UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MINERIA DE DATOS

Resúmenes de técnicas de minería de datos

Maestro: Mayra Berrones

Nombre: Matrícula:

Tania Sarahi Rossel Castillo 1810461

Ciudad San Nicolás de los Garza – 02 de octubre del 2020.

PREDICCIÓN

Elementos para un buen modelo de predicción

- Definir adecuadamente nuestro problema
- Recopilar datos
- Elegir una medida o indicador de éxito
- Preparar los datos

Árboles 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 preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Los árboles se pueden clasificar

- Árboles de regresión: en los cuales la variable respuesta y es cuantitativa
    - Consiste en hacer preguntas de tipo ¿$xk \leq c$? para cada una de las covariables, todas las observaciones dentro de un hiper-rectángulo tendrán el mismo valor estimado $\hat{y}$
- Árboles de clasificación en los cuales la variable respuesta y es cualitativa.
    - Consiste en hacer preguntas del tipo ¿$xk \leq c$? para las covariables cuantitativas o preguntas del tipo ¿$xk = nivelj$? para las covariables cualitativas

La estructura básica de un árbol de decisión es por diferentes tipos de nodos:

- Primer nodo o nodo raíz: se produce la primera división
- Nodos internos o intermedios: vuelven a dividir el conjunto de datos en función de variables
- Nodos terminales u hojas: se ubica en la parte inferior del esquema, indica la clasificación definitiva
    - 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»

La información de cada nodo es la siguiente:

- Condición: Si es un nodo donde se toma alguna decisión.
- Gini: Es una medida de impureza, cuando vale 0 es totalmente puro
    - $gini = 1 - \sum_{k=1}^{n} p_c^2$

- 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

Bosques aleatorios

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

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 de los modelos no-paramétricos:

- 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 cruzada

El model assessment se emplea para estimar el test error rate de un modelo y evaluar su capacidad predictiva. También se puede emplear para seleccionar el nivel de flexibilidad adecuado.

Métricas de eficacia.

- Error cuadrático medio: Diferencia entre el estimador y lo que se estima
- Curva roc: Sirve para conocer el rendimiento global de la prueba

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, ítems o atributos que tienden a ocurrir de forma conjunta.

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.

Aplicaciones

- 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.

Tipos de reglas de asociación

- Asociación Cuantitativa: tipos de valores
  - ❖ Asociación Booleana: Ausencia o presencia de un ítem
  - ❖ Asociación Cuantitativa: Describe asociaciones entre ítems cuantitativos o atributos.
- Asociación Multidimensional: dimensiones de datos
  - ❖ 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: Niveles de abstracción
  - ❖ 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: Número de veces con que A y B aparecen juntos en una base de datos de transacciones $\frac{Frecuencia\ en\ que\ A \cap B\ aparecen\ en\ las\ transacciones}{Total\ de\ transacciones}$ con un soporte bajo pudo haber aparecido por casualidad

- Confianza: Mide la fortaleza de la regla $\frac{Soporte\ (A=>B)}{Soporte\ (A)}$ con una confianza baja 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 $\frac{Soporte\ (A=>B)}{Soporte\ (A)*Soporte\ (B)}$

  Mayor a 1 representa una relación fuerte y frecuencia mayor que el azar
  Igual a 1 representa relación del azar
  Menor a 1 representa relación débil y frecuencia menor que el azar

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.

Principales usos:

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

Tipos básicos de análisis

- Centroid Based Clustering: Cada cluster es representado por un centroide y se construyen basados en la distancia de punto de los datos hasta el centroide. Se realizan varias iteraciones utilizando K-medias
- Connectivity Based Clustering: se definen agrupando a los datos más similares o cercanos, su principal característica es que un cluster contiene a otros. Algoritmo usado 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. Algoritmo usado Gaussian mixture models.
- Density Based Clustering: Los clusters son definidos por áreas de concentración. Se trata de conectar puntos cuya distancia entre sí es considerada pequeña. Los cluster que se encuentran fuera de los puntos relacionados dentro de una distancia limitada es considerada como irregular

Método de K-medias

K representa el número de clusters y es definido por el usuario, después se eligen k datos aleatorios que serán los centroides representativos para cada cluster, así mismo se analiza la distancia de cada dato al centroide más cercano, después se obtiene media de cada cluster y este será el nuevo centro y este proceso se repetirá hasta que los cluster no cambien.

Varianza de los clusters

- Disminuye al aumentar K
- Si solo hay un elemento en el cluster la varianza es cero.
- Entre menor 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 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

Es la representación gráfica de información y datos.

Al utilizar elementos visuales proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos.

También es esencial para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Existen multitud de técnicas y aproximaciones para la visualización según sea la naturaleza del dato de la información. Según la complejidad y elaboración de la información podemos tener la siguiente clasificación

Tipos de visualizaciones

- Elementos básicos de representación de datos
  - ❖ 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.
- 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.
- Infografías: Se utilizan para contar “historias” se construye a través símbolos, leyendas, dibujos, imágenes sintéticas, etc.

Algunos Software

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

Importancia de la visualización de datos

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

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.

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

Tipos de regresiones lineales:

Regresión lineal simple

Cuando el análisis de regresión solo cuenta con una variable regresora, tiene como modelo $y = \beta_0 + \beta_1 x + e$

Regresión lineal múltiple

Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos.

Se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo: $y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k + e$

La cantidad 'e' en la ecuación es una variable aleatoria normalmente distribuida con E(e)=0 y Var(e)= σ^2

Aplicaciones

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

CLASIFICACIÓN

Es la técnica más comúnmente aplicada, que organiza o mapea un conjunto de atributos por clase dependiendo de sus características.

Funciones

Se estima un modelo usando los datos recolectados para hacer predicciones futuras.

Técnicas de clasificación

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

Regla de Bayes

Si tenemos una hipótesis H sustentada para una evidencia $E \rightarrow P(H|E) = \frac{P(E|H)*P(H)}{P(E)}$

Donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B

Redes Neuronales

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 en tres capas:
    - De entrada
    - Oculta
    - De salida.
- Internamente pueden verse como una gráfica dirigida.

Árboles de Decisión

Son una serie de condiciones organizadas en forma jerárquica

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.

PATRONES SECUENCIALES

Pertenecen a la categoría de Predictivas, se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, el orden de acontecimientos es considerado, también son eventos que se enlazan con el paso del tiempo.

Para los patrones secuenciales

- 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”.
- 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.
- Utiliza reglas de asociación secuenciales, reglas que expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos en el tiempo.

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 una cantidad de elementos
- La longitud de una secuencia es una cantidad de ítems
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S
- Los patrones secuenciales son subsecuencias de una secuencia que tienen un soporte mínimo

Algunas áreas de aplicación

- Medicina
- Análisis de mercado, distribución y comercio
- Aplicaciones financieras y banca
- Aplicaciones de seguro y salud privada

Tipos de base de datos

- Temporales
- Documentales
- Relacionales

Resolución de problemas

- Agrupamiento de patrones secuenciales: Se define como 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.
    - Cada M patrones se mezclan agrupamientos pueden ser:
        - Mezcla por cercanía
        - Mezcla por tamaño
        - Mezcla forzada
- Clasificación con datos secuenciales: Expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo
- Reglas de asociación con datos secuenciales: Se presenta cuando los datos contiguos presentan algún tipo de relación

DETECCIÓN DE OUTLIERS

Son los valores que se “escapan al rango en donde se concentran la mayoría de las muestras”. Según Wikipedia son las muestras que están distantes de otras observaciones, y el objetivo de esto es localizar las anomalías

Es importante detectar los outliers debido a que pueden afectar considerablemente los resultados que pueda obtener un modelo de machine learning

Los Outliers pueden significar varias cosas:

- ERROR: Si tenemos un grupo de “edades de personas” y tenemos una persona con 160 años, seguramente sea un error de carga de datos. En este caso, la detección de outliers nos ayuda a detectar errores.
- LIMITES: En otros casos, podemos tener valores que se escapan del “grupo medio”, pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo de ML.
- Punto de Interés: puede que sean los casos “anómalos” los que queremos detectar y que sean nuestro objetivo.

Puede haber variedades de outliers desde 1 hasta n dimensiones

Una grafica de detección de outliers sencilla son los Boxplot

Una vez detectados los outliers según la lógica de negocio podemos actuar de una manera u otra.

Por ejemplo, podríamos decidir:

- Las edades fuera de la distribución normal, eliminar.
- El salario que sobrepasa el límite, asignar el valor máximo (media + 2 sigmas).
- Las compras mensuales, mantener sin cambios.

Python

Se puede utilizar una media conocida para la detección de outliers que es la media de la distribución más 2 sigmas como frontera. Pero existen otras estrategias para delimitar outliers.

Una librería muy recomendada es PyOD. Posee diversas estrategias para detectar Outliers. Ofrece distintos algoritmos, entre ellos Knn que tiene mucho sentido, pues analiza la cercanía entre muestras, PCA, Redes Neuronales