Vanessa Ortiz de la Rosa, Matricula: 1810699

Resumen

CLUSTERING

Definición: 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, proceso de agrupar datos en clases o clusters de tal forma que los objetos de un cluster tengan una similitud alta entre ellos.

Usos: Investigación de mercado, Identificar comunidades, prevención del crimen y Procesamiento de imágenes.

Tipos básicos de análisis

Centroid Based Clustering

Los clusters son representados por un centroide, estos se construyen en base a la distancia de los datos al centroide, se itera hasta el mejor resultado.

Connectivity Bases Clustering

Los clusters se definen agrupando los datos más similares, la característica principal es que un cluster contiene otros clusters(Jerarquía).

Distribution Based Clustering

En este método cada cluster pertenece a una distribución normal.

Density Based Clustering

Los clusters son definidos por áreas de concentración, se trata conectan puntos cuya distancia entre si es consideradamente pequeña.

Método-K medias

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

Se escoge el valor de k

- Elegimos k datos aleatorios que pasarán a ser los centroides representativos de cada cluster.
- Analizamos la distancia de cada dato al centroide más cercano, a su cluster.
- Obtener el promedio de cada cluster y este será el nuevo centro.
- Se itera el algoritmo 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 codo y representa el número de k a utilizar.

VISUALIZACION

Es la representación gráfica de información y datos, al utilizarse se proporciona una manera rápida de ver y poder comprender tendencias, datos atípicos e incluso patrones.

Las técnicas de visualización se diferencian según la naturaleza de la información de los datos.

Tipos de visualizaciones

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, grafica de pastel.
- Mapas: burbujas, mapa temático, mapa de calor, de agregación.
- Tablas: con anidación, dinámicas, de transiciones, etc.

Cuadros de mandos: Composición de visualizaciones individuales que sostienen una coherencia y relación de temática entre ellas, se utilizan en organizaciones para análisis de conjuntos de variables y toma de decisiones.

Infografías: Destinadas a las construcciones de narrativas a partir de los datos se utilizan para contar historias, se construye mediante la disposición de información en que las visualizaciones se mezclan con otros elementos como pueden ser símbolos, leyendas, dibujos entre otras.

Importancia de la visualización de datos en cualquier empleo

Es de utilidad poder usar los datos para tomar decisiones y poder tener elementos visuales para contar historias con los datos, la visualización esta justo en el centro del análisis.

OUTLIERS

Detección de valores atípicos, que también suelen denominarse outliers. Detectan la existencia de valores observados que no siguen el mismo comportamiento que los demás. Estos son consecuencia de errores en el procedimiento al introducir los datos, cuando se detectan el analista sigue un criterio de eliminación o una forma de incluirlos.

Categorías

- Aquellas observaciones que provienen de un error de procedimiento, por ejemplo, un error de codificación, error de entrada de datos, etc. Estos datos atípicos, si no se detectan mediante filtrado, deben eliminarse o recodificarse como datos ausentes.
- Aquellas observaciones que ocurren como consecuencia de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Éstos generalmente se retienen en la muestra, salvo que su significancia no sea relevante.
- Datos atípicos comprende las observaciones extraordinarias para las que el investigador no tiene explicación, las cuales normalmente se eliminan del análisis.
- Casos atípicos la forman las observaciones que se sitúan fuera del rango ordinario de valores de la variable. Suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población

¿Cómo detectar los outliers?

Mediante el uso de un histograma, es posible observar dichas observaciones, pero se depende mucho del dominio.

Soluciones

a) Ignorar

b) Eliminar la columna

c) Eliminar la fila

d) Reemplazar el valor por: nulo, máximo o mínimo

d) Hacer que los anómalos sean “muy alto” o “muy bajo”

Aplicación

- 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.

REGLAS DE ASOCIACION

Reglas de asociación: Tipo de análisis que extrae información por coincidencias, su objetivo es encontrar relaciones dentro de un conjunto de transacciones (ítems o atributos que ocurren de forma conjunta)

Se define como:

A => B, donde A y B, ítems individuales

Ventajas de las reglas de asociación

- Seleccionar combinaciones de artículos que ocurren con mayor frecuencia y poder medir la fuerza e importancia de estas combinaciones.

Aplicaciones

- Definir patrones de navegación dentro de la tienda.
- Distribución de mercancías en tiendas.
- Análisis de información de ventas.
- Segmentación de clientes con base en patrones de compra

Tipos de reglas de asociación

Asociación Cuantitativa

- 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

Dada una regla “Si A => B”, la confianza de esta regla es el cociente del

soporte de la regla y el soporte del antecedente solamente

Confianza mide la fortaleza de la regla.

- En lenguaje de probabilidad, confianza es una probabilidad condicional:
- Regla con baja confianza: es probable que no exista relación entre antecedente y consecuente.

REGRESION

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.

Cuando el análisis de regresión sólo se trata de una variable represora, se llama regresión lineal simple.

La regresión lineal simple tiene como modelo:

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

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

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.

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x \quad \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

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 represores, o variables predictivas bajo el modelo:

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

Se realiza la estimación por mínimos cuadrados.

Aplicaciones

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

CLASIFICACION

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

Funciona: Porque 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 -> $p(H|E)= (p(E|H)* p(H))/p(E)$

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

condicionada al suceso B

Redes neuronales

Trabajan directamente con números nominales, estos deben enumerarse.

- Se usan en Clasificación, Agrupamiento, Regresión

Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.

Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol.

Utilidad

- Problemas que mezclen datos categóricos y numéricos.
- Clasificación, Agrupamiento, Regresión

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

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.

El patrón secuencial 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.

Son eventos que se enlazan con el paso del tiempo.

Procedimiento

- 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 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.

Aplicaciones

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

Resolución de problemas

Agrupación 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: Éstos 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.

PREDICCION

Requisitos previos: Definir el problema, recopilar datos, elegir el indicado y preparar los datos

Árbol de decisión

Modelo predictivo divide el espacio de los predictores, agrupa las observaciones con valores similares para la variable respuesta o dependiente, cada subregión contenga el mayor porcentaje posible de los individuos de cada población.

Se clasifican en:

1. Árboles de regresión en los cuales con variable de respuesta cuantitativa
2. Árboles de clasificación en los cuales con variable de respuesta cualitativa

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

- Primer nodo o nodo raíz.
- Nodos internos o intermedios.
- Nodos terminales u hojas.

Árbol de clasificación

Consiste en hacer preguntas del tipo ¿$x_k \leq c$? o ¿$x_k = nivel_j$? para las covariables cuantitativas cualitativas, el espacio de las covariables es divido en hiper-rectángulos y todas las observaciones dentro de un hiper-rectángulo tendrán el mismo valor grupo estimado.

La partición del espacio se hace de manera repetitiva para encontrar las variables y los valores de corte de tal manera que se minimice la función de costos.

Hay dos tipos de nodo:

- Nodos de decisión.
- Nodos de predicción.

Árbol de regresión

Consiste en hacer preguntas de tipo ¿$x_k \leq c$? para cada una de las covariables, el espacio de las covariables se divide en hiper- rectángulos y las observaciones que queden dentro del mismo, tendrán el mismo valor estimado.

La partición

1. Dado un conjunto de covariables (características), encontrar la covariable que permita predecir mejor la variable respuesta.
2. Encontrar el punto de corte sobre esa covariable que permita predecir mejor la variable respuesta.

Bosques aleatorios

Técnica de aprendizaje basada en arboles de decisión, tiene como ventaja ofrecer un mejor rendimiento de generalización la cual se consigue compensado los errores de las predicciones de los árboles de decisión, con la finalidad de asegurarnos de que los arboles sean distintos, cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging.