Centro de Investigación en Matemáticas (CIMAT).
Unidad Monterrey

Ciencia de Datos Tarea 0

Diego Godinez Bravo

20 de enero de 2024

1. Para los siguientes conjuntos de datos, formula al menos una pregunta de investigación que puedas contestar como un problema de aprendizaje no supervisado, y al menos una pregunta de investigación que puedas contestar como un problema de aprendizaje supervisado. En ambos casos, indica las variables involucradas (en el segundo caso, las variables dependientes e independientes). ¿Son suficientes los datos/variables que tienes o te interesaría incluir otros? Si es así, menciona de dónde podrías obtenerlos.

a) Esta base de datos (BD) contiene el registro de todos los delitos de alto impacto (con los crímenes violentos y no violentos por separado), ocurridos en la Ciudad de México durante 2014, 2015 y 2016.

1. **Aprendizaje No Supervisado**

   Pregunta de Investigación: ¿Existe una división territorial de la Ciudad de México basada en los diferentes tipos de delitos considerando las diferentes épocas del año?

   Variables involucradas: Tipo de delito (columna DELITO SEPARADO); ubicación (columnas CUADRANTE, CX, CY); fecha (columnas FECHA, MES)

2. **Aprendizaje Supervisado**

   Pregunta de Investigación: ¿Se puede predecir el tipo de delito que predominará en un cierto periodo de tiempo y región específica de la Ciudad de México básandonos en la ubicación y hora del día?

   Variables Dependientes: Tipo de delito (columna DELITO SEPARADO)

Variables Independientes: Ubicación (columnas CUADRANTE, CX, CY); hora del día (columna HORA); mes del año (columna MES)

b) Esta base de datos contiene registros de divorcios del año 2000 al 2015 en la ciudad de Xalapa, Ver. Las columnas que contiene ésta base de datos son las siguientes.

1. **Aprendizaje No Supervisado**

   Pregunta de Investigación: ¿En que medida las variables demográficas y socioeconómicas influencian la tasa de divorcio anual?

   Variables involucradas: Ingresos mensuales de la pareja; Nivel educativo de la pareja; Ocupación de la pareja; Edad de la pareja; Nacionalidad de la pareja; Lugar de residencia de la pareja

2. **Aprendizaje Supervisado**

   Pregunta de Investigación: ¿Es posible predecir la probabilidad de divorcio de parejas que residen en la ciudad de Xalapa basándonos en las variables socioeconómicas de la pareja?

   Variables Dependientes: Tasa de divorcio

   Variables Independientes: Ingresos mensuales de la pareja; Nivel educativo de la pareja; Ocupación de la pareja; Edad de la pareja

c) Esta BD contiene más de 150 mil registros de letras de canciones de diferentes artistas, incluyendo la variable *valence*, que es un indicador obtenido con la API de spotify, que lo describe como: *A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).*

1. **Aprendizaje No Supervisado**

   Pregunta de Investigación: ¿Hay palabras clave que tengan un impacto significativo en la asignación de una calificación (*valence*) específica a una canción?

   Variables involucradas: Valor de la variable *valence*; Letra de la canción

2. **Aprendizaje Supervisado**

   Pregunta de Investigación: ¿Es posible predecir el valor de la variable *valence* de una canción mediante el análisis de su letra y título?

   Variables Dependientes: Valor de la variable *valence*

   Variables Independientes: Letra y título de la canción; Artista que interpreta la canción

d) Esta BD contiene alrededor de 40 mil registros que corresponden a características de artículos publicados en línea durante un periodo de dos años. Estos artículos en formato texto, fueron preprocesados para extraer diferentes características que se muestran a continuación:

1. **Aprendizaje No Supervisado**

   Pregunta de Investigación: ¿Existen tendencias estacionales en los tipos de artículos que son publicados y adquiridos a lo largo de un año?

   Variables involucradas: Canal donde es publicado el artículo (i.e. 'Lifestyle', 'Entertainment', 'Business', 'Social Media', ...); Día de la semana en que el artículo es publicado; Título de la publicación; Número total de palabras en la publicación; Tasa de palabras con un sentimiento positivo/negativo en la publicación; Palabras clave

2. **Aprendizaje Supervisado**

   Pregunta de Investigación: ¿Se puede predecir el tiempo promedio que un artículo requiere para ser adquirido, considerando sus atributos intrínsecos y la manera en que se publica o promociona?

   Variables Dependientes: Tiempo promedio de adquisición de un artículo específico

   Variables Independientes: Canal donde es publicado el artículo (i.e. 'Lifestyle', 'Entertainment', 'Business', 'Social Media', ...); Día de la semana en que el artículo es publicado; Título de la publicación; Número total de palabras en la publicación; Tasa de palabras con un sentimiento positivo/negativo en la publicación; Palabras clave

   En este caso una variable adicional que se requiere es la fecha en la cual un artículo es adquirido por un cliente, datos los cuales se podrían obtener directamente de la plataforma en la cual los artículos son publicados.

e) Esta BD contiene registros de los viajes realizados desde 2014 a la fecha, por los usuarios del sistema de bicicletas públicas del Área Metropolitana de Guadalajara. La información disponible incluye lo siguiente:

1. **Aprendizaje No Supervisado**

   Pregunta de Investigación: ¿Se ve influenciada la demanda del servicio por variables demográficas y/o variables socioeconómicas relacionadas tanto con los usuarios como con las estaciones disponibles?

   Variables involucradas: Información relevante del viaje (i.e. ID del viaje, información acerca del inicio y término del viaje); información relacionada a las estaciones involucradas en el viaje; información relevante del usuario (e.g. edad, género, ocupación)

2. **Aprendizaje Supervisado**

   Pregunta de Investigación: ¿Es posible predecir el número promedio de viajes que se llevarán a cabo en una zona durante un período de tiempo específico a fin de gestionar la disponibilidad de bicicletas en las estaciones ubicadas en las zonas más transitadas?

Variables Dependientes: Número promedio de viajes

Variables Independientes: Información relevante del viaje (i.e. ID del viaje, información acerca del inicio y término del viaje); información relacionada a las estaciones involucradas en el viaje