UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Actividad 5

Minería de Datos

Margarita Ordaz Ruiz

Maestro: Diana Elizabeth Torres Valdés

Grupo: 003

Matricula: 1802473

REGLAS DE ASOCIACIÓN

Las reglas de asociación son de un tipo de análisis que extrae información por coincidencias, con el fin de encontrar relaciones dentro un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta. Estas reglas nos permiten encontrar las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos y medir la fuerza e importancia de estas combinaciones.

Algunas de las aplicaciones del análisis mencionado anteriormente pueden ser el ordenamiento de productos, definir patrones de navegación dentro de tiendas, sugerir promociones efectivas de pares de productos, generar descuentos específicos para cada cliente.

Tipos de Reglas de Asociación:

- *Asociación Cuantitativa*, Con base en los tipos de valores que manejan las reglas:
    1. *Asociación Booleana*: las cuales son asociaciones entre la presencia o ausencia de un ítem.
    2. *Asociación Cuantitativa*: estas nos describen asociaciones entre ítems cuantitativos o atributos.
- *Asociación Multidimensional,* Con base en las dimensiones de datos que involucra una regla:
    1. *Asociación Unidimensional*: Si los ítems o atributos de la regla se referencian en una sola dimensión.
    2. *Asociación Multidimensional*: Si los ítems o atributos de la regla se referencian en dos o más dimensiones.
- *Asociación Multinivel,* Con base en los niveles de abstracción que involucra la regla:
    1. *Asociación de un nivel*: Los ítems son referenciados en un único nivel de abstracción.
    2. *Asociación Multinivel*: Los ítems son referenciados a varios niveles de abstracción

Métricas de Interés:

- El ***Soporte*** tiene que ver con el número de veces que aparece un itemset en la base de datos.
- Dada una regla “Si A => B”, la ***confianza*** de esta regla es el cociente del soporte de la regla y el soporte del antecedente solamente, la confianza mide la fortaleza de la regla.
- El ***lift*** refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente.

DECTECCIÓN DE OUTLIERS

Un outlier se define como una observación que se desvía mucho del resto de las observaciones apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos, en otras palabras, los outliers en nuestro dataset serán los valores que se "escapan al rango en donde se concentran la mayoría de las muestras".

¿Y por qué nos interesa detectar esos Outliers? Porque pueden afectar considerablemente a los resultados que puedan obtener nuestros modelos. Por eso hay que detectarlos, y tenerlos en cuenta.

Los Outliers pueden significar varias cosas:

- *ERROR*: Si tenemos un grupo de "edades de personas" y tenemos una persona con 160 años, seguramente sea un error de carga de datos. En este caso, la detección de outliers nos ayuda a detectar errores.
- *LIMITES*: En otros casos, podemos tener valores que se escapan del "grupo medio", pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo.
- *Punto de Interés*: puede que sean los casos "anómalos" los que queremos detectar y que sean nuestro objetivo.

Los outliers se pueden detectar mediante gráficas, una gráfica de detección sencilla es el Boxplot o Diagrama de cajas, podemos visualizar las variables y en esa "cajita" veremos donde se concentra el 50 por ciento de nuestra distribución (percentiles 25 a 75), los valores mínimos y máximos (las rayas en "T") y -por supuesto- los outliers, esos "valores extraños" y alejados. También se pueden detectar en gráficos de dispersión cuando son de 1 a 3 dimensiones, y cuando son de más de 3 con ayuda de librerías de Python como lo es PyOD.

Una vez detectados, ¿qué hago? Según la lógica de negocio podemos actuar de una manera u otra. Por ejemplo, podríamos decidir:

- Las edades fuera de la distribución normal, eliminar.
- El salario que sobrepasa el límite, asignar el valor máximo (media + 2 sigmas).
- Las compras mensuales, mantener sin cambios.

REGRESIÓN

La regresión es una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Regresión lineal simple: Cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple. La regresión lineal simple tiene como modelo: $y = \beta_0 + \beta_1 x + e$. Donde "e" es una variable aleatoria normalmente distribuida con E(e)=0 y Var(e)= σ^2. Se utiliza *la estimación por mínimos cuadrados*, y el modelo ajustado por esta estimación queda como: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Regresión lineal múltiple: Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos. En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$. Al igual que en la regresión lineal simple se utiliza *la estimación por mínimos cuadrados*, y el modelo ajustado por esta estimación queda como: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$.

Regresión con variables transformadas: La necesidad de un modelo alternativo para el modelo probabilístico lineal $y = \beta_0 + \beta_1 x + e$, puede ser sugerida ya sea por un argumento teórico o al examinar gráficas de diagnóstico desde un análisis de regresión lineal. Una clase importante de estos modelos se especifica por medio de funciones que sean 'intrínsecamente lineales'.

Una función que relacione y con x es intrínsecamente lineal si por medio de una transformación de x y/o y, la función se puede expresar como $\hat{y}' = \hat{\beta}_0 + \hat{\beta}_1 x'$,donde x'= la variable independiente transformada y y'= la variable dependiente transformada. Cuatro de las funciones intrínsecamente lineales más útiles son la exponencial, la potencia, la logarítmica y la potencia.

Regresión polinómica: En numerosas situaciones, ya sea de un razonamiento teórico o de otro tipo, una gráfica de puntos sugiere que la verdadera función de regresión tiene uno o más picos o valles, es decir, al menos un mínimo o máximo relativos. En tales casos, una función con polinomios $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_k x^k$ puede dar una aproximación satisfactoria a la verdadera función de regresión.

Aplicaciones, la regresión es muy útil para poder hacer pronósticos, buscamos si hay relación entre las variables y con base a eso podemos predecir un valor futuro, esto se puede aplicar en la industria, en medicina, en estadística, informática e incluso en el comportamiento humano.

CLUSTERING

El clustering es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes. Se utiliza para investigación de mercado, identificar comunidades, prevención de crimen, procesamiento de imágenes. Los datos se pueden transformar en variables cuantitativas, binarias y categóricas. Los tipos básicos de análisis son:

Centroid Based Clustering: Cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más usado de este tipo es el de K-medias.

Connectivity Based Clustering: Los clusters se definen agrupando a los datos más similares o cercanos. La característica principal es que un cluster contiene a otros clusters, es decir, representan una jerarquía. Un algoritmo usado de este tipo es Hierarchical clustering.

Distribution Based Clustering: En este método cada cluster pertenece a una distribución normal, la idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. Un algoritmo de clustering perteneciente a este tipo es Gaussian mixture models.

Density Based Clustering: Los clusters son definidos por áreas de concentración. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters.

VISUALIZACIÓN

La visualización de datos es la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. Es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Existen multitud de técnicas y aproximaciones para la visualización según sea la naturaleza del dato de la información. Según la complejidad y elaboración de la información podemos tener la siguiente clasificación:

1. *Elementos básicos de representación de datos:*
    - Gráficas: barras, líneas, columnas, puntos, "tree maps", tarta, semi-tarta, etc.
    - Mapas: burbujas, mapa temático, mapa de calor, de agregación.
    - Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.
2. Un *cuadro de mando* es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Son ampliamente utilizados en las organizaciones para análisis de conjuntos de variables y toma de decisiones.
3. *Infografías* se utilizan para contar "historias". Esta se narra mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

También se pueden utilizar softwares para la visualización de datos como lo son HTML5, CSS3, SCV, WebGL.

La visualización de datos es muy importante en cualquier empleo, los conjuntos de habilidades están cambiando para adaptarse a un mundo basado en los datos. Para los profesionales es cada vez más valioso poder usar los datos para tomar decisiones y usar elementos visuales para contar historias con los datos para informar quién, qué, cuándo, dónde y cómo. La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

PATRONES SECUENCIALES

Los patrones secuenciales se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, describe el modelo de compras que hace un cliente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo.

Se trata de buscar asociaciones de la forma que si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n. El objetivo de la tarea es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos. Utiliza reglas que expresan patrones de comportamiento secuencial.

Estos patrones se desarrollan en áreas como la medicina, la biología, bioingeniería, Análisis de mercado, distribución y comercio, aplicaciones financieras y banca, en las aplicaciones de seguro y salud privada y en los deportes. Y los tipos de bases de datos en los que se desarrollan son los temporales, documentales y relacionales.

Y sus principales características son que el orden sí importa, su tamaño es su cantidad de elementos (itemsets), su longitud es su cantidad de ítems y su soporte es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.

Para la resolución de problemas se utiliza la *agrupación de patrones secuenciales*, la cual es la tarea de separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí, después la *clasificación con datos secuenciales*, y las *reglas de asociación con datos secuenciales* se presenta cuando los datos contiguos presentan algún tipo de relación.

CLASIFICACIÓN

La clasificación es una técnica de minería de datos, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características. Sirve para estimar un modelo usando los datos recolectados para hacer predicciones futuras.

Existen algunas técnicas para la clasificación como lo son, la *clasificación Bayesiana*, la cual como su nombre lo dice, se basa en el teorema de Bayes, las *redes neuronales*, esta técnica trabaja directamente con números y en caso de que se desee trabajar con datos nominales, se deben de enumerar, estas redes consisten generalmente de tres capas: de entrada, oculta y de salida, y no sólo se utilizan para la clasificación, sino que también se utiliza para agrupamiento y regresión.

El árbol de decisión es otra técnica, estas son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos, al igual que las redes neuronales, es útil para clasificación, agrupamiento, regresión, sin embargo, esta técnica puede traer problemas como lo son que las reglas no necesariamente formen un árbol, y que no cubran todas las posibilidades.

Otras técnicas que normalmente se utilizan son el Support Vector Machines (SVM) y la Clasificación basada en asociaciones.

PREDICCIÓN

Para hacer un buen modelo de predicción es necesario tener en cuenta algunos elementos antes de empezar con la predicción, estos elementos son primero definir adecuadamente nuestro problema, después recopilar los datos, elegir una medida o indicador de éxito y preparar los datos, es decir, elegir los datos que se consideren más relevantes para la resolución de nuestros problemas. El siguiente paso es dividir los datos, de la base de datos limpia, el 70% de los datos son un conjunto de entrenamiento, 15% son para la validación y 15% es para hacer las pruebas.

Para las predicciones se utilizan técnicas como árboles de aleatorios como lo son el árbol de decisión, el cual es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente, también tenemos el árbol de clasificación, el árbol de regresión.

Otra técnica utilizada son los bosques aleatorios o Random Forest, esta es una técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión.