UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MINERÍA DE DATOS

RESÚMENES
TÉCNICAS DE MINERÍA DE DATOS

FLOR KARINA JUÁREZ RODRÍGUEZ
1802920

M.C. MAYRA CRISTINA BERRONES REYES

SEMESTRE 7
LICENCIADO EN ACTUARÍA

SAN NICOLÁS DE LOS GARZA A 02 DE OCTUBRE DE 2020

CLUSTERING

El clustering es una técnica de aprendizajes de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándose en similitudes. Esta técnica tiene usos como la investigación de mercado, para identificar comunidades, prevención de un crimen, procesamiento de imágenes, entre otros.

Tipos básicos de análisis:

- Centroid Based Clustering
- Connectivity based Clustering
- Distribution Based Clustering
- Density Based Clustering

Centroid Based Clustering

Cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia de un punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más usado de este tipo es el de K-medias.

Connectivity based Clustering

Los cluster se definen agrupando a los datos más similares o cercanos (esto porque los puntos más cercanos están más relacionados que ortos puntos más lejanos). La característica principal es que un clster contiene a otros clusters (representan una jerarquía). Un algoritmo usado de este tipo es Hierarchical clustering

Distribution Based Clustering

En este método cada cluster pertenece a una distribución normal. La idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. Un algoritmo de clustering perteneciente a este tipo es Gaussian mixture models.

Density Based Clustering

Los clusters son definidos por áreas de concetración. Se trata de conectar puntos cuya distancia entre si es considerada pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia limitada y considera como iregular a las áreas esparcidas entre clusters.

Método K-medias

Este método es un algoritmo de clustering basado en centroides donde K representa el número de clusters y es definido por el usuario. Una vez se escoja el valor de k se realizan los siguientes pasos:

1. Centroides. Se eligen k datos aleatorios que pasarán a ser los centroides representativos de cada cluster.
2. Distancias. Se analiza la distancia de cada dato al centroide más cercano, perteneciendo a su cluster.
3. Media. Obtener la media de cada cluster y este será el nuevo centro.
4. Iterar. Repetir el proceso hasat que los clsuters no cambien.

Varianza de los clusters

La varianza de cada cluster disminuye al aumentar k. Si sólo hay un elemento en el cluster la varianza es de cero. Entre menos sea la suma de las varianzas de los clusters, mejor es nuestro clustering.

Método del codo

Consiste en graficar la reducción de la varianza total a medida que k aumenta. En un punto la redución de la varianza no disminuirá de forma significativa entre un valor k y otro. Ese punto es llamado elbow plot o como y representa el número de k a utilizar.

REGLAS DE ASOCIACIÓN

Las reglas de asociación se derivan de 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. Nos permiten encontrar las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos transaccional. Una regla de asociación se define como una implicación del tipo: $Si\ A \Rightarrow B$, siendo A un antecedente y B la consecuencia.

Algunas aplicaciones son: define patrones de navegaciones dentro de una tienda, promociones de pares de productos, soporte para la toma de decisiones, analiza información de ventas, ayuda en la distribución de mercancías en tiendas y segmenta clientes con base en patrones de compra.

Tipos de reglas de asocación

Con base en los tipos de valores que manejas las reglas:

- Asociación Booleana: asocaciones entre la presencia o ausencia de un ítem.
- Asociación Cuantitativa: descrine asocaciones entre ítems cuantitativos o atributos.

Con base en las dimensiones de datos que nos involucra una regla:

- Asociación Unidimiensional: los ítems de la regla se referencian en una sola dimensión.
- Asociación Multidimensional: los ítems de la regla se referencian en dos o más dimensiones.

Con base en los niveles de abstracción que involucran la regla:

- Asociación de un nivel: los ítems son referenciados en un nivel único de abstracción.
- Asociación Multinivel: los ítems son referenciados a varios niveles de abstracción.

Métricas de interés

Soporte. Dada una regla $Si\ A \Rightarrow B$, el soporte se define como el número de veces o frecuencia (relativa) con que A y B aparecen juntos en una base de datos transacciones.

- Regla de bajo soporte: puede haber aparecio por casualidad

Confianza. Dada una regla $Si\ A \Rightarrow B$, la confianza es el cociente del soporte de la regla y el soporte del antecedente solamente. La confianza mide la fortaleza de la regla.

- Regla con baja confianza: es probable que no exista relación entre antecedente y consecuente.

Lift. Refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente.

DETECCIÓN DE OUTLIERS

Un outlier o dato extremo es una observación que se desvía de las otras o en otro sentido, datos que parecen inconsistentes con el conjunto de datos. Los métodos para detección de outliers pueden ser clasificados como univariantes y multivariantes. En la práctica las variables tienen valores inusuales, muy grandes o muy pequeños; estos outliers pueden ser causados por medidas incorrectas, datos erróneos o por venir de una población diferente a la mayoría de los datos.

Los outliers aumentan la varianza del error y reducen la potencia de las pruebas estadísticas. Si son distribuidos aleatoriamente, pueden quebrantar la normalidad. En la prueba de hipótesis las probabilidades pueden influencias el error tipo I (rechazas una hipótesis que es verdadera) y tipo II (no rechazar una hipótesis que es falsa). También pueden alterar las estimaciones causando sesgos.

El proceso de detección incluye un proceso de minería de datos que utiliza herramientas basadas en algoritmos de tupo no supervisado. El proceso de detección consta de dos enfoques según su forma: local y global. Los enfoques globales incluyen un conjunto de técnicas en las que se asigna una puntuación a cada anomalía con relación al conjunto de datos globales. Por otro lado, los enfoques locales, representan las anomalías en un dato determinado con respecto a su vecindad directa; es decir, a los datos cercanos en cuanto a la similaridad de sus características. De acuerdo con los conceptos antes mencionados, el enfoque local detecta valores atípicos que son ignorados cuando se utiliza un enfoque global, en especial en aquellos con densidad variable.

Tipos de outliers

- Casos atípicos que surgen de un error de procedimiento
- Observación que ocurre como consecuencia de un acontecimiento extraordinario
- Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables
- Datos extraordinarios para las que el investigador no tiene explicación.

Identificación de outliers

Se debe examinar la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico. Esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas. Para muestras pequeñas (de 80 o incluso menos observaciones), las pautas sugeridas identifican como atípicos aquellos casos con valores estándar de 2.5 o superiores. Cuando los tamaños muestrales son mayores, las pautas sugieren que el valor umbral sea 3.

Esta técnica se puede aplicar en el seguramiento de ingresos en las telecomunicaciones, detección de fraudes financieros, seguridad y detección de fallas.

VISUALIZACIÓN

La visualización de datos es la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. La visualización de datos es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Tipos de visualizaciones

Existen multitud de técnicas y aproximaciones para la visualización según sea la naturaleza del dato de la información. Podemos establecer la siguiente clasificación de tipos de visualización según complejidad y elaboración de la información.

1. Elementos básicos de representación de datos.

Es el caso más sencillo, a continuación, se señalan algunos tipos de visualizaciones básicas:

- Gráficas: barras, líneas, columnas, puntos, "tree maps", tarta, semi-tarta, etc.
- Mapas: burbujas, coropletas (o mapa temático), mapa de calor, de agregación (o análisis de drill-down)
- Tablas: con anidación, dinámicas, de drill-down, de transiciones, etc.

2. Cuadros de mando.

Un cuadro de mando es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Son ampliamente utilizados en las organizaciones para análisis de conjuntos de variables y toma de decisiones.

3. Infografías

Las infografías no están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos; es decir, las infografías se utilizan para contar "historias". Esta narrativa no se construye a través de texto, sino mediante la disposición de la información en la que las visualizaciones se combinan con otros elementos como: símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Importancia de la visualización de datos en cualquier empleo

Los conjuntos de habilidades están cambiando para adaptarse a un mundo basado en los datos. Para los profesionales es cada vez más valioso poder usar los datos para tomar decisiones y usar elementos visuales para contar historias con los datos para informar quién, qué, cuándo, dónde y cómo. La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

REGRESIÓN

La regresión es una técnica de minería de datos de la cateogría predictiva. Predice el valor de un atributo en particula basándose en los datos recolectados de otros atriutos. La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Regresión lineal

Cuando el análisis de regresión sólo se traya de una variable regresora, se llama regresión lineal simple y tiene como modelo $y = \beta_0 + \beta_1 x + \varepsilon$, donde la cantidad ε es una variable aleatoria normalmente distribuida con $E(\varepsilon) = 0$ y $vAR(\varepsilon) = \sigma^2$.

→ Estimación por mínimos cuadrados. La estimación de $y = \beta_0 + \beta_1 x + \varepsilon$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x$

Regresión lineal múltiple

Un modelo d eregresión múlitple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos. En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

→ Estimación por mínimos cuadrados.

La función de mínimos cuadrados es $S(\beta_0, \dots, \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$

Las aplicaciones de esta técnica son en la medicina, informática, estadística, comportamiento humano e industria.

CLASIFICACIÓN

La clasificación es la técnica de minería de datos más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características. Se entrena (estima) un modelo usando los datos recolectados para hacer predicciones futuras.

Técnicas de clasificación:

- Clasificación por inducción de árbol de decisión
- Clasificación Bayesiana
- Redes neurales
- Support Vector Machines (SVM)
- Clasificación basada en asociaciones

Regla de Bayes

Si tenemos una hipótesis para $E \rightarrow p(A|B) = (p(B|A) * p(A))/p(B)$ donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B.

Redes neuronales

Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse. Las redes neuronales consisten generalmentene en tres capas: de entrada, oculta y de salida. Internamente puede verse como una gráfica dirigida.

Árbol de decisión

Son una seire de condiciones organizadas en forma jerárquica, a modo de árbol, son útiles para problemas que mezclen datos categóricos y númericos.

→ Problemas con la intrudcción de reglas:

- Las reglas no necesariamente forman un árbol
- Las reglas puedne no cubrir todas las posibilidades
- Las reglas pueden entrar en conflicto

PATRONES SECUENCIALES

Se especializan en analizar los datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias. El patrón secuencial describe el modelo de compras que hace un cliente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo. Son eventos que se enlazan con el paso del tiempo. Son eventos que se enlazan con el paso del tiempo.

Los patrones secuenciales buscan asociaciones de la forma "si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n". El objetivo de la tarea es poder describir de forma concisa relaciones temporales que existen entre los valore de los atributos del conjunto de ejemplos.

Características:

- El orden importa
- Su objetivo es encontrar patrones en secuencia
- Una secuencia es una lista ordenada de ítemsets y cada uno es elemento de una secuencia
- El tamaño de una secuencia es su cantidad de elementos (ítemsets)
- La longitud de una secuencia es su cantidad de ítems
- El soporte de una secuencia es el % de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes (o patrones secuenciales) son las subsecuencias de una secuencia que tienen un soprte mínimo.

Esta técnica se desarrolla en las áreas de medicina, biología, bioingienería, web, análisis de mervado, distribución y comercio, etc. Y los tipos de base de datos que utiliza son las bases temporal, documentales y relacionales.

Agrupamiento de patrones secuenciales. Tarea de separar en grupos a los datos de manera que los miembros de un grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a los objetivos de otros grupos.

Clasificación con datos secuenciales. Éstos expresan patrones de comportamiento secuenciales, es decir, que se dan en instantes distintios (pero cercanos) en el tiempo.

Reglas de asociación con datos secuenciales. Se presentan cuando los datos contiguos presentan algún tipo de relación.

PREDICCIÓN

Árboles aleatorios

Árbol de decisión. Modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Para dividir el espacio muestral en subregiones es precio aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Estructura básica de un árbol de decisión (formados por nodos y se leen de arriba para abajo):

- *Primer nodo o nodo raíz.* En él se produce la primera división en función de la variable más importante.
- *Nodos internos o intermediarios.* Tras la primera división, estos nodos vuelven a dividir el conjunto de datos en función de las variables.
- *Nodos terminales u hojas.* Se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva

Los árboles se pueden clasificar en dos tipos:

→ Árbol de clasificación. Consiste en hacer preguntas del tipo $¿\, x_k \leq c?$ para las covariables cuantitativas o del tipo $¿\, x_k = nivel_j?$ para las covariables cualitativas, de esta forma el espacio de las covariables es dividido en hiperrectángulos y todas las observaciones que queden dentro de un hiperrectángulo tendrán el mismo valor grupo estimado.

Hay dos tipos de nodo:

- *Nodos de decisión.* Tienen una condición al principio y tiene más nodos debajo de ellos.
- *Notos de predicción.* No tienen ninguna condición ni nodos debajo de ellos.

→ Árbol de regresión. Consiste en hacer preguntas del tipo $¿\, x_k \leq c?$ para cada una de las covariables, de esta forma el espacio de las covariables en dividido en hiperrectángulos y todas las observaciones que queden dentro de un hiperrectángulo tendrán el mismo valor grupo estimado.

Pasos para realizar la partición del espacio:

1. Dado un conjunto de covariables, encontrar la covariable que permita predecir mejor la variable respuesta.
2. Encontrar el punto de corte sobre esa covariable que permita predecir mejor la variable respuesta.
3. Repetir los pasos anteriores hasta que se alcance el criterio de parada.

Bosques aleatorios

Random forest. Técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento durante entrenamiento similar, ésta mejora consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarse que los árboles sean distintos se aplica una estrategia llamada bagging.

→ Bagging. Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

Dado que un random forest es un conjunto de árboles de decisión y los árboles son modelos no-paramétricos, los random forests tienen las mismas ventajas y desventajas:

- *Ventaja*. Pueden aprender cualquier correspondencia entre datos de entrada y resultado a predecir.
- *Desventaja*. No son buenos extrapolando porque no siguen un modelo conocido.

Validación de cruzada

Se emplea para estimar el test error rate de un modelo y así evaluar su capacidad predictiva, a este proceso se le conoce como model assessment. También se puede emplear para seleccionar el nivel de flexibilidad adecuado.

Métrica de eficacia para datos numéricos y categóricos

Error cuadrático medio.

- Mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima
- Valor esperado de la pérdida del erro cuadrado

Curva ROC

- Nos sirve para conocer el rendimiento global de la prueba (área bajo la curva)
- Eje X - Falsos positivos
- Eje Y - Verdaderos positivos