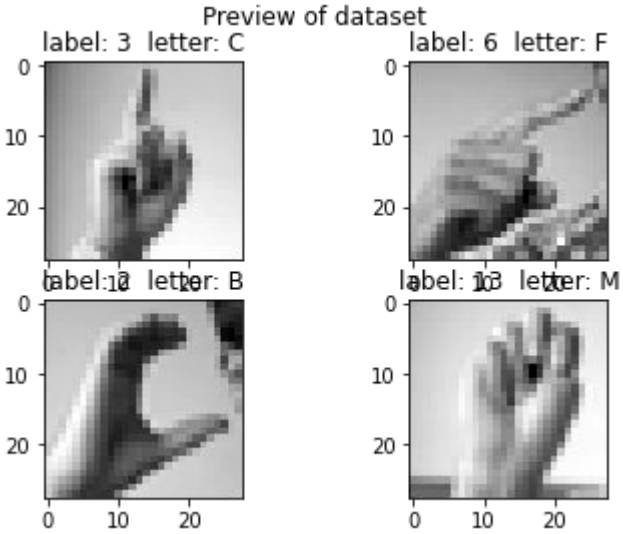
Algunos detalles

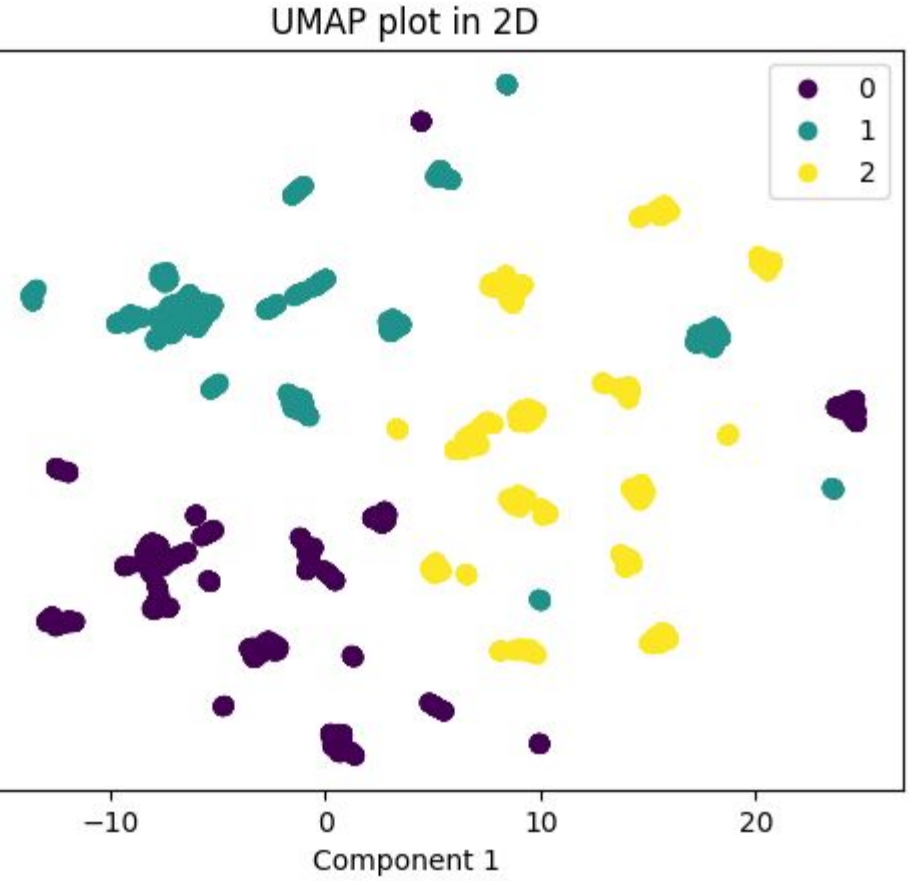
- Cada instancia de entrenamiento y prueba representa una etiqueta (0-25) para cada letra A-Z (no hay casos para 9=J o 25=Z).
- Los datos de entrenamiento (27,455 casos) y de prueba (7,172 casos) tienen una fila de encabezado, $\{pixel_1, pixel_2, ..., pixel_{784}\}$ que representa una única imagen de 28x28 píxeles con valores de escala de grises entre 0-255.

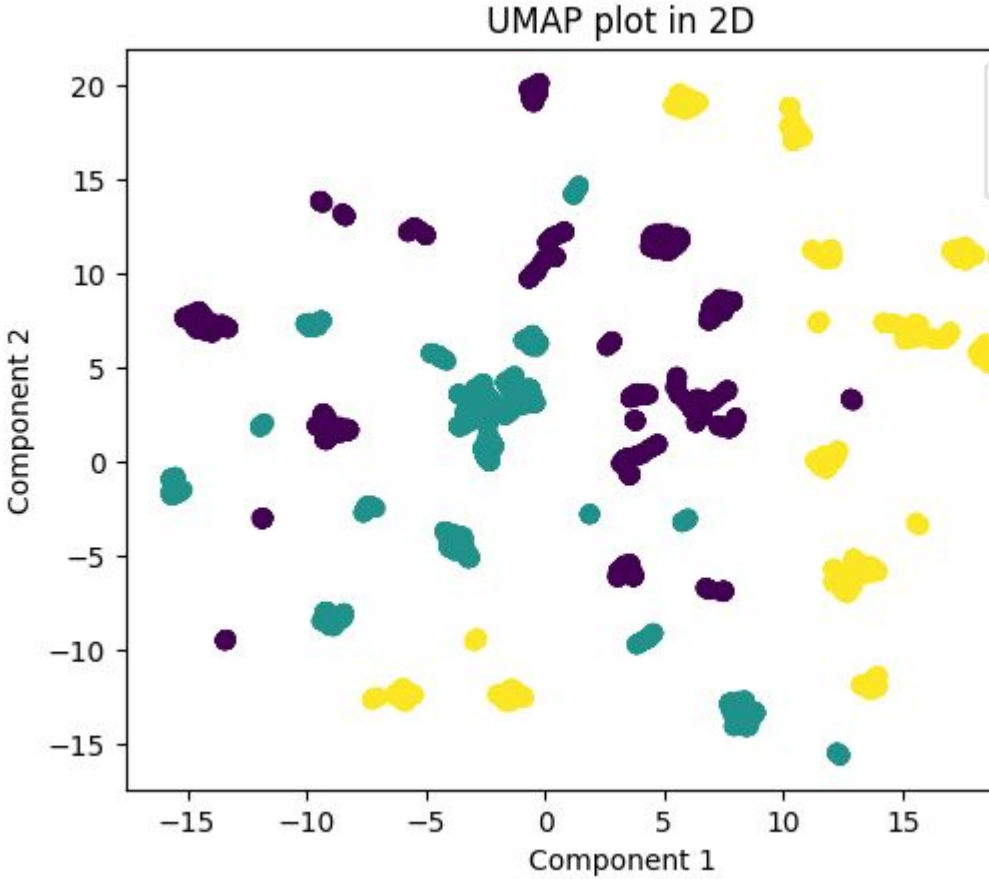
Nota: Se realizó aumento de datos y las nuevas instancias se agregaron a ambos conjuntos de datos

Análisis exploratorio

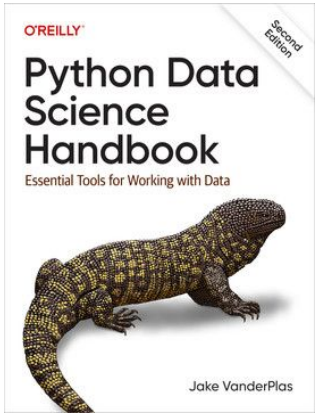





UMAP plot in 2D
0
1
2
−10
0
10
20
Component 1


UMAP plot in 2D
20
15
10
5
0
−5
−10
−15
Component 2
−15
−10
−5
0
5
10
15
Component 1

Extra Libro



GitHub

- 05.09-Principal-Component-Analysis.ipynb

Gracias!

¿Alguna pregunta?

hussein@cicese.mx

https://sites.google.com/view/husseinlopeznava

SCAN ME