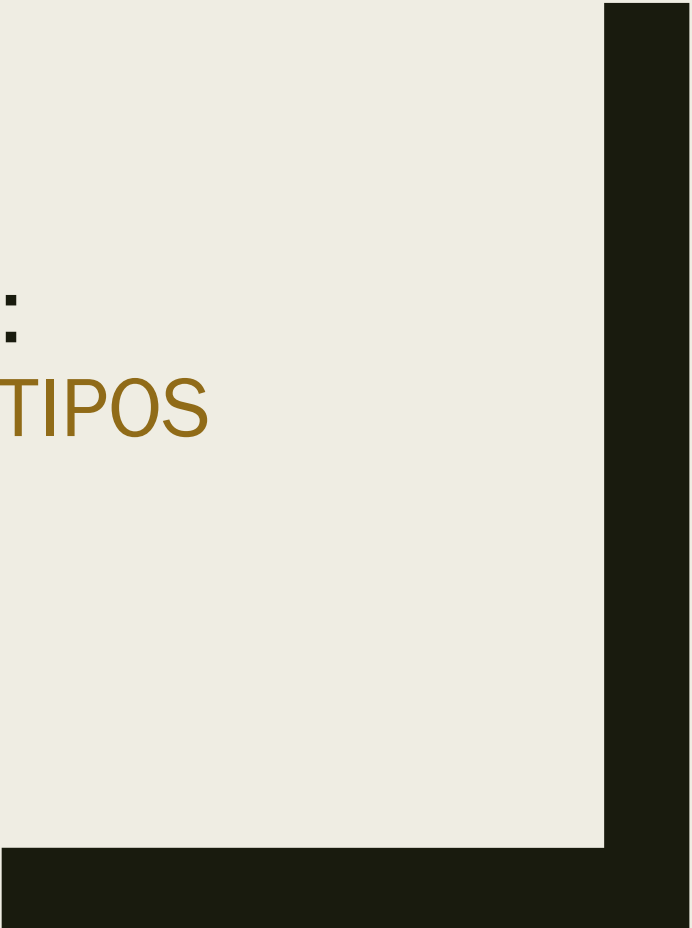




ANALÍTICA AVANZADA DE DATOS: APRENDIZAJE DE CUANTIFICACIÓN VECTORIAL

A. Alejandra Sánchez Manilla
asanchezm.q@gmail.com



Aprendizaje de cuantificación vectorial



El aprendizaje de cuantificación vectorial es una técnica utilizada en el procesamiento de datos y la inteligencia artificial para representar y comprimir información de manera eficiente



Su **objetivo principal** es reducir la dimensionalidad de un conjunto de vectores, manteniendo la estructura y las características más relevantes de los datos originales

Aprendizaje de cuantificación vectorial

El proceso de cuantificación vectorial implica mapear los vectores de alta dimensión a un espacio de menor dimensión

Se utilizan técnicas y algoritmos que buscan encontrar representantes o "**centroides**" que capturen la esencia de los datos originales. Estos centroides son *vectores que representan grupos o clústeres de datos similares*

La asignación de los vectores originales a los centroides se basa en criterios de similitud, como la distancia euclidiana o la similitud coseno.

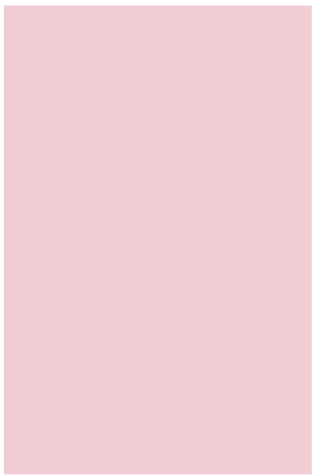
Aprendizaje de cuantificación vectorial



Cuantificación vectorial clásica (CVC):

Utiliza técnicas matemáticas para agrupar los vectores originales en un conjunto discreto de representantes llamados **centroides**

La asignación de vectores a centroides se basa en criterios como la distancia euclidiana o la similitud coseno



Aprendizaje no supervisado:

Este enfoque busca **identificar patrones** y estructuras ocultas en los datos **sin la necesidad de etiquetas** o información previa

Algunos algoritmos: agrupamiento k-means y el análisis de componentes principales (PCA)

Aprendizaje de cuantificación vectorial



Aprendizaje supervisado:

En este caso, se utilizan conjuntos de datos etiquetados para entrenar un modelo que pueda asignar nuevos vectores a categorías predefinidas

Los algoritmos: como las máquinas de vectores de soporte (SVM) y las redes neuronales

Aprendizaje de cuantificación vectorial

Ejemplo 1:

Supongamos que tenemos una imagen en escala de grises de 512x512 píxeles

Cada píxel tiene un valor de intensidad que varía entre 0 y 255

El objetivo es comprimir la imagen reduciendo la cantidad de información necesaria para almacenarla, manteniendo una calidad visual aceptable

1. Preparación de datos:

- La imagen se divide en bloques más pequeños, por ejemplo, bloques de 8x8 píxeles
- Cada bloque se representa como un vector unidimensional de 64 elementos (8x8)
- Los valores de los píxeles se normalizan para que estén en un rango común, como $[0, 1]$

Aprendizaje de cuantificación vectorial

2. Entrenamiento del algoritmo:

- Se aplica el algoritmo de cuantización vectorial, como el algoritmo de agrupamiento k-means, al conjunto de bloques de la imagen
- Se selecciona un número deseado de centroides, por ejemplo, $k=16$
- Los centroides se inicializan aleatoriamente y se actualizan iterativamente hasta alcanzar la convergencia
- Cada bloque de la imagen se asigna al centroide más cercano en función de la distancia euclidiana

Aprendizaje de cuantificación vectorial

3. Representación comprimida:

- En lugar de almacenar los bloques de la imagen original, se almacenan únicamente los índices de los centroides a los que pertenecen
- Estos índices pueden codificarse de manera eficiente, utilizando técnicas de compresión como la codificación Huffman

4. Reconstrucción de la imagen:

- Para descomprimir la imagen, se utiliza la información almacenada de los índices de centroides
- Los centroides correspondientes se utilizan para reconstruir los bloques de la imagen
- Los bloques se concatenan para obtener la imagen completa y se realiza una operación inversa de normalización

Aprendizaje de cuantificación vectorial

El resultado de este proceso es una imagen comprimida que utiliza menos información para su representación, pero que conserva una calidad visual aceptable

La cantidad de compresión y calidad dependerá del *número de centroides* seleccionados y del algoritmo de cuantización vectorial utilizado

En este ejemplo, la compresión de imágenes muestra cómo se puede reducir la cantidad de datos necesarios para almacenar una imagen, aprovechando las similitudes entre los bloques de la imagen y representándolos mediante centroides y sus índices correspondientes

Aprendizaje de cuantificación vectorial

Ejemplo 2:

Consideremos un conjunto de datos que consiste en imágenes en escala de grises de dígitos escritos a mano, donde cada imagen es representada por un vector de píxeles

Supongamos que tenemos un conjunto de entrenamiento con 1000 imágenes de dígitos del 0 al 9

El *objetivo* es aplicar la cuantificación vectorial para reducir la dimensionalidad de las imágenes, manteniendo la información esencial de cada dígito.

Para ello, utilizaremos el enfoque de aprendizaje no supervisado con el algoritmo de agrupamiento k-means

Aprendizaje de cuantificación vectorial

1. Preparación de datos:

- Cada imagen en escala de grises se convierte en un vector unidimensional, por ejemplo, de tamaño 784 (28x28 píxeles)
- Se normalizan los valores de los píxeles para que estén en un rango común, como $[0, 1]$

2. Aplicación del algoritmo k-means:

- Se elige un número de centroides deseados, por ejemplo, $k=50$
- Se seleccionan aleatoriamente 50 imágenes como centroides iniciales
- Se asigna cada imagen restante al centroide más cercano en función de la distancia euclidiana
- Se actualizan los centroides recalculando la media de los vectores de las imágenes asignadas a cada centroide
- Se repiten los pasos de asignación y actualización hasta que se alcance una convergencia

Aprendizaje de cuantificación vectorial

3. Representación de imágenes comprimidas:

- Una vez que se obtienen los centroides finales, cada imagen se asigna al centroide más cercano
- En lugar de representar cada imagen con su vector original de píxeles, se representa con el índice del centroide al que pertenece
- Así, en lugar de un vector de tamaño 784, ahora cada imagen está representada por un único número entre 1 y 50, indicando su centroide correspondiente

Este proceso de aprendizaje de cuantificación vectorial reduce la dimensionalidad de las imágenes originales de manera significativa

En lugar de trabajar con vectores de alta dimensión, ahora podemos utilizar únicamente los índices de los centroides, lo que resulta en una representación más compacta y eficiente

Aprendizaje de cuantificación vectorial

- La ventaja de este enfoque es que las imágenes siguen conservando las características más importantes para distinguir los dígitos escritos a mano
- Además, al reducir la dimensionalidad, también se reduce el ruido y se simplifica el procesamiento de los datos, lo que puede acelerar las tareas de clasificación, reconocimiento o análisis posteriores
- Cabe mencionar que este es solo un ejemplo básico y que existen diferentes variaciones y mejoras del enfoque de cuantificación vectorial, así como otros algoritmos y técnicas que pueden aplicarse según el problema y el conjunto de datos específico

Aprendizaje de cuantificación vectorial

Aplicaciones en diversas áreas:

- La compresión de datos
- La recuperación de información
- El reconocimiento de patrones y el análisis de grandes conjuntos de datos

Y permite una representación más compacta de la información, lo que a su vez puede acelerar el procesamiento de datos y reducir los requisitos de almacenamiento

Aprendizaje de Cuantificación Vectorial con Ganadores Locales

El Aprendizaje de Cuantificación Vectorial con Ganadores Locales (LVQ, *por sus siglas en inglés: Learning Vector Quantization*) es un algoritmo de aprendizaje supervisado utilizado en el campo del reconocimiento de patrones y la clasificación de datos

Es una variante del algoritmo de cuantización vectorial que utiliza información de etiquetas o clases para entrenar los centroides

El **objetivo principal** del LVQ es encontrar un conjunto de centroides que representen de manera efectiva las clases o categorías de un conjunto de datos.

Cada centroide representa una clase o categoría específica y se utiliza para clasificar nuevos ejemplos basándose en su similitud con los centroides

Aprendizaje de Cuantificación Vectorial con Ganadores Locales

El proceso de aprendizaje del LVQ se puede resumir en los siguientes pasos:

1. Inicialización:

- Se seleccionan aleatoriamente un conjunto de centroides iniciales
- Cada centroide se etiqueta con una clase específica

2. Iteraciones:

- Se toma un ejemplo de entrenamiento del conjunto de datos
- Se calcula la distancia entre el ejemplo y cada uno de los centroides
- Se encuentra el centroide más cercano al ejemplo, conocido como el ganador
- Se actualizan los pesos del ganador y posiblemente de sus vecinos más cercanos, para acercarlos al ejemplo

Aprendizaje de Cuantificación Vectorial con Ganadores Locales

3. Convergencia:

- El proceso de actualización de los centroides se repite hasta que se alcance la convergencia, es decir, cuando los centroides dejan de cambiar significativamente
- Una vez que se ha entrenado el modelo LVQ, se puede utilizar para clasificar nuevos ejemplos
- Dado un ejemplo desconocido, se calcula la distancia con cada uno de los centroides y se asigna a la clase representada por el centroide más cercano
- El LVQ es un algoritmo efectivo para la clasificación de datos en problemas de reconocimiento de patrones

Aprendizaje de Cuantificación Vectorial con Ganadores Locales

- Pero el rendimiento del LVQ depende en gran medida de la calidad y representatividad de los centroides iniciales, así como de los parámetros de configuración adecuados, como la tasa de aprendizaje y la cantidad de vecinos actualizados

El Aprendizaje de Cuantificación Vectorial con Ganadores Locales (LVQ) es un algoritmo supervisado que utiliza centroides etiquetados para clasificar ejemplos y representa una variante del algoritmo de cuantización vectorial en la que se tiene en cuenta la información de clase o etiqueta