# **UMAP**

UMAP (Aproximación y Proyección Uniforme de Manifolds) es una técnica de reducción de dimensionalidad que tiene como objetivo preservar tanto la estructura local como global en datos de alta dimensionalidad. Está basada en un marco matemático con:

* **Suposición de manifold:** UMAP asume que los datos de alta dimensionalidad yacen en un manifold de dimensionalidad inferior incrustado en el espacio original. El objetivo es encontrar una representación de baja dimensionalidad que preserve la estructura del manifold.
* **Conjunto simplicial difuso:** UMAP construye una representación topológica difusa de los datos creando un grafo ponderado de vecinos más cercanos. Este grafo captura la estructura local y global, y se utiliza para aproximar el manifold subyacente.
* **Geometría riemanniana:** UMAP optimiza una función objetivo que mide la similitud entre el conjunto simplicial difuso en el espacio de alta dimensionalidad vs el espacio de baja dimensionalidad. Esta optimización está guiada por principios de geometría riemanniana, para encontrar una incrustación que minimice la distorsión.
* **Optimización estocástica:** UMAP emplea un algoritmo de optimización estocástica, como el descenso de gradiente estocástico (SGD). Este proceso ajusta iterativamente la incrustación para minimizar la discrepancia entre las representaciones de alta y baja dimensionalidad.

Es útil para:

* **Visualización**: UMAP puede utilizarse para visualizar datos de alta dimensionalidad en dos o tres dimensiones.
* **Agrupamiento**: Las incrustaciones de UMAP pueden utilizarse como características de entrada para algoritmos de agrupamiento, ayudando a identificar grupos en datos de alta dimensionalidad.
* **Extracción de características:** UMAP puede utilizarse como técnica de extracción de características para reducir la dimensionalidad de los datos antes de alimentarlos en modelos de aprendizaje automático (mejora el rendimiento del modelo y reduce la complejidad computacional).
* **Detección de anomalías:** Las incrustaciones de UMAP pueden utilizarse para identificar valores atípicos o anomalías en datos de alta dimensionalidad al medir la distancia entre los puntos de datos en el espacio de baja dimensionalidad.

# LDA

Latent Dirichlet Allocation es un modelo probabilístico utilizado para la modelización de temas, que es una técnica para descubrir temas abstractos dentro de una colección de documentos. Los principios matemáticos involucran estadísticas bayesianas y distribuciones de probabilidad.

* **Distribución de Dirichlet:** LDA asume que los documentos están representados como una mezcla de temas, y cada tema es una distribución de probabilidad sobre palabras. Esta distribución se utiliza para modelar la distribución de temas en un documento y la distribución de palabras en un tema. La distribución de Dirichlet se caracteriza por un conjunto de parámetros (alfa para la distribución de documentos-tema y beta para la distribución de tema-palabra).
* **Proceso generativo:** LDA asume un proceso generativo para crear documentos:
  + Para cada documento, elige una distribución sobre temas de la distribución de Dirichlet (distribución de documentos-tema).
  + Para cada palabra en el documento:
    - Elige un tema de la distribución del documento sobre temas.
    - Elige una palabra de la distribución de palabras del tema elegido (distribución de tema-palabra).
* **Inferencia**: Dada una colección de documentos, el objetivo de LDA es inferir la estructura de temas subyacentes. Esto implica estimar los parámetros de la distribución de Dirichlet que mejor expliquen los documentos observados. Esto se hace típicamente utilizando inferencia variacional o muestreo de Gibbs.

LDA es útil para varias tareas de procesamiento del lenguaje natural, incluyendo:

* Modelización de temas: Identificar los temas principales presentes en una colección de documentos.
* Agrupación de documentos: Agrupar documentos similares basados en sus distribuciones de temas.
* Clasificación de texto: Asignar temas a documentos nuevos o no vistos.
* Extracción de características: Representar documentos en un espacio de menor dimensión basado en sus distribuciones de temas para tareas posteriores como clasificación o recomendación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algoritmo | UMAP | LDA | PCA | T-SNE |
| Nombre | Uniform Manifold Approximation and Projection | Latent Dirichlet Allocation (LDA) | Principal Component Analysis |  |
| Tipo |  | Probabilístico | Lineal | Probabilístico |
| Principios matemáticos | Suposición de manifold  Conjunto simplicial difuso  Geometría riemanniana  Optimización estocástica  Topología algebraica  Teoría de grafos | Estadística bayesiana | Algebra linea  Eigenvectors/values |  |
| Aplicaciones | - Visualización  - Agrupamiento  - Extracción de características  - Detección de anomalías | - Encontrar temas en una colección de documentos  - Clasificación de texto  - Lenguaje natural  - Extracción de características | - Extracción de características | - Extracción de características  - Imágenes |
| Implementación | <https://github.com/lmcinnes/umap> |  |  |  |
| Enlaces |  |  |  |  |

<https://plotly.com/python/t-sne-and-umap-projections/>