UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

RESUMEN

NOMBRE: FERNANDO ISAI GONZALEZ CASTILLO

GRUPO: 002

MATRICULA: 1819011

MAESTRA: MAYRA CRISTINA BERRONES REYES

MATERIA: MINERÍA DE DATOS

FECHA: 2 DE OCTUBRE DEL 2020

Clustering:

Esta técnica consiste en agrupar puntos de datos y crear particiones según sus similitudes.

Existen algunos usos del clustering: Investigación del mercado, Identificar comunidades, Prevención de crimen y Procesamiento de imágenes.

Es importante transformar nuestros datos para este método, las variables cuantitativas deben tener la misma medición, las variables binarias solo pueden tomar el valor de 0 y 1 y en las variables categóricos se puede volver variables binarias.

Existen distintos tipos de análisis de clustering:

Centroid Based Clustering: Se escogen puntos al azar y cada cluster es representado por centroides, estos clusters se basan en la distancia de los puntos hasta el centroide, se realizan varias iteraciones hasta llegar al resultado, su algoritmo es el de k-medias, en donde los centroides son k puntos aleatorios que pasan a ser centroides de cada cluster, de cada dato calculamos su distancia con la del centroide y este dato pertenece al cluster del de la distancia mínima, ya que tenemos todos nuestros cluster, obtenemos la media de cada cluster y el resultado será el nuevo centro e iteramos hasta que ya no haya cambios.

Connectivity Based Clustering: Los puntos más cercanos están más relacionados que otros más lejanos, su característica es que un cluster contiene a otros clusters y su algoritmo usado es Hierarchical clustering.

Density Based Clustering: Cada cluster pertenece a una distribución normal, los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal y su algoritmo es Gaussian mixture models.

Density Based Clustering, Son definidos por áreas de concentración, trata de conectar puntos que tengan una distancia pequeña, este cluster contiene a todos los puntos relacionados dentro de una distancia limitada.

Reglas de Asociación:

Las reglas de asociación son utilizadas para encontrar relaciones dentro de un gran conjunto de transacciones entre sus ítems. Esta técnica implica un antecedente y un consecuente, en donde ambos son ítems individuales, un ejemplo de ello es cereal como antecedente y leche como consecuente.

Esta técnica tiene distintas aplicaciones como promociones de pares de productos, distribución de mercancías en tienda y nos ayuda a tomar decisiones, segmentar a los clientes dependiendo de sus compras y definir patrones de navegación dentro de la tienda, además ayuda a acomodar el orden de los productos en una tienda como cuando vemos la mostaza, mayonesa y cátsup en un solo pasillo.

Existen distintos tipos de reglas de asociación dependiendo de su asociación: Asociación booleana, cuantitativa, unidimensional, multidimensional, de un nivel y multinivel, este último se refiere a que existen varios niveles de atracciones, es decir no es lo mismo decir computadora a computadora portátil de la marca Apple.

Para resolver esta técnica se utilizan tres métricas: Soporte, Confianza y Lift.

El soporte nos dice el número de veces o la frecuencia en la que aparecen nuestros productos A y B juntos con respecto al total de transacciones, en probabilidad sería $P(A \cap B)$ =Frecuencia en que A y B aparecen en las transacciones/Total de transacciones, si nuestro resultado es bajo significa que los productos juntos aparecieron por casualidad.

La confianza nos dice la probabilidad de que, si ya escogimos el producto A, escojamos el producto consecuente B, siendo escrita como $P(B/A) = P(A \cap B) / P(A)$, si este resultado es bajo significa que no existe relación entre el antecedente A y consecuente B.

Por último, el lift refleja el aumento de probabilidad del consecuente B cuando ya sabemos que escogimos el antecedente A, siendo escrito como $LIFT = P(A \cap B) / (P(A)*P(B)$, si nuestro resultado es igual a 1, significa que los productos fueron escogidos al azar.

Para delimitar el número de reglas de todos los productos de nuestras transacciones podemos definir un soporte, confianza y lift mínimo, con el fin de reducir el número de reglas y encontrar que productos tienen una mayor relación.

Detección de outliers:

Outliers o datos atípicos, son observaciones que están muy alejadas de las demás observaciones y podrían no estarse comportando de la misma manera que las demás observaciones de nuestra base de datos.

Existen distintas aplicaciones como el aseguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros y por último en la seguridad y detección de fallas, pero no solo se puede aplicar para las anteriores mencionadas, ya que cualquier empresa o persona que cuenta con una base de datos podría llegar a encontrar datos atípicos dentro de sus observaciones, es decir podrían estar siguiendo un patrón distinto a lo de los demás.

En la clase se presentó el ejemplo de un conjunto de salarios de una región en particular, algunas personas tendrían salarios muy similares pero aquellas personas como directivos, gerentes, etc., tendían un salario muy alto, siendo estos salarios nuestros datos atípicos.

El algoritmo para la detección de datos atípicos se llama agrupamiento espacial basado en densidad de aplicaciones con ruido o Density-based spatial clustering of applications with noise (DBSCAN), en donde se tienen dos parámetros: Épsilon, el cual es el radio o distancia máxima con el punto vecino y Puntos mínimos (Minimum Points), este es el número mínimo de puntos en el radio de Épsilon, con estos parámetros se determinan los puntos centrales, los cuales cumplen con los dos parámetros Épsilon y Mínimo de puntos, puntos fronterizos, los cuales cumplen con Épsilon pero no con el mínimo de puntos y puntos outliers, los cuales no cumplen con ninguno de los dos parámetros.

Regresión:

Dentro de la historia de la regresión, en 1805 se registró la primera regresión lineal por el método de los mínimos cuadrados por Legendre y dicho término fue introducido por Francis Galton en su libro “Natural Inheritance”.

Lo que hace esta técnica es predecir el valor de un dato basándose en los otros datos que fueron recolectados. La función de la regresión es encontrar una relación o ecuación matemática mediante el análisis de la variable dependiente (y) y las variables independientes (x’s).

Existen varios tipos de regresiones, entre ellas se encuentran la regresión lineal simple y la regresión lineal múltiple.

La regresión lineal simple solamente involucra a un regresor (variable dependiente) y una variable independiente, en la cual se tiene como modelo:

$$y = B_0 + B_1 x + e$$

En donde B_0 representa una constante, B_1 representa que tanto influye la variable “x” a la variable “y”, este puede influir de manera positiva o negativa y “e” es una variable aleatoria normalmente distribuida con media 0 y varianza como σ^2.

Mientras que la regresión lineal múltiple involucra k regresores y una variable independiente, en el cual se tiene como modelo:

$$y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + \ldots + B_k x_k + e$$

Ambas regresiones se pueden estimar por la técnica de mínimos cuadrados.

Las regresiones se pueden utilizar en la medicina, informática, comportamiento humano, industria y estadística.

Visualización:

La visualización de datos es representar de manera gráfica nuestra información y datos, utilizando elementos visuales como gráficos y mapas, proporcionando una manera más fácil de observar tendencias, valores atípicos y patrones en nuestros datos.

Esta técnica es de gran ayuda para analizar cantidades grandes de datos y para la toma de decisiones.

En los tipos de visualizaciones se encuentran:

Elementos básicos de representación de datos: es el más sencillo donde se utilizan tipos de visualizaciones básicas como las gráficas (barras, líneas, columnas, tarta), mapas (burbujas, mapas de calor, de agregación, coropletas) y tablas (dinámicas, con anidación, de transiciones).

Un cuadro de mando es una composición compleja de visualizaciones individuales que guardan una coherencia y tienen una relación entre ellas, se utilizan para análisis de conjuntos de variables y toma de decisiones.

Las infografías se utilizan en la construcción de narrativas a partir de los datos, esto quiere decir que se utilizan para contar historias. Estas narrativas se construyen mediante la disposición de información en la que las visualizaciones se combinan con símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Existen diversos estándares web que se utilizan para la evolución de aplicaciones web, las cuales son fundamentales para la creación de visualizaciones web basadas en datos, algunas de ellas son:

HTML5, la cual se utiliza en Canvas para dibujar gráficos 2D.

CSS3 la cual permite diferencial el contenido de las páginas web.

SCV, la cual es utilizado para crear gráficos 2D.

WebGL, la cual crea gráficos 3D haciendo uso de Canvas.

La visualización de datos es importante en cualquier empleo ya que se utilizan para tomar decisiones y usar elementos visuales para contar historias con los datos.

Clasificación:

La clasificación es una técnica de la minería de datos comúnmente aplicada, que se encarga de organizar o mapear un conjunto de atributos por clase según las características de los elementos dentro de estas.

Esta técnica lo que hace es estimar un modelo para futuras predicciones utilizando los datos que hemos recolectado.

Hay diversas técnicas de clasificación como la clasificación por inducción de árbol de decisión, clasificación bayesiana, redes neuronales, support vector machines (SVM) y clasificación basada en asociaciones.

La regla de Bayes nos dice que $P(H/E) = \frac{P(E/H) * P(H)}{P(E)}$ en donde P(H) representa la probabilidad del suceso H, P(E) representa la probabilidad del suceso E y P(E/H) representa la probabilidad del suceso E condicionada al suceso H.

Las redes neuronales se trabajan con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse. Estas consisten en 3 capas: de entrada, oculta y salida, además las redes neuronales se usan en la clasificación, agrupamiento y regresión.

Los árboles de decisión son una serie de condiciones las cuales aparecen organizadas de forma jerárquica, en forma de árbol, estos árboles de decisión son útiles para problemas que contengan variables categóricas o cualitativas y variables numéricas o cuantitativas, además se usan en la clasificación, agrupamiento y regresión.

Patrones secuenciales:

Es una técnica de minería de datos que se encarga de analizar los datos y encontrar subsecuencias dentro de un grupo de secuencias.

Los patrones secuenciales describen el modelo de comprar que un cliente o un grupo de personas realizan relacionando las transacciones efectuadas por ellos en el transcurso del tiempo, dichos eventos se enlazan con el paso del tiempo.

En los patrones de secuencia, se buscan asociaciones de que si sucede un evento X en el tiempo t entonces sucederá un evento Y en el tiempo t+n, su objetivo es poder describir las relaciones temporales que existen entre los valores de los atributos del conjunto.

Dentro de sus características se encuentra que: el orden importa, queremos encontrar patrones de secuencia, una secuencia es una lista ordenada por elementos de secuencia, el tamaño y longitud de una secuencia es la cantidad de elementos e ítems respectivamente. Además, el soporte es el porcentaje de secuencias que contienen en un conjunto de secuencias.

Existen diversas áreas donde podemos usar los patrones secuencias como la medicina, biología, web, análisis de mercado, distribución y comercio, deportes y aplicaciones financieras, banca de seguros y salud privada.

Las bases de datos temporales, documentales y relacionados son los distintos tipos de base de datos.

La agrupación de patrones secuenciales se encarga de separar en grupos a los datos, en donde los miembros de un grupo sean similares entre sí y sean diferentes a los objetivos de los otros grupos.

La clasificación con datos secuenciales expresa patrones de comportamientos secuenciales, en donde se dan en tiempos distintos.

La regla de asociación con datos secuenciales presenta la relación que tienen los datos contiguos.

Predicción:

Para hacer un buen modelo de predicción se necesita definir adecuadamente el problema, recopilar datos, elegir una medida de éxito y preparar los datos, además es necesario dividir nuestros datos en el 70% de entrenamiento, 15% de validación y 15% conjunto de pruebas.

Los árboles de decisión dividen los predictores juntando observaciones con valores similares para la variable dependiente. Es necesario tomar distintas decisiones para dividir nuestro espacio muestral en subregiones, después de subdivide en regiones más pequeñas hasta que en una subregión incluyan los datos de la misma clase.

Estos se pueden dividir en dos árboles: Arboles de regresión, en los que la variable dependiente es cuantitativa y también tenemos los árboles de clasificación, en los cuales la variable dependiente es cualitativa.

Los árboles de clasificación hacen preguntas xk<c para las cuantitativas y xk=nivelj para las cualitativas, por lo que el espacio queda dividido en rectángulos y todas las observaciones que queden en el mismo rectángulo tendrán el mismo valor grupo estimado.

En ellas hay dos tipos de nodos: los nodos de decisión, los cuales tienen una condición y debajo tienen más nodos y los nodos de predicción, los cuales no tienen condición ni nodos debajo de ellos.

Los árboles de regresión hacen preguntas xk<c, por lo que el espacio queda dividido en rectángulos y todas las observaciones que queden en el mismo rectángulo tendrán el mismo valor $\hat{y}$ estimado.

También tenemos los Bosques Aleatorios o Random Forest, los cuales compensa los errores de las predicciones de distintos arboles de decisión, para que sean diferentes arboles es necesario que creemos cada uno con una muestra aleatorio de los datos de entrenamiento, esta técnica se denomina bagging.

Esta técnica crea diferentes modelos usando muestras aleatorias con reemplazo y después combina o junta los resultados y predice los nuevos datos usando un "voto mayoritario", en donde se clasifica como positivo si la mayoría de los árboles predijo la observación como positiva.