Resumen general.

Reglas de asociación:

Es una técnica utilizada en la inteligencia artificial, utilizada en el data meaning. Donde el data menaning es para encontrar alguna tendencia o patrón en grandes bases de datos.

Conjunto de elementos:

Es una colección de 1 o mas artículos, por ejemplo, (leche, pan, mermelada). K-itemset, un conjunto de elementos que contiene k elementos.

Recuento de soporte:

Frecuencia de ocurrencia de un itemset.

Confianza:

Mide que tan frecuentes ítems en Y aparecen en X.

El ojetivo de las reglas de asociación es que dado un conjunto T, encontrar todas las reglasteniendo:

- Umbral mínimo de soporte.
- Umbral mínimo de confianza.

Reglas de asociación de la minería en 2 pasos

Generación de elementos frecuentes:

Generar todos los conjuntos de elementos posibles.

Principio a priori:

Si un conjunto de elementos es frecuente, también los subconjuntos deben de serlo.

El algoritmo consta de 2 etapas.

- Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado limite.
- Convertir esos itemsets frecuentes en reglas de asociación.

Detección de outliers (Valores atípicos):

Su objetivo es, encontrar patrones que ven un resumen de las relaciones ocultas dentro de los datos (estudia el comportamiento de valores extremos que difieren del patron general de una muestra).

- ¿Qué es un valor atipico?

- Son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos.

- ¿Por qué son ocasionados?

- Errores de entrada de datos y procedimiento.
- Acontecimientos extraordinarios.
- Valores extremos y/o faltantes.
- Causa no conocida.

Como identificar un dato atípico:

Cuando en una gráfica tienes una cierta conformación de datos que se comportan de una manera semejante y se ve un dato que sobresale de los demas.

¿Cómo se calculan los valores atípicos?

Existen diferentes técnicas para detectarlos (métodos univariantes y métodos multivariantes).

Los métodos **univariantes** se centran en el análisis de una única característica o cualidad de un conjunto de datos.

Los métodos **multivariantes** son observaciones que se consideran extrañas por el valor ue toman en el conjunto de variables.

Técnica para la detección de valores atípicos:

- Prueba de Grubbs
- Prueba de Dixon.
- Prueba de Tukey.
- Análisis de valores.
- Regresión simple.

¿Qué hacer si se detecta un valor atipico?

Se puede eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable.

Si no se debe a un error, eliminarlo o sustituirlo puede 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 outliers.

- Detección de fraudes financieros.
- Nutrición y salud.
- Negocios.

Regresión:

¿Qué es una regresión?

Es un modelo matemático para determinar el grado de dependencia entre una o mas variables, es decir, conocer si existe relación entre ellas.

Existen 2 tipos de regresión:

1.- Regresión lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.

2.- Regresión lineal múltiple: cuando 2 o mas variables independientes influyen sobre una variable dependiente.

Regresión en la minería de datos:

Se encuentra en la categoría predictivo con el objetivo de analizar los datos de un conjunto y en base a esto, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

La categoría predictiva no es 100% precisa ni confiable, lo que se busca es ver cómo se comporta el modelo con nuevos conjuntos de datos.

Análisis de regresión:

Permite examinar la relación entre 2 o mas variables e identificar cuales son las que tienen mas impacto en un tema de interés.

También a su vez 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.

Clustering:

¿Qué es el clustering?

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.

Cluster:

Un **cluster** es una colección de objetos de datos. Similares entre sí dentro del mismo grupo.

Algoritmos de cluster

k-medias:

Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Pasos:

- Se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones.
- Se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:
    - Determina las coordenadas del centroide.
    - Determina la distancia de cada objeto a los centroides.
    - Agrupa los objetos basados en la menor distancia.
- Finalmente quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

Existe un algoritmo mas avanzado que se llama x-means, este algoritmo 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, dando de esta manera más flexibilidad al usuario.

Predicción

Es una técnica que se utilizara para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento.

- Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo
- Los valores son continuos
- Las predicciones son a menudo sobre el futuro

Variables independientes: atributos ya conocidos

Variables de respuesta: lo que queremos saber

Aplicaciones.

- Revisar los historiales crediticios de los consumidores y las comprar pasadas para predecir si serán un riesgo crediticio en el futuro
- Predecir si va a llover en función de la humedad actual
- Predecir el precio de venta de una propiedad
- Predecir la puntuación de cualquier equipo durante un partido de futbol

Técnicas

- **Modelos estadísticos simples como regresión**
- **Regresión lineal:** Determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra.
- **Regresión lineal multivariante:** Pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella

- **Estadísticas no lineales como series de potencias**
    - **Regresión no lineal:** Se denomina regresión no lineal porque las relaciones entre los parámetros dependientes e independientes no son lineales
    - **Regresión no lineal mujltivariante.**

- **Redes neuronales, RBF, etc.**
- Utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión, consiste en 3 capas: entrada, oculta y salida

Todo basado en ajustar una curva a través de los datos, es decir, encontrar una relación entre los predictores y los pronosticados

Patrones Secuenciales

Características:

- Ordenada.
- Encontrar patrones secuenciales.
- El tamaño es una cantidad de elementos.
- La longitud es la cantidad de ítems.
- El soporte 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.

Ventajas:

- Flexibilidad y eficiencia.

Desventajas:

- Utilización y sesgado por los primeros patrones.

Tipos de datos:

- ADN y proteínas.
- Recorrido de clientes en un supermercado.
- Registro de accesos a una página web.

Aplicaciones (agrupamiento de patrones secuenciales)

Medicina (predecir si un compuesto químico causa cáncer).

Análisis de mercado (comportamiento de compras).

Aplicaciones (clasificación con datos secuenciales)

Web (reconocimiento de spam de un corro eléctrico).

Análisis de secuencias

Base de datos, secuencia, elemento y evento.

Visualización de datos

Es la presentación de información en formato ilustrado o grafico.

Es importante conocer que existen diferentes tipos de visualización de datos ya que uno de los grandes retos que enfrentan los usuarios de empresas es que tipo de elementos visual se debe utilizar para representar la información de la mejor forma.

Ejemplos de visualizaciones de datos:

Los gráficos presentan de manera sencilla la información

Los mapas son otro tipo de información de visualización de datos

Infográficas son una colección de imágenes, texto y gráficos

Los cuadros de mando, muestra los indicadores de los negocios

Los softwares son una técnica muy fácil de utilizar como R o Excel

Aplicaciones:

- Comprender información con rapidez
- Identificar relaciones y patrones
- Identificar tendencias emergentes

Clasificación

Tareas predictivas:

predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos

- regresión
- predicción
- patrones de secuencia
- clasificación

La clasificación ordena por clases tomando en cuenta las características de los elementos que contiene:

- Empareja o asocia datos a grupos predefinidos.
- Encuentra modelos que describen y distinguen clases o conceptos para futuras predicciones.
- Probablemente se considera la tarea más familiar y más popular de la minería de datos.

algunos métodos:

- Análisis discriminante (clasificar por los colores)
- Arboles de decisión (traza un camino)
- Reglas de clasificación (buscan términos no clasificados, y pueden transcribir datos)
- Redes neuronales artificiales (puede a ver múltiples caminos)