U N I V E R S I D A D D E G U A D A L A J A R A

CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERÍA
DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Seminario de Solución de Problemas de Sistemas Basados en Conocimiento

Tarea No. 3

Nombre: Hurtado González Edgar Arturo
Código: 212597894

Vectorización para Machine Learning

La vectorización es esencial en Machine Learning porque los algoritmos de aprendizaje automático, en su mayoría, trabajan con datos numéricos. Sin embargo, los datos del mundo real vienen en diversas formas: texto, imágenes, audio, tablas, etc., y no todos son directamente utilizables por estos algoritmos.

La vectorización convierte estos datos en vectores numéricos, que son conjuntos de números organizados en una secuencia unidimensional. Esta transformación permite que los algoritmos de Machine Learning procesen y analicen la información de manera más efectiva. Aquí hay más detalles sobre cómo se aplica la vectorización en diferentes tipos de datos:

Texto:

- **Bolsa de Palabras (Bag of Words):** Convierte el texto en vectores que representan la frecuencia de las palabras en un documento. Por ejemplo, si tienes un conjunto de oraciones, puedes crear vectores que indiquen cuántas veces aparece cada palabra en cada oración.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** Similar a la bolsa de palabras, pero pondera la importancia de las palabras en función de su frecuencia en el documento y en el corpus completo.
- **Embeddings de Palabras:** Transforma las palabras en vectores densos de números reales, donde la similitud semántica se refleja en la proximidad de los vectores en un espacio de alta dimensión.

Imágenes:

- **Aplanamiento de imágenes:** Convierte las imágenes en vectores unidimensionales colocando los píxeles en una secuencia.
- **Características extraídas por CNN:** Utiliza redes neuronales convolucionales para extraer características relevantes de las imágenes y luego representarlas como vectores numéricos.

Datos estructurados:

- **Datos numéricos:** Si ya tienes datos numéricos o en forma de tablas, la vectorización puede ser simplemente representar estos datos como vectores.
- **Normalización y transformación:** Puedes aplicar técnicas como normalización (escalamiento de los datos para que estén en el mismo rango) o transformaciones como PCA (Análisis de Componentes Principales) para reducir la dimensionalidad y representar los datos de manera más eficiente.

En resumen, la vectorización es el proceso fundamental que permite que los datos no estructurados se conviertan en una forma que los algoritmos de Machine Learning puedan entender y procesar, lo que facilita el aprendizaje de patrones y la toma de decisiones basadas en datos.

Label Encoding

El Label Encoding es una técnica utilizada en Machine Learning para convertir variables categóricas en números enteros. En el aprendizaje automático, muchos algoritmos requieren que los datos de entrada sean numéricos, y el Label Encoding es una forma de convertir categorías en números de manera que el algoritmo pueda interpretarlas. Algunos puntos clave sobre el Label Encoding serían:

Proceso de Label Encoding:

1. **Identificación de variables categóricas:** Selecciona las columnas o características que contienen valores categóricos, por ejemplo, las categorías de productos, colores, etiquetas de clases, etc.
2. **Asignación de números enteros:** Para cada categoría única en la variable categórica, se asigna un número entero. Por ejemplo, si tienes una columna de colores con ['rojo', 'verde', 'azul'], podrías asignar 'rojo' a 0, 'verde' a 1 y 'azul' a 2.

Consideraciones:

- **Orden implícito:** El Label Encoding puede generar un orden implícito en los datos cuando se convierten las categorías en números. Algunos algoritmos podrían interpretar esto como un orden o jerarquía entre las categorías, lo cual no siempre es deseado. Por ejemplo, si se codifican etiquetas de clases como 0, 1 y 2, un algoritmo podría asumir que la clase 2 es "mayor" o "mejor" que la clase 0, lo cual no es necesariamente cierto.
- **No es adecuado para todas las situaciones:** El Label Encoding puede funcionar bien cuando hay una relación ordinal entre las categorías (por ejemplo, 'bajo', 'medio', 'alto'), pero no es apropiado para variables categóricas donde no existe un orden inherente entre las categorías.

El Label Encoding es útil en algunos casos, pero es importante tener en cuenta sus limitaciones, especialmente si se utilizan algoritmos sensibles a las magnitudes numéricas o cuando se trabaja con variables categóricas sin orden específico. Otras técnicas como One-Hot Encoding o el uso de embeddings pueden ser preferibles en ciertas situaciones.

One-Hot Encoding

El One-Hot Encoding es otra técnica utilizada en Machine Learning para manejar variables categóricas, especialmente útil cuando las categorías no tienen una relación de orden o jerarquía entre sí. Esta técnica convierte cada categoría en una representación binaria en la que cada categoría se convierte en una nueva columna y se le asigna un valor binario (1 o 0) según su presencia en la observación original.

Proceso de One-Hot Encoding:

1. **Identificación de variables categóricas: Al** igual que con el Label Encoding, se seleccionan las características que contienen valores categóricos.
2. **Creación de nuevas columnas:** Para cada categoría única en la variable categórica, se crea una nueva columna. Cada fila recibe un 1 en la columna correspondiente a su categoría y 0 en todas las demás columnas.

Consideraciones:

- **Aumento en la dimensionalidad:** El One-Hot Encoding puede aumentar significativamente la dimensionalidad de los datos, especialmente si hay muchas categorías. Esto puede llevar a conjuntos de datos más grandes y a un aumento en la complejidad computacional.
- **Esparsidad de los datos:** Si hay muchas categorías, la representación One-Hot puede generar matrices con muchos ceros, lo que se conoce como "esparsidad". En tales casos, podrían ser preferibles técnicas como embeddings para representar las categorías de manera más compacta.

El One-Hot Encoding es una técnica útil y ampliamente utilizada para trabajar con variables categóricas en el aprendizaje automático, especialmente cuando no hay una relación ordinal entre las categorías y se necesita una representación binaria distintiva para cada una. Sin embargo, su uso debe ser evaluado en función del contexto y las características específicas de los datos.

Estas técnicas son fundamentales para preparar y transformar datos para su uso en modelos de Machine Learning, pero es esencial comprender sus implicaciones y aplicarlas de manera adecuada según el tipo de datos y el contexto del problema a resolver.