Clasificación

Es la organización o mapeo de un conjunto de atributos por clase dependiendo de sus características, conociendo las características de los datos podemos hacer predicciones a futuro. Los modelos de clasificación pueden ser clasificados dependiendo de la certeza en la predicción, y existen cuatro casos: decir que es verdadera cuando es verdadera, decir que es verdadera cuando es falsa, decir que es falsa cuando es verdadera y decir que es falsa cuando es falsa, de aquí las dos predicciones incorrectas son llamadas Error tipo I o falso positivo, donde se dice que es verdadera, cuando es falsa y el Error tipo II o falso negativo, donde se dice que es falsa cuando es verdadera. Otros conceptos importantes son la certeza, denotada por el número de predicciones correctas entre el número de predicciones y su complemento llamado tasa del error.

- Redes neuronales: son redes que tienen la apariencia de neuronas. Se componen de una capa de entrada, una capa de salida y una capa oculta, en estas capas se encuentran ciertos nodos virtuales donde ocurre la clasificación. Su funcionamiento se basa en las conexiones, y su aprendizaje se da con las repeticiones hasta que el algoritmo encuentre la salida deseada y una vez que el algoritmo haya aprendido del problema la predicción será mejor.

- Árboles de decisión: son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos. Las principales desventajas son que las reglas no necesariamente forman un árbol, también puede que n se lleguen a cubrir todas las posibilidades y las reglas pueden entrar en conflicto.

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
$y = \beta 0 + \beta 1x + e$

- 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:
$y = \beta 0 + \beta 1x + \beta 2x2 + \cdots + \beta kxk + e$

La cantidad ‘e’ en la ecuación es una variable aleatoria normalmente distribuida con E(e)=0 y Var(e)= σ2

Aplicaciones

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

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 gráfica 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.

Patrones secuenciales

Es una clase espacial de dependencia en la que el orden de los acontecimientos es considerado, son eventos que se relacionan con el paso del tiempo. Se trata de buscar asociaciones de la forma "si sucede el evento x en el instante del tiempo t entonces sucederá el evento y en el instante t+n". Su objetivo 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, las cuales expresan patrones de comportamiento secuencial, es decir, que se dan en instantes distintos de tiempo.

Sus características son: que el orden importa, el tamaño de una secuencia es su cantidad de elementos (itemset), 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 y las secuencias frecuentes son subsecuencias de una secuencia que tiene soporte mínimo.

Agrupación de patrones secuenciales: separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre si, y al mismo tiempo sean diferentes a los objetivos de otros grupos, parecido al Centroide Base Clustering visto en el tema de Clustering.

Reglas de asociación con datos secuenciales: se presentan cuando los datos contiguos presentan algún tipo de relación, expresan patrones de comportamiento secuenciales, es decir, que se den en instantes distintos, pero cercanos, de tiempo. Tiene relación con el tema reglas de asociación visto anteriormente.

Visualización

Es la representación gráfica de información y datos, proporcionan una manera accesible de ver y comprender tendencias, valores y patrones de los datos, sobre todo cuando lo que necesitamos analizar son grandes cantidades de información y tomar decisiones basadas en estos datos. Se dividen en tres categorías: los elementos básicos en la representación de datos, como gráficas de líneas, de barras, puntos, mapas y tablas; los cuadros de mando, que son composiciones complejas de visualizaciones individuales que guardan una relación temática entre ellas; y por último las infografías, que son utilizadas para contar historias. Los estándares principales para la visualización de datos son: HTML5, CSS3, SCV y WebGL

. El poder mostrar los datos que tenemos de manera visual es muy importante hoy en día, ya que cada vez es más común que las empresas utilicen sus bases de datos para la toma

de decisiones y por medio de estas representaciones gráficas podemos llegar a saber el quien, que, donde, cuando y porque de cualquier conjunto de datos y la problemática que se quiera resolver con ellos. La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

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 o 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. o 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 o Nodos de decisión: tienen una condición al principio y tienen más nodos debajo de ellos o Nodos de predicción: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo»

Clustering

Es una técnica de aprendizaje de maquina no supervisada, esto quiere decir que la máquina podrá aprender por medio de los datos que le demos y que sea no supervisada se refiere a que no hay una interpretación de los datos por parte de una persona, que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes, a los subconjuntos creados por esta técnica se le denomina clúster, cada clúster está conformado por datos que comparten características específicas que a su vez son diferentes entre cada clúster.

-Tipos Básicos de Análisis

Centroid Base Clustering: Cada clúster es representado por un centroide y son construidos en base a la distancia del punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado. El algoritmo más utilizado es el de k-medias.

Conectivity Base Clustering: Los clústeres se definen agrupando a los datos más similares o cercanos, los puntos más cercanos están más relacionados que otros puntos más lejanos. Su característica principal es que el clúster contiene a otros clústers, representando así una jerarquía. El algoritmo utilizado es el Hierarchical Clustering.

Reglas de Asociación

Las reglas de asociación son 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, definido como: “si A(antecedente) entonces B(consecuencia)” o “A => B”, en donde A y B son ítems individuales.

Las reglas se dividen en diferentes tipos dependiendo de las características que se están tomando en cuenta para realizarlas, estos son: con base en el tipo de valores que manejan las reglas (Booleana o Cuantitativa), con base en las dimensiones de datos que involucra la regla (Unidimensional o Multidimensional) y con base en los niveles de abstracción que involucra la regla (de un nivel o Multinivel). Estas reglas son utilizadas para encontrar las combinaciones de artículos que ocurren con mayor frecuencia dentro de una base de datos y a su vez medir la fuerza e importancia de estas combinaciones, por lo que son aplicadas con diferentes fines como:

- Definir patrones de navegación dentro de una tienda, promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas,

distribución de mercancías en tiendas y segmentación de clientes con base en patones de compra.