Nombre: Miguel Alejandro Noriega Ortega #1798528

Método clasificación

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

Existen distintos tipos de técnicas de clasificación como lo son:

Clasificación por inducción de árbol de decisión: Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol, Útiles para problemas que mezclen datos categóricos y numéricos y Útiles en Clasificación, Agrupamiento y Regresión

Clasificación Bayesiana: El teorema de Bayes, en la teoría de la probabilidad, expresa la probabilidad condicional de un evento aleatorio A dado B en términos de la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de solo A.

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 y Regresión

Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida. E internamente pueden verse como una grafica dirigida.

Support Vector Machines (SVM) : Las máquinas de vectores de soporte o máquinas de vector soporte, son un conjunto de algoritmos de aprendizaje supervisado, Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

Nombre: Miguel Alejandro Noriega Ortega #1798528

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.

CENTROID BASED CLUSTERING

Cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia de 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.

MÉTODO K-MEDIAS

Algoritmo de clustering basado en centroides. K representa el número de clusters y es definido por el usuario.

Una vez que escogemos el valor de k:

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

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 numero de k a utilizar

Nombre: Miguel Alejandro Noriega Ortega #1798528

Outliers

Son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por:

a) Errores de procedimiento.

b) Acontecimientos extraordinarios.

c) Valores extremos. Por ejemplo, una muestra de datos del número de cigarrillos consumidos a diario contiene el valor 60 porque hay un fumador que fuma sesenta cigarrillos al día.

d) Causas no conocidas.

Los datos atípicos distorsionan los resultados de los análisis, y por esta razón hay que identificarlas y tratarlos de manera adecuada, generalmente excluyéndolos del análisis.

Para cada grupo de registros, o para un conjunto completo de registros, se utiliza la desviación estándar de un campo numérico específico o un múltiplo de la desviación estándar para establecer los límites superior e inferior de los valores atípicos.

Todos los registros con un valor en el campo numérico que sea superior al límite superior, o inferior al límite inferior, se consideran valores atípicos y se incluyen en los resultados de la salida.

La desviación estándar es una medida de la dispersión de un conjunto de datos; es decir, cuán dispersos están los valores. El cálculo de valores atípicos utiliza la desviación estándar de la población.

Nombre: Miguel Alejandro Noriega Ortega #1798528

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, 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.

Los Problemas que resuelve son:

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.

Clasificación con datos secuenciales: Se presenta cuando los datos contiguos presentan algún tipo de relación.

Reglas de asociación: Éstos expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo.

Nombre: Miguel Alejandro Noriega Ortega #1798528

Predicción

Árbol de decisión: Modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.

Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

✓ Primer nodo o nodo raíz: en él se produce la primera división en función de la variable más importante.

✓ Nodos internos o intermedios: tras la primera división encontramos estos nodos, que 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.

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

Árbol de regresión: Consiste en hacer preguntas de tipo ¿$xk \leq c$? 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 $\hat{y}$

Random Forest: 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.

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.

Nombre: Miguel Alejandro Noriega Ortega #1798528

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 "con un antecedente y 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 y 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.

Nombre: Miguel Alejandro Noriega Ortega #1798528

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: Cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple. La regresión lineal simple tiene como modelo: $\mathbf{y = \beta_o + \beta_1 x + e}$

Estimación por mínimos cuadrados La estimación de $y = \beta_o + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza:

$$\widehat{\boldsymbol{B_0}} = \overline{\boldsymbol{y}} - \widehat{\boldsymbol{B_1}}\boldsymbol{x}$$

Regresión Lineal Múltiple Un modelo de regresión múltiple 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:

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

Nombre: Miguel Alejandro Noriega Ortega #1798528

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.

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 drilldown)

•*Tablas:* con anidación, dinámicas, de drilldown, de transiciones, etc.

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.

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".

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.