Universidad Austral - Maestría en Ciencia de Datos

Introduccion al Data Mining

Trabajo Practico #1 v2

Autor: Andres Montes de Oca

Fecha: 2 de Diciembre de 2022

1) Que es el Data Mining?:

La minería de datos tiene varias definiciones según distintos autores y según sus fechas de publicación. En resumen, podemos decir que Data Mining es el proceso de descubrir insights desde mis datos, los cuales no son visibles a simple vista, se necesita "escarbar" para poder encontrar información relevante. Data Mining no es una moda, yo lo vería mas como una técnica que ha ido evolucionando y transformándose a lo largo del tiempo. En principio se la definía erroneamente como un proceso automatico en la extracción de insights, pero con el tiempo eso fue cambiando cuando se empezó a dar cuenta que la intervención y el raciocinio humano eran necesarios. El data mining toma sus bases en la Estadística, Machine Learning y Sistemas de Gestion de Base de Datos (SGBDs), y hace frente a los problemas que genera el inmenso volumen de Datos que disponemos actualmente (alta dimensionalidad, heterogeneidad, escalabilidad, etc). Si bien Data Mining y KDD muchas veces se usan como sinónimos, estos no lo son. KDD es un concepto mucho mas amplio, el cuales incluye todos los pasos o procesos necesarios para la extracción de Inisghts desde los datos raw. Data Mining es solo uno de estos pasos, es el cual se encarga de la búsqueda de patrones de interés en nuestros datos

2) Tareas de Data Mining:
   - "Dividir los clientes de una compañía de acuerdo con su género"
     Falso, esto es una simple consulta a la BBDD, un variable binomial como agrupamiento
   - "Dividir los clientes de una compañía de acuerdo con su rentabilidad"
     Posible Verdadero, en este caso al tomar una variable continua para el agrupamiento sin estar divididos por categorías, puedo utilizar un algoritmo como Clustering para la creación de los grupos
   - "Calcular el total de ventas de una compañía"
     Falso, es solo una sumarizacion de datos
   - "Clasificar una base de datos de estudiantes basado en los números de identificación."
     Falso, los números identificatorios deberían ser únicos, no necesito ninguna técnica ni algoritmo para poder claisificar los estudiantes por ID, ya vienen así en la BBDD
   - "Predecir los resultados de arrojar un par de dados"
     Falso, esto puede ser hecho con un modelo estadístico, teniendo en cuenta las distintas Esperanzas posibles en los resultados
   - "Predecir el precio futuro de las acciones de una compañía usando registros históricos"
     Verdadero, aquí puedo aplicar una técnica de Regresión analizando el cambio histórico en el precio de las acciones para predecir su comportamiento
   - "Monitorear los latidos del corazón de un paciente para buscar anomalías"
     Verdadero, si en mis datos tengo registros de pacientes con anomalías previamente detectadas, puedo correr un modelo de clasificación que busuqe la relación entre todas las variables y me prediga (con un cierto nivel de accuracy) si el paciente en estudio es propenso a tener una anomalia o no

- “Monitorear ondas sísmicas para detectar actividades de terremotos”
  Verdadero, exactamente como el punto anterior, mientras tenga datos históricos de terremotos previos, puedo generar un modelo de Clasificación
- “Extraer las frecuencias de una onda de sonido”
  Verdadero, aquí puedo correr técnicas de aprendizaje no supervisado (como PCA), para la eliminación del ruido en la deteccion de ondas de sonido

3) Ejemplo Data Mining:

Se me ocurre como ejemplo una empresa que se dedique al servicio de auxilio mecánico express en CABA, donde su servicio se caracterice por ser rápido. La empresa debería disponer de información histórica de auxilios prestados anteriormente, para poder realizar algún tipo de modelo predictivo. El modelo nos debería ayudar en el planeamiento de la distribución de las unidades de auxilio mecanico por todo el ámbito de la ciudad. Una tarea fundamental del Data Mining para este ejemplo podría ser el Análisis de Correlación entre eventos de distinto tipo, como los climáticos (la temperatura), o los sociales (algún evento multitudinario), y la demanda de los servicios de auxilio.

4) Consutlor Data Mining para Google:

Podríamos aplicar técnicas de clustering para poder agrupar a nuestros usuarios según el tipo de búsqueda que hacen. O si hacen compras on-line, aplicar algún modelo de clasificación que nos ayude en la elección del producto a ofrecerle como recomendado, teniendo en cuenta los patrones de búsqueda. La verdad es que no termino de entender la pregunta.

5) Importancia en la privacidad de los datos:

- “Datos censales del período 1980-1998.”
  En los datos censales no creo que la privacidad sea un aspecto relevante, siempre y cuando en los mismos no se publiquen datos personales sobre los habitantes. Ademas son datos agregados
- “Direcciones de IP y tiempos de visita de usuarios Web que visitan su sitio web”
- En este caso la privacidad si es importante, ya que por dirección ip puedo llegar a geolocalizar a una persona.
- “Imágenes satelitales de una región.”
  Según en nivel de detalle, si es que i la calidad de las mismas pueden comprometer la privacidad de las personas.
- “Nombres y direcciones de e-mail recolectadas de la web.”
  Aquí también la privacidad es importante, ya que si bien las direcciones de correo por si solas no brindan mucha información personal, pueden ser el punto de partida para tareas malintencionadas hacia los usuarios.

6) Diferencias entre Crisp-DM y SEMMA:

CRISP-DM tiene una fase adicional al comienzo, relacionada con el entendimiento de los requerimientos de negocio, y otra fase adicional al final relacionada con la implementación o puesta en producción de mi modelo. Pero ambas metodologías también tienen varias etapas en común, como la recolección de datos con sus EDAs, transformaciones de variables, selección de el/los modelos a utilizar, y sus evaluaciones de performance.