UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
Facultad de Ciencias Físico-Matemáticas

Minería de Datos
Resúmenes

ALUMNO: Cristian Antonio Jaramillo Arriaga 1680776.

GRUPO: 012.

PROFESOR: Mayra Cristina Berrones Reyes.

25 de septiembre de 2020.

ÍNDICE

Regresión Lineal

En la regresión buscamos una variable aleatoria simple digamos Y, en teoría el valor de esta variable aleatoria está influenciado por los valores tomados por una o más variables.

Y se denomina como: “Variable Dependiente” o “Respuesta”

Las variables influyentes: “Variables Independientes”, “Variable Predictoras” o regresoras.

En el caso de la regresión lineal asumimos que Y(costo) es una función lineal de x (superficie) y entonces el modelo lineal se escribe como

$$Ye = \alpha + \beta * x$$

$$\text{Alquiler mensual} = \alpha + \beta * \text{Superficie}$$

Con datos históricos, podríamos crear un modelo lineal y obtener las posibles valores de alpha y Beta.

$\alpha = 86.96, \beta = 2.37$

Modelo lineal $Ye = 86.96 + 2.37 * x$

Usando esta ecuación, podremos encontrar el alquiler de cualquier casa, por ejemplo una de 110 m2:

$$Ye = 86.96 + 2.37 * 110 = 347.66$$

Reglas de Asociación

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, *ítems* o atributos que tienden a ocurrir de forma conjunta. Por ejemplo:

- La cesta de la compra en un supermercado.
- Los libros que compra un cliente en una librería.
- Las páginas web visitadas por un usuario.
- Las características que aparecen de forma conjunta.

A cada uno de los eventos o elementos que forman parte de una transacción se le conoce como ítem y a un conjunto de ellos itemset. Una transacción puede estar formada por uno o varios items, en el caso de ser varios, cada posible subconjunto de ellos es un itemset distinto. Por ejemplo, la transacción T = {A,B,C} está formada por 3 items (A, B y C) y sus posibles itemsets son: {A,B,C}, {A,B}, {B,C}, {A,C}, {A}, {B} y {C}.

Apriori

Fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, tiene dos etapas:

- Identificar todos los *itemsets* que ocurren con una frecuencia por encima de un determinado límite (*itemsets* frecuentes).
- Convertir esos *itemsets* frecuentes en reglas de asociación.

Clustering

Es una técnica dentro de la disciplina de Inteligencia Artificial, identifica de manera automática agrupaciones (o clústeres de elementos) de acuerdo a una medida de similitud entre ellos.

Métricas de distancia

Una métrica de distancia es una función $\boldsymbol{d}\ (\boldsymbol{x}, \boldsymbol{y})$ que especifica la distancia entre elementos de un conjunto de números reales no negativos.

Dos elementos son iguales bajo una métrica particular si la distancia entre ellos es cero.

Las funciones de distancia representan un método para calcularla cercanía entre dos elementos.

Algoritmo k-means

En el algoritmo k-means, n objetos se agrupan en k agrupaciones en función de características, donde k < n y k es un número entero positivo.

La agrupación de objetos se realiza minimizando la suma de cuadrados de distancias, es decir, una distancia euclidiana entre los datos y el centroide del grupo correspondiente.

Outliers

Es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.

Método de detección: Desviación estándar

En estadística, si una distribución de datos es aproximadamente normal, aproximadamente el 68% de los valores de los datos se encuentran dentro de una desviación estándar de la media y aproximadamente el 95% están dentro de dos desviaciones estándar, y aproximadamente el 99,7% se encuentran dentro de tres desviaciones estándar.

Método de detección: Boxplots

Los diagramas de caja son una representación gráfica de datos numéricos a través de cuantiles. Es una forma muy simple pero efectiva de visualizar valores atípicos. Los bigotes inferiores y superiores pueden verse como los límites de la distribución de datos. Cualquier punto de datos que se muestre por encima o por debajo de los bigotes, puede considerarse atípico o anómalo.

Método de detección: DBScan Clustering

DBScan es un algoritmo de agrupación en clústeres que utiliza datos agrupados en grupos. También se utiliza como un método de detección de anomalías basado en la densidad con datos unidimensionales o multidimensionales. También se pueden utilizar otros algoritmos de agrupación como k-medias y agrupación jerárquica para detectar valores atípicos.

Predicción

El análisis predictivo consiste en la tecnología que aprende de la experiencia para predecir el futuro comportamiento de individuos para tomar mejores decisiones.

Modelo predictivo

Se podrá utilizar para predecir qué probabilidades hay de que una persona –en función de los datos que se disponga de la misma– reaccione de una manera determinada (si comprará un producto, si cambiará de voto, si contratará un servicio…). Una vez introducidos los datos de la persona y se aplique el modelo predictivo se obtendrá una calificación que indicará la probabilidad de que se produzca la situación estudiada por el modelo.

Técnicas aplicables al análisis predictivo

Técnicas de regresión

- regresión lineal.
- Árboles de clasificación y regresión.
- Curvas de regresión adaptativa multivariable.

Técnicas de aprendizaje computacional

- Redes neuronales.
- Máquinas de vectores de soporte.
- Naïve Bayes.
- K-vecinos más cercanos.