Reducción de Dimensionalidad: Desbloquear Insights de Datos Complejos

Una técnica fundamental para transformar datos complejos en información comprensible y procesable

Teo Ramos, María Dunaeva

Made with GAMMA

¿Por Qué Importa la Reducción de Dimensionalidad?

Maldición de la Dimensionalidad

A medida que aumentan las dimensiones, los datos se vuelven dispersos. Dos consecuencias directas:

- Las distancias pierden significado (todos los puntos están "lejos" entre sí)
- Los algoritmos de ML necesitan exponencialmente más datos para funcionar bien

Overfitting (sobreajuste)

Demasiadas características pueden llevar a modelos que memorizan en lugar de generalizar

Velocidad de Cómputo

Reducir dimensiones acelera significativamente el procesamiento y análisis

Reducir las dimensiones revela patrones ocultos y hace que los datos complejos sean más manejables y comprensibles

Made with GAMMA

¿Qué es la Reducción de Dimensionalidad?

Es el proceso de transformar datos de alta dimensión a un espacio de menor dimensión, preservando la estructura y varianza más significativas. Es como tomar una fotografía (2D) de un objeto tridimensional: perdemos información, pero mantenemos lo esencial.

Beneficios prácticos:

- Visualización de datos complejos en 2D o 3D
- Reducción de ruido y overfitting
- Aceleración de algoritmos (menos features = menos cómputo)
- Identificación de patrones ocultos

Made with GAMMA

Procedimiento

01

Análisis del Espacio Original

Identificar las características más relevantes en los datos de alta dimensión

02

Transformación Matemática

Aplicar técnicas para proyectar los datos a un espacio de menor dimensión

03

Preservación de Información

Mantener la estructura y patrones más importantes de los datos originales

Made with GAMMA

Dos Enfoques Principales

Selección de Características

Elegir las características originales más relevantes y descartar las redundantes o irrelevantes

- Filtrado por correlación
- Selección univariante
- Métodos wrapper

Extracción de Características

Crear nuevas características que resuman y capturen la información esencial de los datos originales

- Combinaciones lineales
- Proyecciones no lineales
- Embeddings aprendidos

Ambos enfoques buscan reducir la redundancia y mejorar la interpretabilidad de los datos complejos.

Made with GAMMA

Análisis de Componentes Principales (PCA)

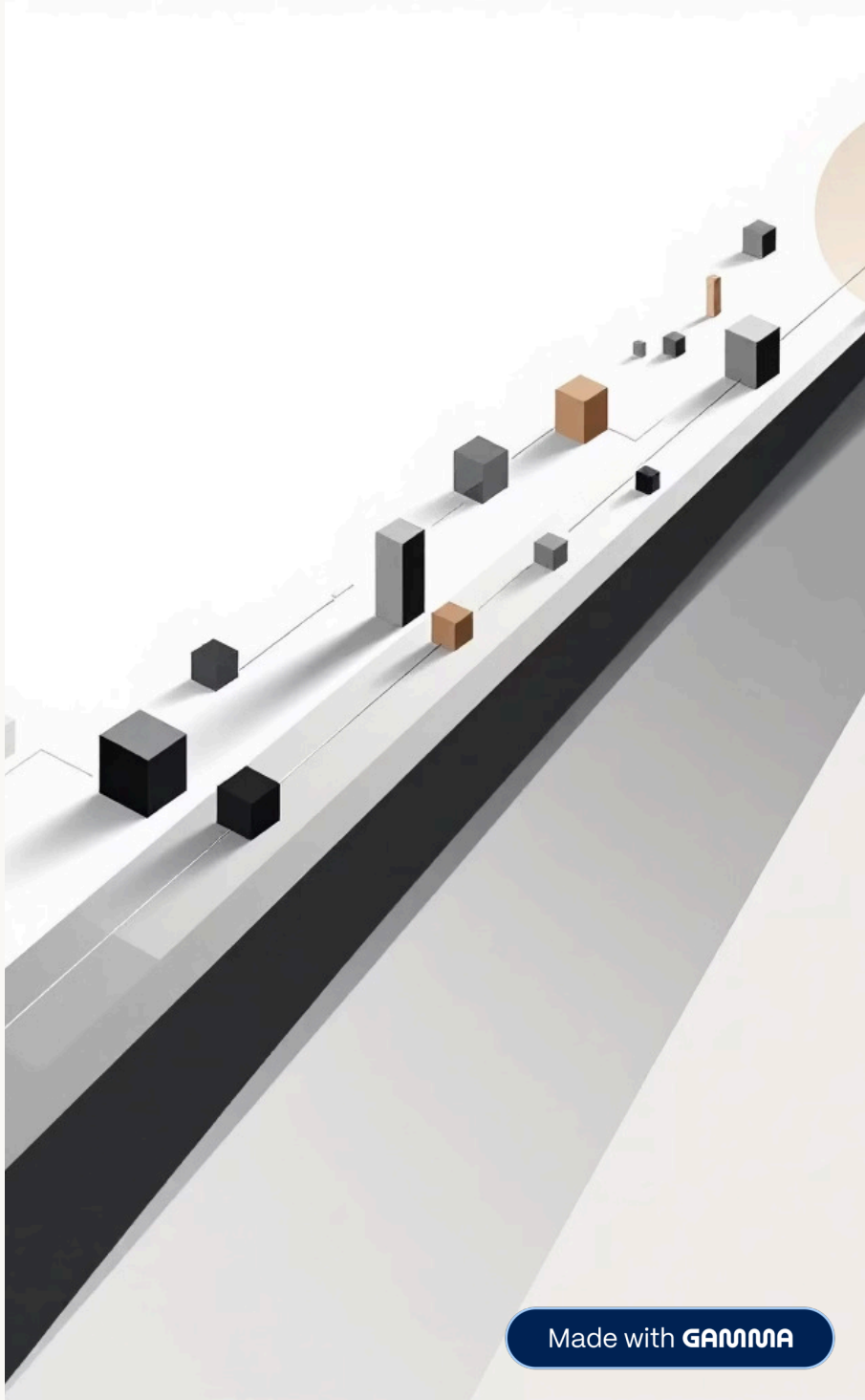
Método lineal que encuentra las direcciones (ejes ortogonales) que capturan la máxima varianza en los datos.

Características:

- Método lineal (combinaciones lineales de features originales)
- Los componentes son ortogonales entre sí
- Ordena los componentes por varianza explicada
- Interpretable: cada componente es una combinación de variables originales

Cuándo usarlo:

- Datos con correlaciones lineales fuertes
- Necesitas interpretabilidad
- Quieres eliminar multicolinealidad
- Visualización exploratoria rápida



Made with GAMMA

PCA

Velocidad

Computacionalmente eficiente para grandes conjuntos de datos

Interpretabilidad

Los componentes principales tienen interpretación matemática clara

Estructura Global

Preserva relaciones lineales y estructura global de los datos

Limitación:

- Asume relaciones lineales
- Sensible a escala (requiere normalización)
- No preserva distancias locales complejas

Made with GAMMA

Técnicas No Lineales: t-SNE y UMAP

t-SNE

Preserva vecindarios locales, excelente para visualizar clusters y estructuras complejas

UMAP

Más rápido que t-SNE, preserva tanto estructura local como global de manera equilibrada

Ambas técnicas sobresalen en la visualización de manifolds de datos complejos y no lineales

Made with GAMMA

t-SNE (t-Distributed Stochastic Neighbor Embedding)

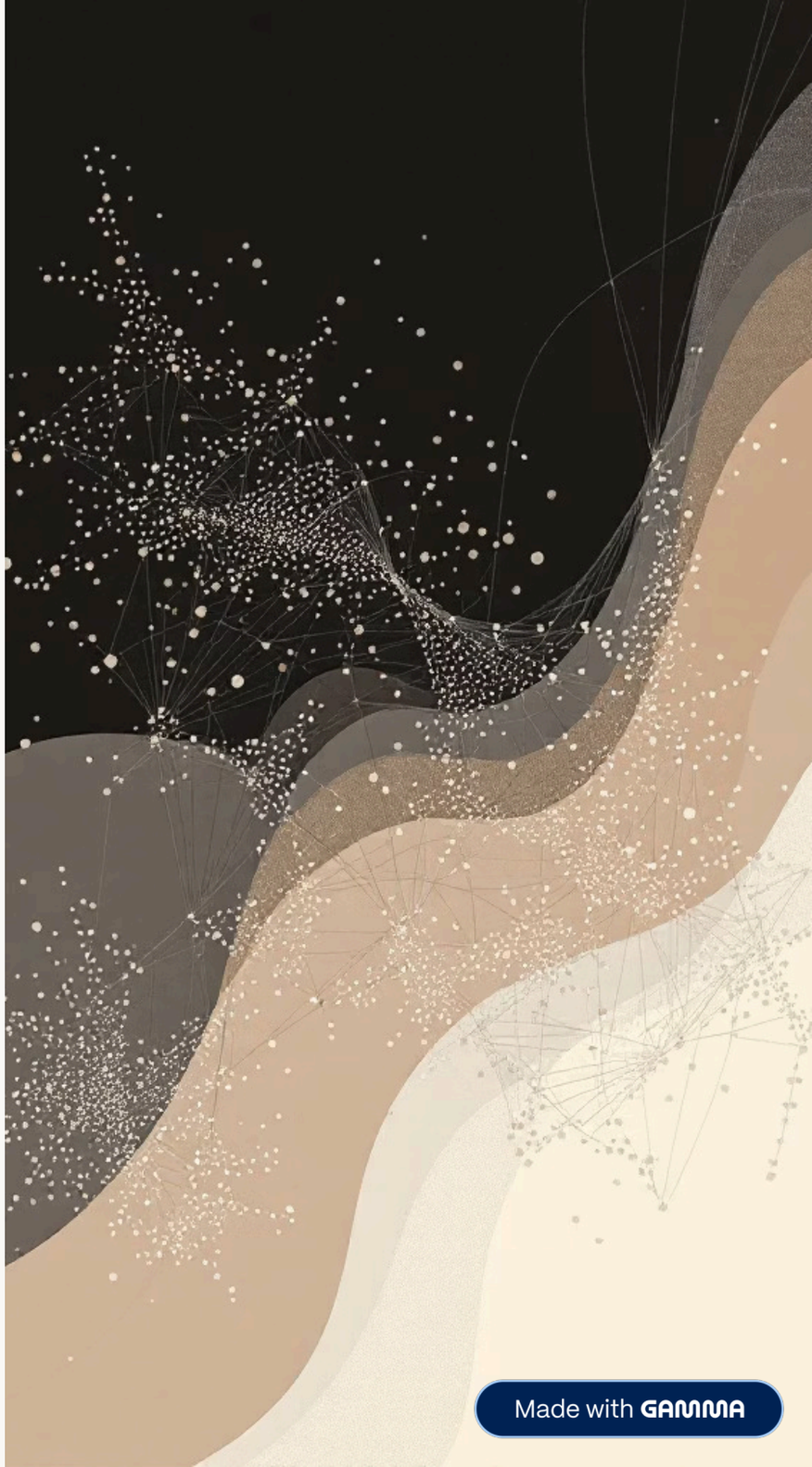
Preserva las relaciones de vecindad local. Puntos cercanos en el espacio original permanecen cercanos en el espacio reducido. Usa probabilidades para modelar similitudes.

Características:

- Método no lineal
- Excelente para visualización (especialmente 2D/3D)
- Preserva estructura local, no global
- No es determinista (diferentes ejecuciones dan resultados distintos)

Cuándo usarlo:

- Visualización de clusters y patrones
- Datos con estructura no lineal compleja
- Exploración inicial de datasets complejos (imágenes, texto embeddings)
- Cuando las distancias globales no importan tanto



Made with GAMMA

t-SNE

Exploración Visual

Ideal para visualizar datos de alta dimensión en 2D o 3D, revelando patrones ocultos

Agrupamiento Local

Preserva relaciones locales entre puntos para identificar clústeres y similitudes con gran precisión

Separación No Lineal

Captura estructuras no lineales, útil para datos complejos como imágenes o texto

Intuición de Clases

Facilita la interpretación visual de clases o etiquetas en tareas de clasificación, incluso cuando no están explícitas

Limitación:

- No sirve para nuevos datos (no es un modelo que puedas aplicar a test set)
- Computacionalmente costoso con muchos datos
- Los hiperparámetros (perplexity) impactan mucho el resultado
- Las distancias en el output no son interpretables

Made with GAMMA

UMAP (Uniform Manifold Approximation and Projection)

Similar a t-SNE pero con fundamentos matemáticos diferentes (topología). Preserva mejor la estructura global mientras mantiene la local.

Características:

- Método no lineal más moderno
- Más rápido que t-SNE
- Preserva estructura local Y global razonablemente
- **Puede transformar datos nuevos** (a diferencia de t-SNE)
- Mejor conservación de distancias

Cuándo usarlo:

- Necesitas velocidad con grandes volúmenes
- Quieres mantener estructura global
- Necesitas aplicar la transformación a datos nuevos
- Proyectos en producción

Made with GAMMA

UMAP

Velocidad

Rápido en entrenamiento y proyección, incluso con grandes volúmenes de datos

Estructura Local y Global

Preserva tanto la estructura local (vecindades) como la estructura global del conjunto de datos

Versatilidad

Funciona bien como técnica de visualización, pero también como preprocesamiento para modelos de machine learning.

Parametrizable

Ofrece control sobre el equilibrio entre estructura local y global mediante parámetros como `n_neighbors` y `min_dist`.

Limitación:

- Menos años de investigación que PCA/t-SNE
- Hiperparámetros también requieren tuning (n_neighbors, min_dist)
- Menor adoptación (aunque creciendo rápidamente)

Made with GAMMA

Comparando PCA, t-SNE y UMAP

PCA

Mejor para: Varianza lineal global, análisis de componentes

Limitado en: Visualización de clusters complejos

t-SNE

Excelente para: Separación local de clusters

Puede distorsionar: Distancias globales entre grupos

UMAP

Equilibra: Estructura local y global

Ventajas: Escalable y versátil para diversos tipos de datos

Made with GAMMA

Buenas prácticas

Errores comunes

❌ Aplicar reducción antes de train-test split (data leakage)
❌ No escalar los datos con PCA
❌ Sobre-interpretar distancias en t-SNE
❌ Usar t-SNE para datos nuevos
❌ No experimentar con hiperparámetros

✅ Fit solo en training, transform en test
✅ Normalizar primero
✅ Entender las limitaciones de cada técnica
✅ Probar múltiples configuraciones

Preprocesamiento

1. **Escala tus datos**: PCA especialmente lo requiere. StandardScaler o MinMaxScaler.
2. **Maneja valores faltantes**: Imputación antes de reducir dimensionalidad.
3. **Elimina outliers extremos**: Pueden dominar los componentes principales.

Evaluación

- **PCA**: Gráfico de varianza explicada acumulada (elegir componentes que expliquen 80-95%)
- **t-SNE/UMAP**: Evaluación visual principalmente. ¿Se separan los clusters conocidos?
- **Validación**: Usa los datos reducidos en tu tarea objetivo (clasificación, clustering) y mide performance

Made with GAMMA

Más técnicas

Autoencoders (Deep Learning):

- Redes neuronales que comprimen y reconstruyen datos
- El cuello de botella actúa como representación reducida
- Muy potentes para datos complejos (imágenes, audio)
- Requieren más datos y cómputo

Linear Discriminant Analysis (LDA):

- Similar a PCA pero supervisado (usa las etiquetas)
- Maximiza separación entre clases
- Útil para clasificación
- Limitado a C-1 dimensiones (C = número de clases)

Factor Analysis:

- Busca variables latentes que explican correlaciones
- Más enfocado en interpretabilidad que PCA
- Común en ciencias sociales y psicometría

Made with GAMMA

Aplicaciones del Mundo Real

Análisis de Imágenes

Compresión y visualización de embeddings de modelos como ResNet y CLIP para reconocimiento de patrones visuales

Datos de Texto

Reducción de dimensiones en embeddings de palabras y documentos para clustering y búsqueda semántica

Bioinformática

Visualización de datos de expresión génica y reducción de ruido en análisis genómicos

Estas aplicaciones demuestran la versatilidad y el poder de la reducción de dimensionalidad en diversos dominios científicos y tecnológicos.

Made with GAMMA

Aprovechando la Reducción de Dimensionalidad

Herramienta Esencial

Indispensable para dar sentido a datos complejos de alta dimensión

Elección Estratégica

Seleccionar la técnica basándose en la estructura de datos y objetivos del análisis

Resultados Transformadores

Desbloquear insights ocultos, mejorar modelos y crear visualizaciones impactantes

La reducción de dimensionalidad es el puente entre la complejidad de los datos modernos y la comprensión humana.

Made with GAMMA

Visualizando la Reducción de Dimensionalidad con CIFAR-10

Proyecciones PCA

Los embeddings de imágenes reducidos a 2D revelan clusters claros de clases

Mapas t-SNE

Muestran direcciones de varianza pero menor separación de clusters

UMAP

Destacan agrupaciones locales pero pueden distorsionar la forma general

Cada técnica ofrece una perspectiva única de los mismos datos, revelando diferentes aspectos de su estructura subyacente.

Made with GAMMA