Apuntes Semana 14-29/05/2025

Elaborado por:
Jeremy Chacón Beckford-2021131338

***Abstract*—El presente documento recopila los apuntes de la semana catorce sobre aprendizaje no supervisado. Se abordaron conceptos fundamentales relacionados con clustering, reducción de dimensionalidad mediante PCA, y los fundamentos matemáticos de vectores propios y valores propios. Se discuten diferentes algoritmos, criterios de convergencia y análisis vectorial aplicado al preprocesamiento de datos.**

Noticias sobre Inteligencia Artificial

A. Actualización del modelo DeepSeek

Se anunció el lanzamiento de **DeepSeek-V3.1**, descrito como el modelo más avanzado de la empresa hasta la fecha. Entre sus principales características se destacan:

- Presenta un rendimiento comparable al de modelos privados de alto nivel.
- Forma parte del ecosistema *open-source*, lo cual facilita el acceso a desarrollos avanzados para empresas sin grandes equipos de investigación.
- Promueve el desarrollo colaborativo y acelerado de modelos de inteligencia artificial accesibles.

B. Paper: *Harnessing the Universal Geometry of Embeddings*

Se menciona un artículo reciente titulado *Harnessing the Universal Geometry of Embeddings*, Este trabajo destaca:

- Los espacios latentes generados por modelos de lenguaje tienden a presentar geometrías similares entre sí.
- Se propone una técnica para traducir representaciones vectoriales entre distintos espacios latentes.
- Se introduce el concepto de una **representación universal** útil para tareas de alineamiento y comparación entre modelos preentrenados.

Aprendizaje no supervisado

En el aprendizaje no supervisado no se utilizan etiquetas. El modelo busca identificar patrones, estructuras o agrupaciones en los datos únicamente a partir de las variables de entrada.

- Entrada: conjunto de vectores de características.
- Salida: agrupaciones, relaciones internas, reducción de dimensionalidad.

Clustering

Definición

El **clustering** es un procedimiento para identificar grupos en los datos, basado en la similitud de los *features*. Se divide el espacio del conjunto de datos en N grupos que sean lo más homogéneos posibles internamente, y diferentes entre sí. *No hay una técnica perfecta de clustering, solo distintas aproximaciones y algoritmos que se ajustan a diferentes tipos de datos.*

Múltiples definiciones de clustering

- Proceso de encontrar grupos en los datos.
- Dividir los datos en grupos homogéneos.
- Dividir los datos en grupos donde los *features* de cada grupo estén lo más cerca posible (similares).
- Dividir los datos en grupos donde los *features* de cada grupo estén lo más cerca posible, y los de diferentes grupos estén lo más lejos posible.

Algoritmos discutidos

K-Means

- Requiere definir un valor inicial de K (número de clusters).
- Todas las muestras se asignan a un único clúster, con base en el centroide más cercano.
- Minimiza una función de costo basada en el error cuadrático entre cada muestra del dataset y el elemento representativo del clúster (centroide).
- Proceso iterativo que continúa hasta alcanzar un mínimo local o cumplir un criterio de parada.

Criterio de convergencia

- Cuando las muestras no cambian de clúster tras una iteración.
- Cuando los centroides no cambian.
- Cuando el valor de la función objetivo no cambia (las distancias se estabilizan).
- Cuando se alcanza un número máximo de iteraciones.

Observaciones:

- Sensible a la inicialización de centroides.
- Sensible a *outliers*.
- Puede generar resultados distintos en ejecuciones diferentes.

Tipos de pertenencia a clústeres

- **Clustering determinístico**: la muestra pertenece completamente a un único clúster. Ejemplos: *K-Means*, *DBSCAN*.
- **Clustering probabilístico**: la muestra tiene una probabilidad de pertenecer a cada clúster. Ejemplo: *Gaussian Mixture Models*.

Medidas de distancia

- Euclidiana
- Manhattan
- Coseno
- Mahalanobis

Otros algoritmos

- **DBSCAN**: agrupa por densidad, identifica *outliers*.
- **Birch**: eficiente para grandes volúmenes de datos.
- **Gaussian Mixture Models**: asigna pertenencia probabilística.
- **Clustering jerárquico**: método sistemático para crear grupos basados en métricas de similitud. Consiste en asignar cada muestra a un clúster distinto, agrupar iterativamente los pares de clústeres más similares y actualizar las similitudes hasta obtener un único clúster. Este proceso genera un dendograma que representa las fusiones sucesivas.
  Existen dos variantes principales:
  - **Aglomerativo (Bottom-Up)**: cada punto inicia como su propio clúster; los más similares se van uniendo hasta formar uno solo.
  - **Divisivo (Top-Down)**: se parte de un único clúster que contiene todos los puntos y se va dividiendo progresivamente.

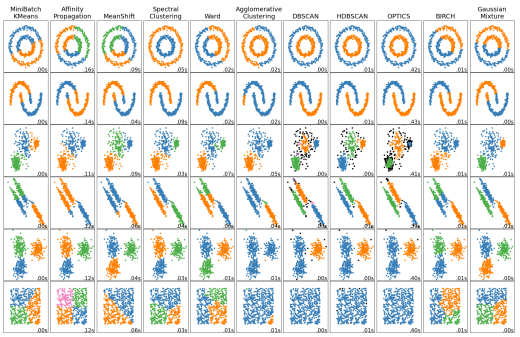

FIGURE 1. Comparación visual de diferentes algoritmos de clustering aplicados sobre conjuntos de datos con formas diversas. Cada columna representa un algoritmo distinto y cada fila un tipo de distribución de datos.

Reducción de dimensionalidad: PCA

Objetivo

Transformar datos de alta dimensión a un espacio con menor número de dimensiones, preservando la mayor cantidad posible de varianza.

Proceso

1) Estandarizar los datos (media cero, varianza uno).
2) Calcular la matriz de covarianza:

$$\mathrm{Cov}(\mathbf{X}) = \frac{1}{n-1}\left((\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{X}-\bar{\mathbf{X}})\right), \quad \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

   Esta matriz mide la relación lineal entre los atributos (features) y debe usarse solo si los datos han sido estandarizados previamente.
3) Obtener los vectores propios (*eigenvectors*) y valores propios (*eigenvalues*).
4) Seleccionar los k vectores con mayores *eigenvalues*.

5) Proyectar los datos al nuevo subespacio. Por ejemplo, si los datos están más dispersos en el eje X que en el eje Y, se prioriza el eje X como componente principal por contener mayor varianza.

Matriz de correlación

La matriz de correlación mide la relación lineal entre los atributos (features), al igual que la matriz de covarianza. Sin embargo, sus valores están acotados entre -1 y 1, lo que facilita su interpretación:

- $+1$: Altamente correlacionados positivamente.
- -1: Altamente correlacionados negativamente.
- 0: No existe correlación lineal.

Cuando dos variables están correlacionadas, tienden a explicar la misma característica del conjunto de datos. En cambio, si no están correlacionadas, se consideran ortogonales entre sí (no redundantes).

Características

- Los componentes principales son ortogonales entre sí.
- Representan combinaciones lineales de los atributos originales.
- Los *eigenvalues* reflejan cuánta varianza captura cada componente.
- Se puede perder información dependiendo del número de componentes seleccionados.

Definición formal

Un eigenvector $\vec{v}$ y su eigenvalor λ cumplen que:

$$A\vec{v} = \lambda\vec{v}$$

donde A es la matriz de covarianza o correlación, y $\vec{v}$ no es el vector nulo.

Otras técnicas de reducción y detección de anomalías

Además de PCA, se mencionó el uso de **Autoencoders** como técnica alternativa para la reducción de dimensionalidad. Esta técnica, basada en redes neuronales, permite obtener una representación comprimida de los datos conservando sus características esenciales.

En cuanto a la detección de *outliers*, se abordaron dos enfoques:

- **Evento binario**: Detección basada en clasificar directamente si un dato representa una anomalía o no.
- **DBSCAN**: Algoritmo que identifica automáticamente puntos que no pertenecen a ningún clúster como anomalías, gracias a su naturaleza basada en densidad.

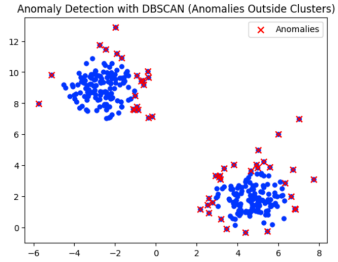

FIGURE 2. Detección de anomalías con DBSCAN. Los puntos rojos representan datos clasificados como outliers fuera de los clústeres principales.

Visualización de datos con t-SNE

Además de PCA, se mencionó la técnica **t-SNE** (*t-Distributed Stochastic Neighbor Embedding*) como una herramienta útil para la visualización de datos de alta dimensión en dos o tres dimensiones. Esta técnica no lineal permite preservar relaciones de proximidad entre puntos, siendo especialmente efectiva para representar estructuras de clústeres complejos.

Se utiliza comúnmente en tareas de exploración de datos y análisis visual, como lo muestra el siguiente ejemplo, donde se representan vectores embebidos del conjunto MNIST:

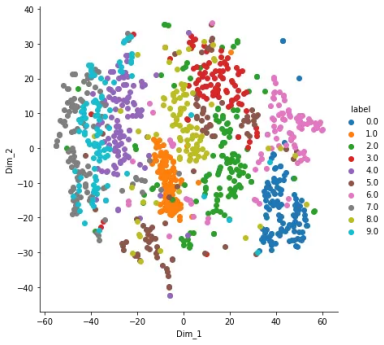

FIGURE 3. Visualización de representaciones embebidas utilizando t-SNE. Cada color representa una clase diferente del dataset.

Conceptos clave

- **Eigenvector**: vector que no cambia de dirección tras una transformación lineal.

- **Eigenvalue**: escalar que representa la magnitud de cambio del eigenvector.
- **Span**: conjunto de combinaciones lineales que conservan la dirección original.
- **PCA**: técnica que extrae componentes principales ortogonales del conjunto de datos.
- **Covarianza**: mide la relación lineal entre dos variables.
- **Clustering**: agrupamiento de datos según métrica de similitud.

Observaciones adicionales

- PCA permite representar un conjunto de datos en un espacio de menor dimensión con pérdida mínima de información.
- La selección del número de componentes puede basarse en el porcentaje de varianza explicada acumulada.
- En clustering, la elección del número de clusters (K) puede apoyarse en heurísticas como el método del codo.
- En espacios de alta dimensión, se recomienda aplicar PCA antes de realizar agrupamientos.

Anuncio importante

Las próximas clases se impartirán de manera virtual.