RESUMEN

MINERIA DE DATOS

Docente: Mayra Cristina Berrones Reyes

Alumno: Jose Luis Buendía Meza 1813456

Grupo: 003

A viernes 2 de octubre del 2020

PATRONES SECUENCIALES

Minería de datos Secuenciales: Extracción de patrones frecuentes relacionados con el tiempo o algún otro tipo de secuencia. Eventos que se enlazan con el paso del tiempo (El orden de acontecimiento es importante).

Características:

- El orden importa.
- Objetivo: encontrar patrones secuenciales.
- El tamaño de una secuencia es su cantidad de elementos.
- La longitud de la secuencia es la cantidad de ítems.
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Aplicaciones

- Medicina: Predecir si un compuesto químico causa cáncer.
- Análisis de mercado: comportamiento de compras.
- Web: Reconocimiento de spam de un correo electrónico.

Secuencias

$s = < e_1\ e_2\ e_3\ e_4 \dots >$ — **|s| es el número de elementos en una secuencia**

$e_i = \{ i_1, i_2, i_3, \dots, i_k \}$ — **Una k-secuencia es una secuencia con k eventos**

Secuencia de visitas en una página web
< {Homepage} {Electronics} {Tablets} {Kindle Fire HD} {Shopping Cart} {Order Confirmation} {Return to Shopping} >

Fases del método GSP (Generalized Sequential Pattern)

Fase 1:
Recorrer la base de datos para obtener todas las secuencias frecuentes de 1 elemento.

Fase 2:

Generación:
Generar k-secuencias candidatas a partir de las (k-1)-secuencias frecuentes.

Poda:
Podar k-secuencias candidatas que contengan alguna (k-1)-secuencia no frecuente.

Conteo:
Obtener el soporte de las candidatas

Eliminación:
Eliminar las k-secuencias candidatas cuyo soporte real esté por debajo del Umbral de soporte mínimo de frecuencia

REGRESIÓN

Regresión: modelo matemático para determinar el grado de dependencia entre una o más variables, es decir conocer si existe relación entre ellas.
- En **Minería de Datos** la Regresión se encuentra dentro de la categoría Predictivo.

Análisis de Regresión: Permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés.

- *Variable(s) dependiente(s)*: Es el factor más importante, el cual se está tratando de entender o predecir.
- *Variable(s) independiente(s)*: Es el factor que tú crees que puede impactar en tu variable dependiente.

El **análisis de regresión** nos permite explicar un fenómeno y predecir cosas acerca del futuro, por lo que nos será de ayuda para tomar decisiones y obtener los mejores resultados.

Regresión Lineal en Python 3

La idea consiste en obtener una ecuación de la forma: $y = mx + b$ que se ajuste mejor a los datos que se tengan.

Sabemos que: $m = \frac{\sum x \sum y - n\sum(xy)}{(\sum x)(\sum x)^2 - n\sum x^2}$ y $b = \bar{y} - m\bar{x}$

Para determinar qué tan bueno es nuestro ajuste existen diferentes parámetros estadísticos, pero en este caso utilizaremos el coeficiente de determinación: $R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ donde;

$$\sigma_x = \sqrt{\frac{\sum(x^2)}{n} - \bar{x}^2}, \quad \sigma_y = \sqrt{\frac{\sum(y^2)}{n} - \bar{y}^2}$$
$$\sigma_{xy} = \frac{\sum(xy)}{n} - \bar{x} \cdot \bar{y}$$

CLASIFICACIÓN

Clasificación: Es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

Algunos métodos:

- **Análisis discriminante**: método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos.
- **Reglas de clasificación**: buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación.
- **Arboles de decisión**: método analítico que a través de una representación esquemática facilita la toma de decisiones.
- **Redes neuronales artificiales**: (también conocido como sistema conexionista) es un modelo de unidades conectadas para trasmitir señales.

Características de los métodos:

- Precisión en la predicción
- Eficiencia
- Robustez
- Escalabilidad
- Interpretabilidad

CLUSTERING

Clustering: También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares. Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases.

Un **cluster** es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.

Análisis de cluster: dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

Aplicaciones

- **Estudios de terremotos**: los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.
- **Marketing**: ayudar a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.
- **Aseguradoras**: identificación de grupos de asegurados de seguros de automóviles con un alto costo promedio de reclamo.

Algunos **Algoritmos de Clustering**

- **Simple K-Means**: Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar.
- **X-Means**: (Variante mejorada del K-Means) Su ventaja fundamental es que ya no se tiene que ingresar el numero deseado de clusters, a X-Means se le define un límite inferior K-min (número mínimo de clusters) y un límite superior K-Max (número máximo de clusters) y este algoritmo es capaz de obtener en ese rango el número óptimo de clusters.
- **Cobweb**: Pertenece a la familia de algoritmos jerárquicos. Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos.

OUTLIERS

Detección de Outliers: Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

Valores atípicos: Observaciones cuyos valores son muy diferentes al comportamiento de las demás observaciones del mismo grupo de datos; Los datos atípicos distorsionan los resultados del análisis y por esta razón hay que identificarlos y tratarlos adecuadamente.

¿Cómo se calculan los valores atípicos?

Existen distintas técnicas para detectarlos y se pueden dividir en dos categorías diferentes:

- Métodos univariantes de detección de Outliers.
- Métodos multivariantes de detección de Outliers.

Técnicas para la detección de Outliers:

1. Prueba de Grubbs
2. Prueba de Dixon
3. Prueba de Tukey
4. Regresión Simple
5. Análisis de valores atípicos de Mahalanobis

¿Qué hacer cuando son detectados los valores atípicos?

Se pueden eliminar o sustituir si fueron debidos a un error de captura o en la medición de la variable; Si no se deben a un error, eliminarlo o sustituirlo pueden modificar las inferencias que se realicen a partir de esa información, debido a que:

- Introduce un sesgo
- Disminuye el tamaño muestral
- Puede afectar a la distribución y a las varianzas

Aplicaciones de la minería de datos en la detección de Outliers:

- Detección de fraudes financieros
- Tecnología informática y telecomunicaciones
- Nutrición y Salud
- Negocios

PREDICCIÓN

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento.

Aplicaciones

1. Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro
2. Predecir si va a llover en función de la humedad actual.
3. Predecir la puntuación de cualquier equipo durante un partido de fútbol.

Técnicas

La mayoría de las técnicas de predicción se basan en modelos matemáticos:

1. Modelos estadísticos simples como regresión

2. Estadísticas no lineales como series de potencias

3. Redes neuronales, RBF, etc.

Tipos de métodos de regresión

- ***Regresión lineal***: En el Análisis de regresión simple, se pretende estudiar y explicar el comportamiento de una variable que notamos y, y que llamaremos variable dependiente o variable de interés, a partir de otra variable, que notamos x, y que llamamos variable explicativa, variable de predicción o variable independiente.
- ***Regresión lineal multivariante***: Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella
- ***Regresión no lineal***: Método para encontrar un modelo no lineal para la relación entre la variable dependiente y un conjunto de variables independientes.

Redes Neuronales: Utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión. Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

REGLAS DE ASOCIACIÓN

Reglas de asociación: Búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos de transacciones, bases de datos relacionales y otros repositorios de información disponibles.

Aplicaciones

- Análisis de datos de la banca.
- Cross-marketing (poner la crema batida junto a las fresas).
- Diseño de catálogos.

Objetivo: Dado un conjunto de transacciones T, el objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo:

- Umbral mínimo de soporte
- Umbral mínimo de confianza
- Lista todas las reglas de asociación posibles.
- Compruebe el soporte y la confianza para cada regla.
- Elimine las reglas que fallan en los umbrales mínimos.

Técnicas

- **RAM (Enfoque de dos pasos):**
  Generación de elementos frecuentes: Generar todos los conjuntos de elementos cuyo soporte ≥ min sup.
  Generación de reglas: Generar reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuente.
- Principio de Apriori: si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes.
  El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos. Esto se conoce como la propiedad *anti-monótona* de soporte.

VISUALIZACIÓN DE DATOS

La Visualización de datos nos sirve para representar gráficamente los elementos más importantes de nuestra base de datos. La visualización de datos es la presentación de información en formato ilustrado o gráfico. Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Tipos

- **Gráficos:** Este es el tipo más común y conocido, que utilizamos en nuestro día a día con las hojas de cálculo, para representar datos de manera sencilla, como Gráficos Circulares, Líneas, Columnas y Barras aisladas o agrupadas, Burbujas, áreas, Diagramas de Dispersión y Mapas de tipo Árbol.
- **Mapas:** Todos conocemos la visualización de datos en mapas para conocer, por ejemplo, la localización de nuestra flota de vehículos en tiempo real o bien la de las tiendas de un supermercado o los cajeros automáticos de nuestro banco en un mapa.
- **Infografías:** Una infografía es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente. son excelentes para ayudarnos a procesar más fácil, la información compleja.
- **Cuadros de Mando (Dashboards):** En el entorno empresarial, un cuadro de mando es una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos, etc.

La mayoría de los analistas de datos utilizan software avanzado para explorar y visualizar datos. Y las herramientas de software van desde Hojas de Cálculo sencillas con Excel o Google Sheets a software de analítica más sofisticado, como R.

Aplicaciones

- **Comunique la Historia:** Una vez que una empresa ha descubierto nuevos insights a partir de la analítica visual, el paso siguiente consiste en comunicar esos insights a otras personas.
- **Identifique tendencias emergentes:** El uso de la visualización de datos para descubrir tendencias en los negocios y en el mercado puede dar a las empresas una ventaja sobre la competencia, y eventualmente tener un impacto en la base de operación.