UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Materia: Minería de Datos

Resúmenes: Técnicas de Minería.

Profesora Mayra Cristina Berrones Reyes

Alumno:

Alfonso Llanos Morales

1887939

7° Semestre

Licenciatura en Actuaría

02/10/2020

TÉCNICAS DESCRIPTIVAS:

Clustering

Es una técnica de aprendizaje de maquina no supervisada, esto quiere decir que la máquina podrá aprender por medio de los datos que le demos y que sea no supervisada se refiere a que no hay una interpretación de los datos por parte de una persona, que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes, a los subconjuntos creados por esta técnica se le denomina clúster, cada clúster esta conformado por datos que comparten características especificas que a su vez son diferentes entre cada clúster.

-Tipos Básicos de Análisis

Centroid Base Clustering: Cada clúster es representado por un centroide y son construidos en base a la distancia del punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más utilizado es el de *k-medias*.

Conectivity Base Clustering: Los clústers se definen agrupando a los datos más similares o cercanos, los puntos más cercanos están más relacionados que otros puntos más lejanos. Su característica principal es que el clúster contiene a otros clústers, representando así una jerarquía. El algoritmo utilizado es el *Hierarchical Clustering*.

Distribution Base Clustering: En este método cada clúster pertenece a una distribución normal, esto es debido a que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. Al utilizar probabilidades tenemos la ventaja de utilizar de mejor manera los datos atípicos. El algoritmo utilizado es el *Gaussian mixture models*.

Density Base Clustering: Los clústers están definidos por áreas de concertación, trata de conectar puntos con distancias muy pequeñas, aunque las formas que se obtengan no sean regulares.

-Método K-medias

Algoritmo de clustering basado en centroides. K representa el número de clusters y es definido por el usuario. Los clusters son creados basándose en el Centroid Base Clustering como ya se mencionó anteriormente. El proceso del algoritmo es definir los k centroides, crear los clusters basándonos en las distancias de los puntos a los centroides, obtener la media de cada clúster, el cual será el nuevo centroide y repetir el proceso hasta que no ocurran cambios. Podemos decir que el mejor clustering es aquel en el cual la suma de las varianzas de sus clustering es menor, pero existe un problema y es debido a que entre

mayor sea k, menor serán las varianzas, por lo que se llegará al punto donde cada k será igual al numero de puntos en la base de datos, por lo que su varianza será de cero. La solución a este problema es la aplicación del *método del codo*, que consiste en graficar la reducción de la varianza total o la varianza como tal a medida que aumenta k y luego ubicar el punto en el cual la reducción de la varianza ya no es tan significativa como antes, el punto es llamado *elbow plot* o codo, donde después del punto, la reducción de la varianza no estaría tan relacionada a la agrupación de mis datos sino al proceso natural de reducción de la varianza debido al aumento de k.

Reglas de Asociación

Las reglas de asociación son un tipo de análisis que extrae información por coincidencias con el objetivo de encontrar relaciones dentro de un conjunto de transacciones, en concreto ítems o atributos que tienden a ocurrir de forma conjunta, definido como: "si A(antecedente) entonces B(consecuencia)" o "A => B", en donde A y B son ítems individuales. Las reglas se dividen en diferentes tipos dependiendo de las características que se están tomando en cuenta para realizarlas, estos son: con base en el tipo de valores que manejan las reglas (Booleana o Cuantitativa), con base en las dimensiones de datos que involucra la regla (Unidimensional o Multidimensional) y con base en los niveles de abstracción que involucra la regla (de un nivel o Multinivel). Estas reglas son utilizadas para encontrar las combinaciones de artículos que ocurren con mayor frecuencia dentro de una base de datos y a su vez medir la fuerza e importancia de estas combinaciones, por lo que son aplicadas con diferentes fines como: definir patrones de navegación dentro de una tienda, promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancías en tiendas y segmentación de clientes con base en patones de compra.

Ahora que sabemos que es una regla de asociación y cómo se construye, necesitamos quedarnos con las reglas que sean más significativas para nuestras transacciones, por lo que utilizaremos las siguientes métricas de interés: el soporte, la confianza y el lift. El soporte se define como la frecuencia con la cual A y B aparecen juntos en la base de datos de transacciones y esta definido como Soporte(A => B) = P(A∩B), si el soporte es bajo significa que la regla pudo haberse dado por mera casualidad. La Confianza mide la fuerza de la regla y está definida como Confianza(A => B) = Soporte(A => B)/Soporte(A) = P(B/A) = P(A∩B)/P(A), si la confianza es baja es probable que no exista relación entre antecedente y consecuente. Por último el Lift que refleja el aumento de la probabilidad de que ocurra el consecuente, cuando sabemos que ocurrió el antecedente y está definido como Lift(A => B) = Soporte(A =>B)/{Soporte(A)*Soporte(B)} = P(A∩B)/{P(A)*P(B)} y dependiendo del valor del Lift podemos sacar diferentes conclusiones, si el Lift es muy cercano a 1 podemos

suponer que la regla fue hecha por el azar, si el >1 representa una relación fuerte y una frecuencia mayor a la del azar, y que es probable que los productos sean complementos; en cambio si el Lift es <1 representa una relación débil y una frecuencia menor a la del azar, por lo que es probable que los productos sean sustitutos.

Detección de outliers

Los outliers o datos atípicos son los datos que se encuentran alejados del resto de datos, debido a que no siguen el patrón que la mayoría sí. La observación que se desvía mucho del resto de las observaciones aparecerá como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de datos. Se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos. La detección de outliers es fundamental ya que de mantener estos datos dentro de nuestra base de datos pueden ocasionar distorsiones en los parámetros que calculemos, por ejemplo, si en un conjunto de datos que representan el salario de los trabajadores de una zona específica, siendo todos profesionistas con salarios dentro de lo normal parecido, pero junto con esos datos se encuentra el salario de una persona del sector informal, con un salario mucho menor, este dato podría hacer que el valor de la media sea menor que el de los profesionistas, dejando de representarlos o siendo menos significativa, por lo que dependiendo de ciertas características decidiremos si quitar este dato atípico de la base de datos. Uno de los algoritmos más utilizados para la detección de outliers es el *Density-based spatial clustering of applications with noise (DBSCAN)*, esto lo hace midiendo las distancias entre los puntos, esta distancia es denotada por épsilon, y formando clúster, dejando los ouliers fuera. El algoritmo clasifica en tres tipos de puntos, los puntos centrales o *Core del clúster*, que son los que cumplen con el mínimo número requerido de puntos dentro de la distancia de su épsilon; los puntos fronterizos o *Border del clúster*, siguen siendo parte del clúster, pero no cumplen el mínimo numero de puntos dentro de la distancia de su épsilon; y los puntos no asignados a un clúster o *Noise*, que son los outliers.

Visualización

Es la representación gráfica de información y datos, proporcionan una manera accesible de ver y comprender tendencias, valores y patrones de los datos, sobre todo cuando lo que necesitamos analizar son grandes cantidades de información y tomar decisiones basadas en estos datos. Se dividen en tres categorías: los elementos básicos en la representación de datos, como gráficas de líneas, de barras, puntos, mapas y tablas; los cuadros de mando, que son composiciones complejas de visualizaciones individuales que guardan una relación temática entre ellas; y por último las infografías, que son utilizadas para contar historias. Los estándares principales para la visualización de datos son: HTML5, CSS3, SCV y WebGL.

El poder mostrar los datos que tenemos de manera visual es muy importante hoy en día, ya que cada ves es más común que las empresas utilicen sus bases de datos para la toma de decisiones y por medio de estas representaciones gráficas podemos llegar a saber el quien, que, donde, cuando y porque de cualquier conjunto de datos y la problemática que se quiera resolver con ellos. La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

La función *describe()* nos da información básica de una columna especificada de la base de datos, datos como la media, los cuartiles y la desviación estándar. Otro punto a recalcar es que las librerías *matplotlib.pyplot* y *seaborn* son unas de las principales para graficar datos en Python.

TÉCNICAS PREDICTIVAS:

Regresión

La regresión es una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en os datos recolectados de otros atributos. La regresión se encarga de analizar el vinculo entro una variable dependiente y una o varias independientes, encontrando una relación matemática. Existen diferentes tipos de regresiones, como la regresión lineal simple, regresión lineal múltiple, regresiones polinómicas o regresiones linealizables.

La regresión lineal simple solo tiene una variable regresora y tiene la forma: $y = \beta_0 + \beta_1 x + e$, donde "$\beta_0$" y "$\beta_1$" son los coeficientes de la ecuación, "*y*" es la variable dependiente, "*x*" es la variable independiente o regresora y "*e*" es el error de la estimación. Cabe destacar que "β_0" es el coeficiente de la intersección con el eje x de la ecuación de regresión. Para el cálculo de los coeficientes se utiliza la estimación por mínimos cuadrados, en la cual se busca reducir el error "*e*" al mínimo, para que así la ecuación se apegue lo más posible a los datos reales, de manera gráfica podemos representar al error como la diferencia entre el

valor real de un dato y el valor que tiene nuestra ecuación de regresión para el mismo dato, por lo que se buscará una línea recta que se encuentre lo más cerca posible de todos los ítems de la base de datos.

La regresión lineal múltiple está dada por k variables regresoras y tiene la forma: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + e$, *donde* $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de las $x_1, x_2, x_3, \ldots, x_k$ regresores, "*y*" es la variable dependiente y *"e"* es el error de estimación.

En Python para poder realizar las regresiones primero tendremos que importar la librería "matplotlib.pyplot" y ya con ella podremos representar en gráficos la dispersión de nuestros datos con plt.scatter() y aplicando las formulas para la obtención de las betas por mínimos cuadrados y con un plt.plot() podremos graficar la ecuación de la regresión.

Clasificación

Es la organización o mapeo de un conjunto de atributos por clase dependiendo de sus características, conociendo las características de los datos podemos hacer predicciones a futuro. Los modelos de clasificación pueden ser clasificados dependiendo de la certeza en la predicción, y existen cuatro casos: decir que es verdadera cuando es verdadera, decir que es verdadera cuando es falsa, decir que es falsa cuando es verdadera y decir que es falsa cuando es falsa, de aquí las dos predicciones incorrectas son llamadas *Error tipo I* o *falso positivo*, donde se dice que es verdadera, cuando es falsa y el *Error tipo II* o *falso negativo*, donde se dice que es falsa cuando es verdadera. Otros conceptos importantes son la *certeza*, denotada por el numero de predicciones correctas entre el numero de predicciones y su complemento llamado *tasa del error*.

Redes neuronales: son redes que tienen la apariencia de neuronas. Se componen de una capa de entrada, una capa de salida y una capa oculta, en estas capas se encuentran ciertos nodos virtuales donde ocurre la clasificación. Su funcionamiento se basa en las conexiones, y su aprendizaje se da con las repeticiones hasta que el algoritmo encuentre la salida deseada y una vez que el algoritmo haya aprendido del problema la predicción será mejor.

Árboles de decisión: son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos. Las principales desventajas son que las reglas no necesariamente forman un árbol, también puede que n se lleguen a cubrir todas las posibilidades y las reglas pueden entrar en conflicto.

Patrones secuenciales

Es una clase espacial de dependencia en la que el orden de los acontecimientos es considerado, son eventos que se relacionan con el paso del tiempo. Se trata de buscar asociaciones de la forma "si sucede el evento x en el instante del tiempo t entonces sucederá el evento y en el instante t+n". Su objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos. Utiliza reglas de asociación secuenciales, las cuales expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos de tiempo.

Sus características son: que el orden importa, el tamaño de una secuencia es su cantidad de elementos(itemset), la longitud de una secuencia es su cantidad de ítems, el soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S y las secuencias frecuentes son subsecuencias de una secuencia que tiene soporte mínimo.

Agrupación de patrones secuenciales: separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre si, y al mismo tiempo sean diferentes a los objetivos de otros grupos, parecido al *Centroid Base Clustering* visto en el tema de clustering.

Reglas de asociación con datos secuenciales: se presentan cuando los datos contiguos presentan algún tipo de relación, expresan patrones de comportamiento secuenciales, es decir, que se den en instantes distintos, pero cercanos, de tiempo. Tiene relación con el tema *reglas de asociación* visto anteriormente.

Predicción

-Metodología de la partición de datos

Para comenzar se utiliza la estrategia de división de los datos, la forma más común es dividir los datos de la siguiente manera: 70% conjunto de entrenamiento, con el cual vamos a construir el modelo, siendo la base de nuestro modelo; un 15% como conjunto de pruebas, para la aplicación de prueba y error; y un 15% de conjunto de validación que se procuran usar solo una vez cuando ya tengamos nuestro modelo bien perfeccionado y tiene el objetivo de validar el modelo.

-Árboles aleatorios

Es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Primeramente, se dividirá

la gráfica o cuadrícula de las dispersiones de los datos en subregiones y cada subregión tendrá una respuesta. Si la subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos de la misma clase. Los árboles de decisión están formados por nodos, el nodo raíz es el principal y de este se desglosan los nodos internos y así sucesivamente hasta las respuestas que vendrían siendo las hojas de nuestro árbol. Los árboles aleatorios se dividen en dos tipos: los árboles de clasificación, el los cuales la variable de respuesta "y" es cualitativa y los árboles de predicción, en los cuales la variable de respuesta "y" es cuantitativa. Otro concepto muy importante es el *Gini* o medida de impureza, esta impureza hace referencia a que tan mezcladas están las clases en cada nodo, cuando el *Gini* vale 0 significa que ese nodo es totalmente puro.

-Bosques aleatorios

Es una técnica de aprendizaje automático supervisada basa de árboles de decisión, la principal ventaja es que obtiene un mejor rendimiento de generalización y el error cuadrático medio se reduce ya que se promedian los errores de cada árbol. Para asegurarnos que cada uno de los árboles es distinto, cada uno se entrena con una muestra aleatoria con remplazo generada de los datos de entrenamiento, estrategia denominada bagging. Por lo que después de realizar esta técnica tendremos una gran cantidad de árboles de decisión y por último se va a predecir el resultado usando el "voto mayoritario", donde se clasificará como positivo si la mayoría de los árboles predicen la observación positiva.

-Validación cruzada

Su función consiste en correr varias veces el modelo para asegurar que el modelo sea generalizado, se estima el test error rate del modelo y a las diferentes técnicas de realizarlo son las siguientes: Validación simple, que es dividir en porcentajes los datos; el Leave One Out Cross-Validation, que es dejar un dato fuera del conjunto y analizar los cambios ocasionados por la presencia o ausencia del dato, y así sucesivamente con los demás, se irán sacando y metiendo cada uno de los datos; por último el (LOOCV) K-Fold Cross-Validation, que es muy similar al anterior pero con grupos de datos.