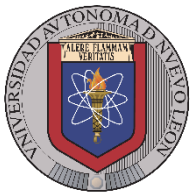

Universidad Autónoma de Nuevo León
Facultad De Ciencias Físico Matemática

Etapa 1 Resúmenes de exposiciones

Minería de Datos

Maestra:
Mayra Cristina Berrones Reyes

Alumno:
Francisco García Sánchez Armáss

Matricula:
1816358

Grupo: 003
Aula: AV12

San Nicolás de los Garza, Nuevo León, México a 26 de octubre del 2020

Resúmenes

Tema 1: Reglas de Asociación

Este tema me agrado mucho, ya que me toco exponerlo y me llamo mucho la atención que sirve para la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras entre conjuntos de elementos u objetos en bases de datos de transacciones y que además se puede ver en distintas aplicaciones como lo son: Análisis de datos de la banca, Cross-marketing, Diseño de catálogos (el más común). Es importante saber cuándo utilizar un umbral mínimo de soporte de todos los ítems que se nos presenten e igual un umbral mínimo de confianza. Para dicho tema quisimos adentrarnos en el algoritmo “a priori” el cual nos ayuda a resolver alguna problemática que tengamos en algún problema de regla de asociación.

Después nos habla sobre la reducción de candidatos, el cuál si tenemos un conjunto de elementos frecuentes, entonces todos sus subconjuntos también deben de ser frecuentes, por lo que se sigue el algoritmo a priori: X, Y: (X C Y) => s(X) >= s(Y).

La generación de reglas me llamo la atención ya que es algo fácil de entender y es necesario para cuando nosotros tengamos una base de datos en este caso supongamos un catálogo sobre los alimentos más vendidos en diversos supermercados, se pueden formar diversas transacciones de las cuales deben de cumplir con el umbral mínimo de soporte, una vez que se haya cumplido lo anterior se pasa a la generación de reglas, que no es más que nada, que ajustar los itemset que cumplieron con la condición del soporte dividirlos entre cada uno de los itemset de manera individual.

El ejemplo que se nos mostró se pudo ver claramente como teniendo un conjunto de transacción que en este caso fueron 7 transacciones divididas en todas la posibles combinaciones sin repetirse, el cual el problema nos define que debe de cumplir con que el soporte sea igual o superior a 3/7, para poder hacer esto debemos de realizar el primer itemset que es para k=1, si todos pasan vamos con k=2 y así nos vamos sucesivamente, aunque para este caso en k=3 ninguno cumple con la condición de k=3 entonces hasta hay le paramos, ya una vez definido nuestros posibles itemset frecuentes pasamos a la “generación de reglas” .

Finalmente me gustó este tema para cuando más adelante desarrolle o en este caso tome una base de datos ya sea de una banca internacional o nacional, en la que pueda ver primeramente todas las posibles transacciones que pueda sacar, después definir un soporte mínimo, para último hacer la generación de reglas y definir los mejores itemset.

Tema 2: **Clasificación**

La clasificación se encarga de predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos, lo cuales "la clasificación" esta dentro de 4 grandes subramas de predictivo, los cuales los otros 3 son.

- Predicción
- Regresión
- Patrones secuenciales

La clasificación es una técnica de la minería de datos, la cual es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene, algunos métodos que se pueden utilizar son:

- Análisis discriminante: método para encontrar una combinación lineal de rasgos.
- Reglas de clasificación: buscan términos no clasificados de forma periódica.
- Arboles de decisión: método analítico que a través de una representación esquemática facilita la toma de decisiones.
- Redes neuronales artificiales: es un modelo de unidades conectadas para transmitir señales.

Las características finales que podemos ver en clasificación son:

1. Precisión en la predicción
2. Eficiencia
3. Robustez
4. Escalabilidad
5. Interpretabilidad

Tema 3: **Detección de outliers**

Este tema lo considero importante ya que se encarga de estudiar el comportamiento de valores extremos que difieren del patrón general de una muestra, me llamo la atención que se tocara que es un valor atípico el cual pues vienen siendo errores de entrada de datos, acontecimientos extraordinarios, causas desconocidas, se nos mencionan diversas técnicas para la detección de valores atípicos los cuales considero interesantes ya que se nos presentan 6 distintas técnicas, las cuales son: Prueba de Grubbs, Prueba de Dixon, Prueba de Tukey, Análisis de valores, Regresión simple (Mínimo de cuadrados).

Considero que es importante entender esto, ya que una vez que maneje una base de datos debo de ver que datos son atípicos, para ello una vez detectados puedo hacer lo siguiente que es eliminar o sustituir si se corrobora que los datos atípicos se debe a un error de captura o medición de la variable, si no pues puede pasar que cometí un error como: Introduce un sesgo, Disminuir el tamaño muestral, Puede afectar a la distribución y a las varianzas, La mejor opción quitarle peso a esas observaciones. Las distintas aplicaciones

que puedo encontrar al realizar la detección de outliers son: Detección de fraudes financieros que es muy común en diversos sitios web, tecnología informática y telecomunicaciones, nutrición y salud, negocios.

En el ejercicio práctico que presentaron mis compañeros pude ver una base de datos de desempleo por cada mes desde 1948 a 1978, en el cual debemos de encontrar datos atípicos, en la materia de "métodos estadísticos" trabajamos con un el análisis de regresión el cual con ayuda de "Minitab" nos mostraban una gráfica sobre como los datos se dispersaban, sin embargo deben de seguir o la gran mayoría de los datos deben de estar juntos y aquellos datos que estaban alejados serian nuestros datos atípicos.

Finalmente me gusto retomar conceptos ya vistos, en materias pasadas ya que logro comprender como es que visto desde la perspectiva de la probabilidad o estadística aplicarla en la programación

Tema 4: Patrones secuenciales

En minería de datos secuenciales, los cuales es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el tiempo. Reglas de asociación secuencial representan patrones en distintos lapsos del tiempo. Las características de los patrones secuenciales me llamo la atención ya que no importa el orden, su objetivo es encontrar patrones secuenciales, el tamaño de una secuencia es su cantidad de elementos, la longitud de la secuencia es la cantidad de ítems, etc. Sus ventajes son la flexibilidad y eficiencia ya que no necesitamos correr mucho los datos, solo ocupamos correrlos una vez, las desventajas que se pueden enfrentar es el sesgado de los primeros valores los cuales pueden ser de cómo acomodemos los valores. Los tipos de datos de patrones secuenciales, me llamo la atención el ADN y proteínas, recorrido de clientes en un supermercado, etc. Las aplicaciones que puede darnos es la medicina: predecir un compuesto químico causa cáncer, análisis de mercado: comportamiento de compras y la web: en el reconocimiento de spam de un correo electrónico.

Los patrones que se suelen ver para entender al hacer ejercicios es la siguiente:

- lsl es el número de elementos de una secuencia.
- Una k-secuencia es una secuencia de k eventos
- Una subsecuencia es una secuencia que está dentro de otra. Pero se deben de cumplir ciertas normas el cual es el "orden".

Análisis de secuencias:

1. Base de datos: clientes que han ido al supermercado.
2. Secuencia: pasillo que sigue el cliente
3. Elementos de transacción: ver por qué pasillos pasaron los clientes
4. Item: los productos que obtuvo la persona al ir al supermercado

Me agrado que se empleara un algoritmo, cual fue “Algoritmo GSP”, nos menciona que, al tener diversas secuencias, debemos de juntar en diversos candidatos que en este caso es de 4 y un umbral mínimo de 3, para obtener la candidata de 4 secuencia final (Menciona como se va a comportar el patrón).

Tema 5: **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. Tiene una relación con otras técnicas ya que 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. Los datos históricos se utilizan para construir un modelo que explica el comportamiento observado en los datos. Cuando este modelo se aplica a nuevas entradas de datos, el resultado es una predicción del comportamiento futuro de los mismos.

Aplicaciones:

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

Técnicas de predicción:

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

Redes neuronales: utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión. Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

Tema 6: **Regresión**

Este tema, para mí ya es conocido ya que lleve la materia de “métodos estadísticos” y me resultan conocidos diversos términos empezando con que es la regresión, el cual es un modelo matemático para determinar el grado de dependencia entre una o más variables (quiere decir que si existe una correlación entre ellas)

La regresión se divide en 2: La regresión lineal (cuando solo comparamos con 2 posibles valores, una variable independiente que ejerce influencia sobre la otra variable dependiente. La regresión múltiple (cuando tenemos dos o más variables independientes y una sola variable dependiente). Realizamos un Análisis de regresión que nos ayuda a visualizar, el comportamiento de la gráfica de probabilidad normal, vs ajustes, el

histograma, y el de orden, aunque si no usamos variables de tiempo no es necesario esta última.

Ahora nos habla sobre diversas funciones que debemos de utilizar para decir que tan bueno es nuestro ajuste, o si es un buen modelo de regresión lineal, para el ejercicio práctico que mostraron mis compañeros considero que es algo importante ya cuando trabajemos con una base de datos, como, por ejemplo:

Se nos da una base de datos de la población de México, el cual tenemos el peso y la estatura, estas dos variables nos sirven para crear una regresión lineal, ya que poseemos una variable dependiente y una variable independiente, lo cual podemos hacer una Análisis de regresión viendo cada una de las gráficas y primero asegurándonos que cumpla con el criterio de que es una regresión lineal, el cual el p-valor <alfa.

Tema 7: Visualización de datos

La visualización de datos es la representación de información en formato ilustrado o gráfico. El cual, al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos, como se nos menciono en la exposición es importante conocer diferentes tipos de visualización de datos, ya que uno de los grandes retos que enfrentan los usuarios de empresas, es que tipo de visual se debe de utilizar para representar la información de la mejor forma, lo cuales se dividen de la sig. Forma:

- **Gráficos:** es el tipo más común y conocido, los cuales pueden ser gráficos circulares, líneas, columnas, barras aisladas, agrupadas, burbujas, áreas, diagramas de dispersión y mapa de tipo árbol.
- **Mapas:** la cual la herramienta más conocida para la visualización de mapas es Google maps.
- **Infografías:** es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente.
- **Cuadros de Mando:** En el entorno empresarial, un cuadro de mando es una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos, etc. y que nos dice lo que está pasando en la empresa (idealmente en tiempo real) para poder tomar decisiones adecuadas.

Las aplicaciones que podemos entender de la visualización de datos son:

- **Comprender la información con rapidez** Mediante el uso de representaciones gráficas de información de negocios, las empresas pueden ver grandes cantidades de datos de formas claras y cohesivas y sacar conclusiones a partir de esa información.

- **Identificar relaciones y patrones**. Incluso muy grandes cantidades de datos complicados comienzan a tener sentido cuando se presentan de manera gráfica; las empresas pueden reconocer parámetros con una correlación muy estrecha.
- **Identifique tendencias emergentes**. El uso de la visualización de datos para descubrir tendencias en los negocios y en el mercado puede dar a las empresas una ventaja sobre la competencia, y eventualmente tener un impacto en la base de operación.

Tema 8: Clustering

En este último tema nos habla como cluster que es una colección de objetos de datos. Similares entre i dentro del mismo grupo. Disimilar a los objetos en otros grupos. Una vez que tenemos el "cluster" se gráfica y nos dan una serie de puntos, con estos puntos se puede realizar un análisis de cluster, el cual es dado un conjunto de puntos de datos se trata de entender su estructura, encontrando similitudes entre los datos de acuerdo con las características encontradas en los datos.

Las diversas aplicaciones que puede ofrecernos el Clustering es:

- El estudio de los terremotos: el cuál es importante ya que los epicentros del terremoto observados deben de agruparse a lo largo de fallas continentales.
- En las Aseguradoras: se identifican los grupos asegurados, ya sea por seguro de vida o de daños, los cuales dentro de esas dos categorías nace muchos más seguros.
- En Marketing: Cuando llegan diversas fechas ya sea festivas, cumpleaños, vacaciones, es importante descubrir cómo van existiendo diversos grupos en sus bases de clientes
- En Demografía: Cuando se realizan el conteo del censo, se suelen recabar mucha información que suele ser almacenada en distintas bases de datos.

Casi al final de este tema se nos mencionan 2 tipos de algoritmos que podemos utilizar para revolver el "clustering" el cual uno de ellos me gusto que es el Cobweb, el cual se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos

Finalmente, si lo considero importante, aunque suele ser muy parecido al tema pasado "Regresión" ya que ambos manejan el uso de graficas para ver el patrón que siguen los datos recolectados de una base de datos.