UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Facultad de ciencias Físico matemáticas

# Minería de Datos

## Resúmenes

Eliezer Gamaliel Castillo Alcantar

Matricula: 1684521

**02 de octubre del 2020**

## Reglas de asociación

Esta técnica funciona bien dado que sirve para encontrar patrones puesto que asocia los datos de forma que puedes entender que tan frecuentes son y generas los grupos óptimos.

Conceptos en las reglas de asociación:

- Suporte: Fracción de transacciones que contienen un itemset.
- Confianza: mide que tan frecuente aparecen ítems en y aparecen transacciones que contienen X.
- Conjunto de elementos: una colección de uno o más artículos.
- Recuento de soporte: frecuencia de ocurrencia de un item-set

Aplicaciones:

- Análisis de datos de la banca: brinda un mejor servicio hacia los clientes según a los productos financieros.
- Cross-Marketing: se utiliza para mejorar las ventas en una empresa ya que relaciona los productos que normalmente les interesa a los clientes.

Objetivo:

El objetivo de la minería de reglas de asociación es encontrar todas las reglas o patrones de una base de datos teniendo en cuenta lo siguiente:

- Umbral mínimo de soporte
- Umbral mínimo de confianza

Tipos de reglas de asociación:

- Enfoque en 2 pasos: se dice de dos pasos por que primero se generan elementos frecuentes y después se hacen reglas de asociación.
- Principio “a priori”: si un conjunto de elementos es frecuente, entonces todos sus subconjuntos son frecuentes.

## Clasificación

Es una técnica de la minería de datos utilizada para el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contienen una base de datos.

Características:

- Eficiencia
- Robustez
- Precisión en la precisión

- Interpretabilidad

Datos de la clasificación

- Empareja todos los grupos predefinidos, junta dependiendo del patrón que siguen los datos
- Encuentra modelos que describen clases o conceptos para futuras predicciones
- La clasificación se considera como la técnica más sencilla.

Métodos utilizados

- Análisis discriminante: se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos.
- Arboles de decisión: este a través de una representación esquemática facilita la toma de decisiones.
- Reglas de clasificación: busca términos no clasificados de forma periódica, para posteriormente encontrar una coincidencia se agrega a los datos de la clasificación.

## Outliers

La técnica de detención de outliers es una técnica de minería que estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

Valores atípicos

Son valores anormales comparados con el resto de datos en la base, no tienen su mismo comportamiento.

Técnicas para identificar datos atípicos

- Métodos univariantes de detención de outliers
- Métodos munivariantes de detención de outliers

Técnicas para la detención de valores atípicos

- Regresión simple
- Prueba de grubbs
- Prueba de Dixon
- Prueba de tuckey
- Análisis de valores y atípicos de mahalanobis

Una ves detectados los valores atípicos, se pueden eliminar o sustituir, aunque lo mejor sería quitarles valor a los datos atípicos con técnicas robustas.

Aplicaciones

- Detención de fraudes financieros

- Tecnología informática y telecomunicaciones
- Nutrición y salud

Distintos significados

- Error: error a la carga de datos
- Punto de interés: casos anómalos que detectamos, por ejemplo, alguna enfermedad
- Limites: valores que son mas grandes o menores a la media

## Predicción

Es una técnica que suele usar para proyectar los tipos de datos, para predecir el resultado de un evento. Casi siempre el simple hecho de reconocer y comprender las tendencias historias es suficiente para trazar una predicción un poco precisa de lo que podría ocurrir en el futuro.

Aplicaciones

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Predecir si va a llover mediante como ha llovido en años anteriores.
- Predecir eventos deportivos mediante resultados pasados.
- Predecir precios de propiedades basados en estudios del mercado.

Técnicas:

- Regresión simple
- Estadística no lineal
- Redes neuronales

Tipos de métodos de regresión:

- Regresión lineal
- Regresión lineal multivariante
- Regresión no lineal
- Regresión no lineal multivariante

## Regresión lineal

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir conocer si existe relación entre ellas. Con esto podemos ajustar las variables a un modelo y así poder hacer una predicción de lo que puede pasar en un futuro.

Tipos de regresión:

- Regresión lineal: cuando una variable independiente ejerce influencia sobre una variable dependiente
- Regresión lineal múltiple: Cuando dos o más variables independientes influyen sobre una variable dependiente.

El objetivo de la regresión es analizar los datos del conjunto y predecir lo que puede ocurrir en el futuro sin embargo no es muy preciso. Al mismo tiempo nos ayuda a visualizar con vario tipos de gráficos las relaciones de las variables utilizadas. Este procedimiento nos va dando una serie de factores los cuales son los siguientes:

La R representa el coeficiente de correlación y significa el nivel de asociación entre las variables.

La R^2 representa el coeficiente de determinación, indica porcentualmente el cambio de la dependiente respecto a la independiente.

Se necesita saber si la regresión es significativa para tener idea si existe estas relaciones entre cada uno.

## Patrones secuenciales

Los patrones secuenciales se basan en el análisis de una secuencia y con una base de datos ordenada por tiempo, o espacio, se busca encontrar un patrón que nos permita predecir el comportamiento con las características de tiempo o espacio.

Aplicaciones:

- Medicina: predecir si un compuesto químico causa cancer o no
- Análisis de mercado: comportamiento de compras de los consumidores en tiendas.

Ventajas:

- Es muy eficiente
- Flexible

Desventajas

- Difícil en ciertas ocasiones
- Sesgado por las primeras observaciones

Tipos de datos:

- ADN y proteínas
- Registros de usuarios de cierta página web
- Recorrido de clientes en un super mercado

## Clustering

Se trata del proceso de dividir los datos en grupos de objetos similares. Las técnicas de clustering son las que utilizando algoritmos matemáticos se encargan en agrupar objetos. Usando la información de las variables que pertenecen a cada objeto se mide la similitud o parecido entre los mismos.

Aplicaciones

- Estudio de terremotos
- Planificación de una ciudad
- Aseguradoras

Métodos de agrupación

- Asignación jerárquica frente a punto
- Determinística vs probabilística
- Datos numéricos y/o simbólicos
- Jerárquico vs plano
- De arriba a abajo y abajo a arriba

Algoritmos de clustering

- Simple K-Means: este tipo de algoritmo define el número de clusters que se desean obtener.
- X-Means: es una mejora del k-Means que es tener que seleccionar a priori el número de clustersque se desean obtener, a X-Means se le define un límite inferior k-min y un límite superior k-max y este algoritmo es capaz de obtener ese rango el número óptimo de clusters, dando de esta manera más flexibilidad.
- Cobweb: se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia.

## Visualización de datos

Representa gráficamente los elementos más importantes de una base de datos ya que se utilizan elementos visuales como cuadros, gráficos o mapas los cuales proporciona una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos.

Tipos de visualización de datos

- Gráficos: más común, hojas de cálculo como diagramas de árbol, gráficos de dispersión, etc.
- Mapas: visualización de datos en mapas para poder visualizar sucesos en tiempo real como Google maps.
- Cuadros de mando: cuando de mando es una herramienta de gestión empresarial imprescindible e incluye indicadores.
- Infografías: conjunto de imágenes, gráficos, texto siempre que resume un tema para que se pueda entender fácilmente.

Aplicaciones:

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