UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Resumen de
Técnicas de Minería de Datos

Actividad 03

Fase 1

Docente: Mayra Cristina Berrones Reyes

Keila Rubí Puente González 1807864

02 de octubre de 2020

DESCRIPTIVAS

Clustering

El *Clustering* es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar un conjunto de objetos en subconjuntos de objetos llamados Clusters donde cada Cluster está formado por una colección de objetos que considerados similares entre sí, pero que son distintos respecto a los objetos de otros Clusters.

Al ser una técnica de aprendizaje no supervisada no tiene una clase de respuesta. Es decir, no se cuenta con información sobre la estructura del dominio de salida por lo que después de agrupar observaciones es necesario asociarle un significado o característica distintiva a cada clúster.

Algunos ejemplos del uso del Clustering son:

- Investigación de mercado: Detectar diferentes tipos de clientes según su forma de compra.
- Identificar comunidades en redes sociales de acuerdo con atributos como ubicación, gustos, etc.
- Prevención de crímenes: Se identifican las principales problemáticas de las comunidades y con base en ello se establecen estrategias.
- Procesamiento de imágenes: Reconocer áreas con ciertas características de tierra (GIS).

Su objetivo es encontrar grupos de objetos similares. Antes de aplicar esta técnica en muchos casos será necesario aplicar algún tipo de transformación a los datos:

- Para *variables cuantitativas* se recomienda una transformación si estas presentan diversas unidades de medida siendo la más popular la estandarización.
- Las *variables binarias* no suelen sufrir transformaciones.
- Las *variables categóricas* son convertidas en variables numéricas por binarización. (presencia/ausencia) para las distintas modalidades (One Hot Encoding).

Los cuatro tipos básicos de análisis (métodos) son:

1. *Centroid Based Clustering*: Cada clúster es representado por un centroide compuesto por puntos que se eligen al azar. Los Clusters se construyen basados en la menor distancia del 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,* donde K representa el número de clusters y es definido por el usuario. En resumen, los pasos son:
   Paso 1: Escoger aleatoriamente k puntos distintos que serán los centroides representativos de cada Cluster.

Paso 2: Medir la distancia entre cada punto y los k puntos iniciales de los Clusters y asignar cada punto al Cluster más cercano que tiene.
Paso 3: Calcular la media de cada cluster y este será nuestro nuevo centro.
Paso 4: Iterar hasta que los Clusters no cambien.
Podemos evaluar la calidad de los Clusters evaluando con la suma de las varianzas de cada uno.

2. *Connectivity Based Clustering:* Los Clusters se definen agrupando a los datos más similares o cercanos basándonos en la premisa de que los puntos más cercanos están más relacionados que otros puntos más lejanos.
   La característica principal es que un Cluster contiene a otros clusters, y debido a esta estructura los clusters representan una jerarquía los cuales son representados por un dendrograma. Este método funciona de dos maneras: ascendente; puede comenzar desde los clusters más pequeños que se unen para formar un cluster más grande o descendente; que es comenzar desde el cluster más grande y en cada paso se divide en dos clusters más pequeños.
   Hierarchical clustering es un algoritmo de clustering perteneciente a este tipo.
3. *Distribution Based Clustering:* En este método cada cluster pertenece a una distribución normal, la idea es que los puntos son divididos basados en la probabilidad de pertenecer a la misma distribución normal.
   Es similar al *Centroid Based Clustering* solo que *Distribution Based Clustering* utiliza la probabilidad para "calcular" los clusters en lugar de utilizar la media. El usuario define el número de clusters y es un método iterativo para optimizar clusters.
   Una de sus ventajas frente a otros tipos de clustering es que se tienen probabilidades de pertenecer a un grupo en lugar de asegurar que se pertenece de manera definitiva y permite un mejor manejo de los datos que podrían ser considerados atípicos.
   Gaussian mixture models es un algoritmo de clustering perteneciente a este tipo.
4. *Density Based Clustering:* Los clusters son definidos por áreas de concentración. Este método comienza buscando áreas de puntos concentrados (cercanos) y asigna esas áreas al mismo cluster. Se trata de conectar puntos cuya distancia entre sí es pequeña y considera como irregular a las áreas esparcidas entre clusters. Este tipo de clustering permite conectar áreas con formas arbitrarias, pero tiene dificultad con datos cuya densidad varía y con grandes dimensiones.

Reglas de Asociación

Una regla de asociación se define como una implicación del tipo: “Si A => B “, esto es, antecedente, consecuencia; donde A y B son artículos individuales.

Se derivan de un tipo de análisis que extrae información por coincidencias, su objetivo es encontrar relaciones dentro un conjunto de transacciones, en concreto, artículos o atributos que tienden a ocurrir de forma conjunta.

Por lo tanto, las reglas de asociación nos permiten desde encontrar las combinaciones de artículos que ocurren con mayor frecuencia en una base de datos transaccional hasta medir la fuerza e importancia de estas combinaciones.

Algunos ejemplos de su uso son:

- Definir patrones de navegación dentro de la tienda.
- Promociones de pares de productos: Hamburguesas y Cátsup.
- Soporte para la toma de decisiones.
- Análisis de información de ventas.
- Distribución de mercancías en tiendas.
- Segmentación de clientes con base en patrones de compra.

Entre los tipos de Reglas de Asociación según los tipos de valores que manejan estas reglas están la *Asociación Booleana* y la *Asociación Cuantitativa;* si son con base en las dimensiones de datos que involucra una regla entonces tenemos *Asociación Unidimensional* y *Asociación Multidimensional*; finalmente, si nos basamos en los niveles de abstracción tenemos la *Asociación de un nivel* y la *Asociación Multinivel*.

Las métricas de interés que se presentan son:

- Soporte: se define como el número de veces o la frecuencia (relativa) con que A y B aparecen juntos en una base de datos de transacciones.
- Confianza: mide la fortaleza de la regla.
- Lift: Refleja el aumento de la probabilidad de que ocurra el consecuente, cuando nos enteramos de que ocurrió el antecedente.

Detección de Outliers

Son datos atípicos, ya que los datos suelen seguir un patrón de comportamiento y estos son datos raros con comportamientos inusuales. Se desvían mucho del resto de las observaciones apareciendo como una “observación sospechosa” que pudo ser generada por mecanismos diferentes al resto de los datos.

Pueden ser usados en cualquier lugar donde se manejen grandes cantidades de datos, algunos ejemplos de su uso son:

- Aseguramiento de ingresos en las telecomunicaciones.

- Detección de fraudes financieros.
- Seguridad y la detección de fallas.

Se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos.

Uno de los algoritmos más conocidos es Density-Based Spatial Clustering of Applications with Noise, detecta los otliers eligiendo un punto al azar, tiene dos parámetros, el primero es épsilon y mide la distancia de un punto hasta el punto elegido al azar, el segundo es mínimo de puntos define un cluster. Los separa en puntos centrales que son los que cumplen con la característica de tener el mínimo de puntos, los Border son más pequeños, sin embargo, siguen cumpliendo con la distancia requerida, y Noise son los otliers, los que no cumplieron con la característica.

En k-medias si hemos encontrado que hay puntos que se alejan del punto central, aunque estén dentro del cluster y estos se encuentran cerca de otros puntos lejanos de otros clusters, entonces con esos se forma un nuevo cluster independiente y se investiga como atípico.

Se pueden verificar por medio de gráficas como BoxPlot.

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.

Es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos, se encuentra justo en el centro del análisis y la narración visual.

Los tipos de visualizaciones se pueden clasificar según la complejidad y elaboración de la información:

- Elementos básicos de representación de datos; es el caso más sencillo, 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 drilldown).
  -Tablas: con anidación, dinámicas, de drilldown, de transiciones, etc.
- Cuadros de mando: es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Son utilizados en las organizaciones para análisis de conjuntos de variables y toma de decisiones.

- 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.

Los estándares web que se han ido desarrollando en los últimos años para la evolución de las aplicaciones web, base fundamental para creación de visualizaciones web basadas en datos.

Estándar	Última Versión	Función
HTML5	v5	Canvas: elemento HTML para dibujar gráficos 2D
CSS3	v3	Permite diferenciar el contenido de las páginas web de la presentación de este contenido
SCV	v2	Utilizado para crear gráficos 2D
WebGL	v1	Gráficos 3D haciendo uso de Canvas

Vivimos en un mundo en evolución y 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 porque al usar muchas gráficas son mucho más fácil de ver y de comprender la información.

PREDICTIVAS

Regresión

En 1805, se documentó la primera forma de regresión lineal y este fue el método de los mínimos cuadrados publicado por Legendre.

En 1889, se introdujo el término regresión por Francis Galton en su libro “Natural inheritance” y fue confirmada por su amigo Karl Pearson. Se utilizó en el estudio de variables eugénicas, esto es en el estudio de las estaturas de padres e hijos la cual mostró que los hijos que tenían padres con estatura mayor que la del promedio tendían a “regresar” a la media de la estatura de la población a lo largo de sucesivas generaciones. Esto es, hijos de padres demasiado altos tendían a ser en promedio más bajos que sus padres, e hijos de padres muy bajos tendían a ser en promedio más altos que sus padres.

Gauss desarrolló de manera más profunda el método e incluía una versión del teorema de Gauss-Markov.

Los modelos lineales son una explicación ágil y simplificada de la realidad por parte de la matemática y estadística.

Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos, es decir, analiza el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

La regresión lineal simple es cuando el análisis de regresión solo se trata de una variable regresora.

La regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos, es decir, la respuesta “y” con los k regresores.

Algunos ejemplos de su uso son:

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

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, como su lugar, altura, estatus.

Se estima un modelo usando los datos recolectados para hacer predicciones futuras, es decir, entre más características conozcamos podremos describirlo mejor.

Es fácil medir la certeza de nuestros modelos, es el número de predicciones correctas entre el total de predicciones.

Algunas técnicas de clasificación son las siguientes:

- La regla de Bayes: Esto es si tenemos una hipótesis H sustentada para una evidencia E -> p(H|E) = (p(E|H)*p(H))/p(E), donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B.
- Redes neuronales: Son redes que parecen neuronas, trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben

enumerarse. También consisten generalmente de tres capas, de entrada, oculta y de salida. Internamente pueden verse como una gráfica dirigida.
Se usan en Clasificación, Agrupamiento, Regresión.

- Clasificación por inducción de árbol de decisión: Su imagen se asemeja a un árbol, por la forma en la que está hecho, es un método analítico que nos ayuda a tomar mejores decisiones, son una serie de condiciones organizadas en forma jerárquica, y son útiles para problemas que mezclen datos categóricos y numéricos como en clasificación, agrupamiento, regresión y en problemas con la inducción de reglas, estas no necesariamente forman un árbol además pueden no cubrir todas las posibilidades y las reglas pueden entrar en conflicto.
- Support Vector Machines (SVM).
- Clasificación basada en asociaciones: Esto es necesario para poder tener un buen resultado, podemos hacerlo incluso por medio de Clustering. Es la base principal.

Patrones Secuenciales

Es una clase especial de dependencia en las que el orden de acontecimientos es considerado, esto es que importa. Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, una secuencia es una lista ordenada de itemsets, donde cada itemset es un elemento de la secuencia y el tamaño o longitud de esta es su cantidad de elementos.

El soporte de una secuencia es el porcentaje 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 soporte mínimo

Describe el modelo de compras que hace un cliente particularmente o un grupo de clientes relacionando las distintas transacciones efectuadas por ellos a lo largo del tiempo, es decir, son eventos que se enlazan con el paso del tiempo.

Se trata de buscar 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”, lo cual es una regla de asociación secuencial, esto es, reglas que expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos en el tiempo.

El objetivo de la tarea es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

Las áreas y tipo de base de datos en los que se desarrolla son:

- Medicina.
- Biología, bioingeniería.
- Web.

- Análisis de mercado, distribución y comercio.
- Aplicaciones financieras y banca.
- Aplicaciones de seguro y salud privada.
- Deportes.
- Base de datos temporales.
- Base de datos documentales.
- Base de datos relacionales.

Para la resolución de problemas tenemos:
Agrupamiento de patrones secuenciales: es la 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: expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo, como por ejemplo comprar un auto y gasolina.

Reglas de Asociación con datos secuenciales: Se presenta cuando los datos contiguos presentan algún tipo de relación, como por ejemplo comprar leche y pan.

Algunos métodos representativos son: AprioriAll, GSP, SPADE, FreeSpan, SPAM, PrefixSpan, ISM, IncSp, ISE, IncSpan.

Predicción

Primero se deben tomar en cuenta algunos elementos para hacer un buen modelo de predicción:

- Definir adecuadamente nuestro problema (objetivo, salidas deseadas).
- Recopilar datos.
- Elegir una medida o indicador de éxito.
- Preparar los datos (tratar con campos vacíos, con valores categóricos).

Para tener una mayor precisión es necesario dividir los datos, el 70% es para el modelo, el 15% conjunto de validación y el 15% conjunto de pruebas.

Hacemos “prueba y error” si no funciona en ese momento la prueba empezamos en el entrenamiento nuevamente, después de hacer las correcciones pasamos al conjunto de validación la cual la hacen unos expertos al probar el modelo.

Dentro de los árboles aleatorios tenemos al árbol de decisión el cual es un 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 preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una 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 se pueden clasificar en dos tipos que son árboles de regresión en los cuales la variable respuesta "y" es cuantitativa; y en árboles de clasificación en los cuales la variable respuesta "y" es cualitativa.

Los árboles de clasificación consisten en hacer preguntas del tipo ¿$x_k \leq c$? para las covariables cuantitativas o preguntas del tipo ¿$x_k = nivelj$? para las covariables cualitativas, de esta forma el espacio de las covariables es divido 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 tienen más nodos debajo de ellos.
- Nodos de predicción: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo».

Gini es una medida de impureza. Cuando Gini vale 0, significa que ese nodo es totalmente puro. La impureza se refiere a cómo de mezcladas están las clases en cada nodo.

Los árboles de regresión consisten en hacer preguntas de tipo ¿$x_k \leq c$? para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiperrectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado $\widehat{y}$.

Algunas de las ventajas de los árboles de regresión son:

- Fácil de entender e interpretar.
- Requiere poca preparación de los datos.
- Las covariables pueden ser cualitativas o cuantitativas.
- No exige supuestos distribucionales.

El Random Forest es una técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarnos que los árboles sean distintos, lo que hacemos es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta

estrategia se denomina bagging y consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

La Validación 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.

Las métricas de eficacia para datos categóricos y numéricos son el error cuadrático medio que mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima y curva roc que nos sirve para conocer el rendimiento global de la prueba (área bajo la curva).