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, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta.

Nos ayudan a encontrar las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos, así como medir la fuerza e importancia de estas combinaciones.

Cuenta con muchas aplicaciones diferentes como lo son definir patrones de navegación dentro de la tienda, crear promociones de pares de productos, da un soporte para la toma de decisiones es un análisis de información de ventas, da una idea para la distribución de mercancías en tiendas se crea una segmentación de clientes con base en patrones de compra.

Existen métricas que nos ayudan a interpretar estas reglas como el soporte el cual es el número de veces o la frecuencia (relativa) con que A y B (ítems) aparecen juntos en una base de datos, la confianza se daría con el cociente del soporte de la regla y el soporte del antecedente solamente dando información sobre la relación entre antecedente y consecuente, por ultimo esta el lift el cual es el cociente del soporte de A dado B entre el soporte de A por el soporte de B, dándonos información sobre la fuerza de relación que existe entre estos ítems.

Predicción

Existen diferentes métodos para la predicción como lo es 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, para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Los árboles se pueden clasificar en dos tipos que son los árboles de regresión en los cuales la variable respuesta y es cuantitativa y los árboles de clasificación en los cuales la variable respuesta y es cualitativa.

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos el primer nodo produce la primera división en función de la variable más importante, los nodos internos o intermedios vuelven a dividir el conjunto de datos en función de las variables y los nodos terminales u hojas se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.

Entre los tipos de árbol esta el 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, de esta forma el espacio de las covariables es divido en

hiperrectángulos y todas las observaciones que queden dentro de un hiperrectángulo tendrán el mismo valor grupo estimado.

Un árbol de regresión consiste en hacer preguntas de tipo $¿xk \leq c?$ para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiperrectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado y.

Otro método es el random forest el cual 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. Para asegurarnos que los árboles sean distintos, lo que hacemos es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging.

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.

Entre sus usos esta la investigación de mercado, identificar comunidades, prevención de crimen y el procesamiento de imágenes.

Existen varios tipos básicos de análisis como es el Centroid Based Clustering en donde 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 y el algoritmo más usado de este tipo es el de K-medias.

En el Connectivity Based Clustering los clusters se definen agrupando a los datos más similares o cercanos (los puntos más cercanos están más relacionados que otros puntos más lejanos), la característica principal es que un cluster contiene a otros clusters (representan una jerarquía), un algoritmo usado de este tipo es Hierarchical clustering.

En el Distribution Based Clustering 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.

En el 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 de datos

Esto 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.

Formas básicas para representar datos son gráficas: barras, líneas, columnas, puntos, "tree maps", tarta, semi-tarta. Mapas: burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drilldown) y Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.

El 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.

También existen las infografías que no están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos; es decir, las infografías se utilizan para contar "historias". Esta narrativa no se construye a través de texto, sino mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

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.

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

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: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$.

Cuenta con una variedad de aplicaciones en diferentes campos con muchas posibilidades como lo puede ser en medicina, informática, estadística, comportamiento humano y la industria.

Método Clasificación

La clasificación es la técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características.

Se utiliza de manera que se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras. Existen diferentes técnicas para esto como los son:

Clasificación por inducción de árbol de decisión, Clasificación Bayesiana , Redes neuronales, Support Vector Machines (SVM), Clasificación basada en asociaciones

En la tecina de redes neuronales trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse, se usan en clasificación, agrupamiento, regresión. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida. Internamente pueden verse como una gráfica dirigida.

En el árbol de decisión son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Son útiles para problemas que mezclen datos categóricos y numéricos, útiles en clasificación, agrupamiento, regresión. Problemas con la inducción de reglas: las reglas no necesariamente forman un árbol, las reglas pueden no cubrir todas las posibilidades, las reglas pueden entrar en conflicto.

En la regla de bayes este vincula la probabilidad de A dado B con la probabilidad de B dado A, dándonos diversos datos acerca de nuestra base de datos para tomar decisiones.

Patrones secuenciales

Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, es una clase especial de dependencia en las que el orden de acontecimientos es considerado, el patrón secuencial describe el modelo de compras que hace un cliente particularmente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo, son eventos que se enlazan con el paso del tiempo.

Se trata de 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”, 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 de asociación secuenciales, reglas que expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos en el tiempo.

En la resolución de problemas se da la agrupación de patrones secuenciales la cual se define como la tarea de separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a los objetivos de otros grupos. La clasificación con datos secuenciales que expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos)

en el tiempo. Las reglas de asociación con datos secuenciales que se presenta cuando los datos contiguos presentan algún tipo de relación.

Outliers

Es la minería de datos anómalos, problema de la detección de datos raros o comportamientos inusuales en los datos. Es 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.

Un valor más extremo (outlier) es un valor en un conjunto de datos que es muy diferente de los otros valores. Esto es, los outliers son valores excepcionalmente lejanos del centro.

Se pueden en diversos campos con mucho potencial como el aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros, seguridad y la detección de fallas en algún sistema.

Se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos.

En la mayoría de los casos, los outliers tienen influencia en la media, pero no en la mediana , o la moda . Por lo tanto, los outliers son importantes en su efecto en la media, graficando los datos en una recta numérica como una gráfica de puntos, nos ayuda a identificar a los outliers.

El método más impartido académicamente por su sencillez y resultados es el test de Tukey, que toma como referencia la diferencia entre el primer cuartil y el tercer cuartil o rango intercuartílico. 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).