Técnicas de la minería de datos

Reglas de asociación

Las reglas de asociación se derivan de un tipo de análisis que extrae información por coincidencias, con el objetivo de encontrar relaciones dentro un conjunto de transacciones que tienden a ocurrir de forma conjunta.

Estas reglas se definen como una implicación del tipo: “Si A => B “, donde A y B son individuales.

Estas reglas de asociación tienen diferentes aplicaciones en la vida cotidiana, como promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancías en tiendas, etcétera.

Tipos de reglas de asociación:

-Asociación Booleana, son asociaciones entre la presencia o ausencia de un ítem.

-Asociación Cuantitativa, describe asociaciones entre ítems cuantitativos o atributos.

-Asociación Unidimensional, si los ítems o atributos de la regla se referencian en una sola dimensión.

-Asociación Multidimensional, si los ítems o atributos de la regla se referencian en dos o más dimensiones.

Métricas de interés:

-Soporte, dada una regla “Si A => B”, el soporte de esta regla se define como el número de veces con que A y B aparecen juntos en una base de datos de transacciones.

-Confianza, 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.

-Lift, refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente

Predicción

La técnica de predicción tiene varios modelos, los cuales son:

-Árbol de decisión, que es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable dependiente. Se divide el

espacio muestral en subregiones y se aplica una serie de reglas. Estos árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

-Árbol de clasificación, el cual Consiste en hacer preguntas del tipo ¿$xk \leq c$? para las covariables cuantitativas o preguntas del tipo ¿$xk = nivelj$? para las covariables cualitativas.

-Árbol de regresión, consiste en hacer preguntas de tipo ¿$xk \leq c$? para cada una de las covariables.

-Bosque aleatorio, el cual es una técnica de aprendizaje automático supervisada basada en los árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar, se consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarse que los árboles sean distintos, cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging.

-Validación cruzada, Se emplea para estimar la prueba error rate de un modelo y así evaluar su capacidad predictiva, a este proceso se le conoce como model assessment. También se puede emplear para seleccionar el nivel de flexibilidad adecuado.

Clustering

Es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos. Los principales usos del clustering son: investigación de mercado, prevención del crimen, procesamiento de imágenes, etcétera.

Tipos de análisis básicos del clustering:

-Centroid based clustering, dónde cada cluster es representado por un centroide, además los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. En este tipo de análisis, el algoritmo más común es el algoritmo de K-medias.

-Connectivity based clustering, aquí los clusters se definen agrupando a los datos más similares o cercanos, la característica principal de este tipo de análisis es que un cluster contiene a otros clusters, y el algoritmo más usado Hierarchical clustering.

-Distribution based clustering, en este tipo de análisis a diferencia de los otros cada cluster pertenece a una distribución normal, además los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. El algoritmo más usado en este tipo de análisis es Gaussian mixture models.

-Density based clustering, aquí los clusters son definidos por áreas de concentración, y este cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como irregular a las áreas esparcidas entre clusters.

Regresión

La regresión es una técnica de la categoría predictiva. La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Existen dos tipos de regresiones, regresión lineal simple y regresión lineal múltiple.

-Regresión lineal simple: el análisis de regresión sólo se trata de una variable regresora. La regresión lineal simple tiene como modelo: $y = \beta_o + \beta_1 x + e$. La estimación de y debe ser una recta que proporcione un buen ajuste a los datos observados.

-Regresión lineal 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 bajo el modelo: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$.

También se puede estimar por mínimos cuadrados, donde el estimador es B=(X'X)^-1 X'y, siempre y cuando exista matriz inversa (X'X)^-1.

Las principales aplicaciones de esta técnica son:

*Medicina

*Informática

*Estadística

*Industria

*Comportamiento humano.

Entre muchas otras.

Clasificación

Es la técnica de minería de datos más comúnmente aplicada, que organiza un conjunto de atributos por clase dependiendo de sus características.

Existen diferentes técnicas de clasificación, las cuales son:

-Regla de Bayes, la cual consiste en tener una hipótesis sustentada para una evidencia.

-Redes neuronales, las cuales trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse. Internamente pueden verse como una gráfica dirigida. Estas redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

-Árbol de decisión, la cual consiste en una serie de condiciones que están organizadas de forma jerárquica tipo árbol. Esta técnica es útil para resolver problemas que muestran datos categóricos y datos numéricos, así como para la clasificación y regresión.

Existen varios problemas con inducción de reglas, los cuales son:

-Las reglas no necesariamente forman un árbol.

-Pueden no cubrir todas las posibilidades.

-Pueden entrar en conflicto.

Patrones secuenciales

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

El principal objetivo de estos patrones es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos y con ello buscar asociaciones de la forma “si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n”.

Unas de las principales características de estas reglas son:

-El orden importa

-La longitud de una secuencia es la cantidad de ítems

-Las secuencias frecuentes son las subsecuencias de una secuencia que tienen un soporte mínimo.

Entre otras características.

Para resolver problemas se puede utilizar el agrupamiento de estos patrones secuenciales, la clasificación con los datos obtenidos de dichos patrones y las reglas de asociación con los mismos datos.

Esta técnica es muy usada en diferentes ramas como lo son la medicina, biología, web, deportes, análisis de mercado, aplicaciones financieras, entre otros.

Existen varios métodos en los que se puede representar esta técnica, por ejemplo GSP, SPADE, ISM, ISE, FreeSpan, entre muchos otros

Outliers

Datos atípicos

“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”. Es decir, es la detección de datos raros o comportamientos inusuales en un grupo de datos.

Se pueden aplicar en diferentes lugares de nuestra vida cotidiana, por ejemplo:

-Telecomunicaciones, aseguramiento de los ingresos.

-Detección de fraudes financieros.

-Seguridad y detección de fallas.

Los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de las muestras establecidas.

Los valores atípicos son en ocasiones una cuestión subjetiva, y existen numerosos métodos para clasificarlos. El método más impartido académicamente por su sencillez y resultados es el test de Tukey. En un diagrama de caja se considera un valor atípico el que se encuentra 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo).