Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico-Matemáticas

Minería de datos

Resumen: Técnicas de minería de datos

Nombre: Pilar Abigail Mendoza Alvarez

Matricula: 1815973

Licenciatura: Actuaria

Grupo: 002

Viernes 02 Octubre 2020

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 un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta.

Una regla de asociación se define como una implicación del tipo : " Si A => B " antecedente consecuencia donde A y B son ítems individuales.

Las reglas de asociación nos permiten:

- Encontrar las combinaciones de artículos o ítems que ocurren con mayor frecuencia en una base de datos transaccional.
- Medir la fuerza e importancia de estas combinaciones.

Tipos de Reglas de Asociación

Asociación Cuantitativa. Con base en los tipos de valores que manejan las reglas:

• Asociación Booleana: asociaciones entre la presencia o ausencia de un ítem.

• Asociación Cuantitativa: describe asociaciones entre ítems cuantitativos o atributos.

Asociación Multidimensional

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

• Asociación Unidimensional: Si los ítems o atributos de la regla se referencian en una sola dimensión.

• Asociación Multidimensional: Si los ítems o atributos de la regla se referencian en dos o más dimensiones.

Asociación Multinivel

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

• Asociación de un nivel: Los ítems son referenciados en un único nivel de abstracción.

• Asociación Multinivel: Los ítems son referenciados a varios niveles de abstracción.

Métricas de interés

Soporte $\frac{\text{Frecuencia en que A} \cap \text{B aparecen en las transacciones}}{\text{Total de transacciones}}$

Confianza $(A \Rightarrow B) = P(B/A) = \frac{P(A \cap B)}{P(A)}$

$Lift\ A \Rightarrow B = \frac{\text{Soporte (A=>B)}}{\text{Soporte (A)} * \text{Soporte(B)}} = \frac{P(A \cap B)}{P(A) * P(B)}$

OUTLIERS

Datos atípicos. Problema de la detección de datos raros o comportamientos inusuales en los datos.

"Observación que se desvía mucho del resto de las observaciones apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos"

Aplicaciones

- 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.

Regresión

La regresión es una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

Regresión lineal simple. Sólo se trata de una variable regresora

$y = \beta_o + \beta_1 x + e$

Regresión lineal múltiple. En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo:

$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$

Aplicaciones

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

Metodología de la partición de datos

Elementos para hacer un buen modelo de predicción

Elementos previos

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

Dividir los datos

- 70% Conjunto de entrenamiento
- 15% Conjunto de validación
- 15% Conjunto de pruebas.

Arboles 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 determinar las subregiones de nuestro espacio muestral se deben tomar en cuenta las siguientes reglas para poder tener solamente datos de la misma clase. Los árboles se pueden clasificar en dos tipos que son:

1. Regresión: variable respuesta **y** es cuantitativa.
2. Clasificación: variable respuesta **y** es cualitativa.

La lectura de los árboles de decisión es hacia abajo y está formada por diferentes nodos. Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

✓ Primer nodo o nodo raíz

✓ Nodos internos o intermedios

✓ Nodos terminales u hojas

Arboles de clasificación. Consiste en hacer preguntas del tipo $¿xk \leq c?$ para las covariables cuantitativas o preguntas del tipo $¿xk = nivelj?$

La información de cada nodo es la siguiente:

➢ Condición: Si es un nodo donde se toma alguna decisión.

➢ Gini: Es una medida de impureza. A continuación, veremos cómo se calcula.

➢ Samples: Número de muestras que satisfacen las condiciones necesarias para llegar a este nodo.

➢ Value: Cuántas muestras de cada clase llegan a este nodo.

➢ Class: Qué clase se les asigna a las muestras que llegan a este nodo.

Gini. se refiere a la probabilidad de cada clase. Podemos calcularla dividiendo el número de muestras de cada clase en cada nodo por el número de muestras totales por nodo. ($gini = 1 - \sum_{k=1}^{n} {p_c}^2$)

Árbol de regresión. preguntas de tipo ¿$xk \leq c$?

Bagging. Una forma de mejorar un modelo predictivo es usando la técnica creada por Leo Breiman que denominó Bagging (o Bootstrap Aggregating). Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

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.

METRICAS DE EFICACIA

- Error cuadrático medio
- Curva roc

Clustering

Es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes.

- Investigación de mercado
- Identificar comunidades
- Prevención de crimen
- Procesamiento de imágenes.

TRANSFORMACIÓN DE DATOS.

- Variables cuantitativas
- Variables Binarias
- Variables categóricas

TIPOS BÁSICOS DE ANÁLISIS.

- Centroid Based Clustering: Los clusters se construyen basados en la distancia de punto de los datos hasta el centroide. (Se usa el algoritmo K-medias)
- Connectivity Based Clustering: La característica principal es que un cluster contiene a otros clusters (representan una jerarquía) (Se usa el algoritmo Hierarchical clustering).
- Distribution Based Clustering: En este método cada cluster pertenece a una distribución normal (Se usa el algoritmo Gaussian mixture models).
- Density Based Clustering: Se trata de conectar puntos cuya distancia entre sí es considerada pequeña.

PASOS K-MEDIAS.

1. *CENTROIDES.* Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster
2. *DISTANCIAS.* Analizamos la distancia de cada dato al centroide más cercano, perteneciendo a su cluster.
3. *MEDIA.* Obtener media de cada cluster y este será el nuevo centro.
4. *ITERAR.* Repetimos el proceso hasta que los clusters 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 0.

MÉTODO DEL CODO.

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

Visualización de datos

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.

Patrones secuenciales

Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias. Es una clase especial de dependencia en las que el orden de acontecimientos es considerado. Son eventos que se enlazan con el paso del tiempo.

- "si sucede el evento X en el instante de tiempo t entonces sucederá el evento Y en el instante t+n".
- Describen de forma concisa relaciones temporales que existen entre los valores 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 itemsets, donde cada itemset es un elemento de la secuencia.
- El tamaño de una secuencia es su cantidad de elementos (itemsets).
- 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.
- Las secuencias frecuentes (o patrones secuenciales) son las subsecuencias de una secuencia que tienen un soporte mínimo.

Resolución de problemas

- Agrupación de patrones secuenciales
- Clasificación de datos secuenciales
- Reglas de asociación con datos secuenciales

Métodos representativos

- GSP
- SPADE
- AprioriAll
- FreeSpan
- SPAM
- PrefixSpan
- ISM
- IncSp
- ISE
- IncSpan

Método de 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.

Técnicas de clasificación:

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

Redes neuronales:

Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben enumerarse.

- Se usan en Clasificación, Agrupamiento, Regresión
- Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.
- Internamente pueden verse como una grafica dirigida.

Problemas con la inducción de reglas:

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