UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MINERÍA DE DATOS

RESUMEN
TÉCNICAS DE MINERÍA DE DATOS.

WENDY OLIVIA BAZÚA CORRALES
1887913

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.
Existen cuestiones relativas a la relacion temporal de las variables de entrada o predictores

Relación con otras tecnicas: cualquiera de las tecnicas utilizadas para la clasidicacion y estmiacion pued ser adaptada para su uso en la prediccion mediante el uso de ejempos.
Los datos historicos se utilizan para crear un modelo
Aplicaciones

- Revisar historiales crediticios de los consumidores para predecir si seran un factor de riesgo crediticio en el futuro
- Predecir el precio de venta de una propiedad
- Predecir si va a llover

Tecnicas

- Regresión
- Series de potencias
- Redes neuronales

Tipos de metodos de regresion

- Regresión lineal
- Regresion lineal multivariante
- Regresion no lineal univariable y multivariable

Patrones secuenciales

Características

- El orden importa
- Tiene como objetivo encontrar patrones secuenciales
- El tamano de una secuencia es su cantidad de elementos
- La longitud de la secuencia es la cantidad de items
- El soporte de una secuencia es el porcentaje de …..

Ventajas: Flexibilidad y eficiencia
Desventajas: Utilización (los valores para los parámetros son dificiles de establecer) y sesgado por los primeros patrones.
Tipos de datos:

- ADN y proteínas
- Recorrido de clientes en un supermercado
- Registros de accesos a una pagina web.

Aplicaciones:

- Agrupamiento de patrones secuenciales

    - Medicina: Predecir si un compuesto causa cancer.
    - Analisis de mercado: Comportamiento de compras.
- Clasificacion con datos secuenciales
    - Web: reconocimineto de spam en un correo electrónico

| s | es el numero de elementos de una secuencia.
Una k-secuencia es una secuencia con k-eventos.
Una subsecuencia es una secuencia que esta dentro de otra. Pero cumpliendo ciertas normas:El item del elemento
de la subsecuencia, tiene que estar dentro del evento i de la secuencia.
Analisis de secuencias:

- Base de datos
- Secuencia
- Elemento (transacción)
- Evento (ítem)

Método GSP : Generalized Sequantial Pattern.

Visualización de datos

Nos sirve para representar graficamente los elementos mas inmportantes de nuestra base de datos.
La visualizacion de datos es la presentacion de informacion en un formato ilustrado o grafico.
Nos permite ver y comprender tendencias, valores atipicos, etc.
Tipos:

- Gráficos
    - Circulares
    - Lineas
    - Columnas
    - Barras
    - Burbujar
    - Dispersion
    - Tipo arbol
- Mapas
- Infografías: Colección de imágenes, graficos y textos simples que resummen un tema para que se pueda comprender facilmente.
- Cuadros de Mando (dashboards): Permite saber en todo el momento el estado de los indicadores del negocio.

Aplicaciones: Comprender la informacion con rapidez, identificar relaciones y patrones, identificar tendencias emergentes, comunicar la historia a otras personas.

Importancia: La visualizacion es una herramienta cada vez mas importante para darle sentido a los billones de filas de datos que se generan cada dia.

Clasificación

Es una tarea predictiva. Es una técnica de la minería de datos que ** en el ordenamiento de datos ***
Empareja o asocia datos a grupos predefinidos, encuentra modelos que describen y distinguen las clases o conceptos para futuras predicciones y se considera el método más sencillo.
Metodos:

- Analisis descriminante
- Arboles de decision
- Reglas de clasificacion
- Redes neuronales artificiales.

Características de los metodos:

- Precision en la prediccion
- Eficiencia
- Robustez
- Escalabilidad
- Interpretabilidad

Reglas de asociación

Busca patrones frecuentes, asociaciones, correlaciones o estructuras entre conjuntos de elementos u objetos en bases de datos.
Aplicaciones

- Analisis de datos de la banca
- Cross- Marketing
- Diseño de catálogos

Soporte: Fraccion de transacciones que contiene un itemset.
Conjunto de elementos frecuentes: Conjunto de elementos cuyo soporte es mayor o igual que un umbral de minimo.
Conjunto de elementos: colección de uno o mas artículos.
Recuento de soporte: frecuencia de ocurrencia de un itemset.
Confianza(c): Mide que tan frencuente items en Y aparecen en transacciones que contienen X. Como una probabilidad condicional.

Su objetivo es dado un conjunto de transaccion T, encontrar todas la reglas teniendo un umbral minimo de soporte y de confianza.

Enfoque de dos pasos:

- Generación de elementos frecuentes.
- Generación de reglas.

Cada conjunto de elementos en la red es un conjunto de elementos frecuente candidato.
Principio “apriori” se utiliza para reducir el número de candidatos o conjuntos. Fue uno de los primeros algoritmos creados para la búsqueda de reglas de asociación. Se identifican

los itemsets que ocurren con una frecuencia por encima del límitee y se convierten en reglas de asociacion.

Outliers:

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

Un valor atípico son observaciones cuyos valores difieren mucho a las otras observaciones del mismo grupo de datos. Son causados por errores al capturar los datos, acontecimientos extraordinarios, valores extremos o casusas desconocidas.

Estos datos distorsionan los resultados de los análisis y por eso hay que identificarlos y saber que hacer con ellos.

Se pueden dividir en dos categorias distintas:

- Métodos univariantes: analizan una única característica de los datos.
- Métodos multivariantes

Algunas de las tecnicas para la deteccion de valores atipicos son: la prueba de Grubbs, la de Dixon, la de Turkey, el análisis de valores atípicos de Mahalahobis y la regresión simple.

Algunas de sus aplicaciones son la detección de fraudes financieros, tecnología informática y telecomunicaciones, nutrición y salud y negocios.

Distintos significados:

- Error
- Límites: valores que se escapan de lo normal pero decidimos mantenerlos.
- Punto de interés

Regresión

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, o sea, conocer si hay una relacion entre ellas.

Hay dos tipos principales:

- Regresión lineal: Una variable independiente ejerce influencia sobre otra variable dependiente.
- Regresión lineal múltiple: dos o más variables independientes.

Se encuentra dentro de la categoría predictiva.

El análisis de regresion permite examinar la relación entre dos o más variables e identificar cuales son las que tienen mayor impacto en un tema de interés.

Nos permite explicar un fenómeno y predecir cosas acerca del futuro, esto nos ayuda a tomar decisiones y tener mejores resultados.

Mientras mayor sea R^2 mejor se ajusta el modelo.

Se utiliza una prueba de significancia para ver que tanto se ajusta nuestro modelo a la muestra.

El coeficiente de correlación nos dice que tanto se relacionan nuestras variables entre si.

Clustering

Consiste en la división de los datos en grupos de objetos similares. Usando algoritmos metemáticos que se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto.

Un cluster es una colección de objetos de datos. Similares entre si dentro del mismo grupo. Tienen que ser parecidos pero no exactamente iguales.

El análisis de cluster trata de entender la estructura de un conjunto de datos.

Algunas de sus aplicaciones son el estudio de terremotos, la planificación de alguna ciudad, el marketing, el uso de suelo y aseguradoras.

Métodos de agrupación:

- Asignación jerárquica
- Datos numéricos y/o símbolos
- Determinística y probabilística
- Exclusivo vs. Superpuesto
- Jerárquico vs plano
- De arriba abajo y de abajo a arriba.

Clusters: conjunto de datos.

Simple K-Means: Este algoritmo debe definir el número de clusters que se desean obtener.

X-Means: Es una variante mejorada del K-Means. Su ventaja esta en haber solucionado una de las mayores deficiencias presentadas en K-Means., el hecho de tener que seleccionar el número de clusters.

Cobweb: Se forma un árbol de clasificación donde las hojas representan segmentos y el nodo raíz engloba por completo el conjunto de datos.

EM: parte de los Finite Mixture Models. Intenta encontrar la distribución de probabilidad de los conjuntos. Finalmente se obtienen un conjunto de clusters que agrupan el conjunto original.