UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO – MATEMÁTICAS

Minería de Datos

Resúmenes: Técnicas de Minería de datos

ALUMNO: JESÚS EDUARDO VALENCIA GONZÁLEZ

MAESTRA: Mayra Cristina Berrones Reyes

MATRÍCULA: 1630606

MONTERREY, NL, A 30 DE SEPTIMBRE 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”

En el caso de una regresión lineal asumimos que Y es una función lineal de x, y entonces el

modelo lineal se escribe como: Ye = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}$ * x.

Y la diferencia entre el valor real y el estimado se puede escribir como:

ei = (Yi - Ye(Xi)).

El objetivo es minimizar la suma de los errores al cuadrado sobre todos los puntos del data set:

X = {(Xi,Yi)}.

Error estándar residual: RSE es la desviación estándar del término del error (desviación de la parte de datos que el modelo no es capaz de explicar por falta de información o datos

adicionales).

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.

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}.

Soporte (Frecuencia relativa): Dada regla si A=>B, el soporte de esta se define como el númerode veces o la frecuencia relativa con que A y B aparecen juntos en una BDD Transaccional.

Confianza (Probabilidad empírica): Dada una regla si A=>B, la confianza de esta regla es el cociente de soporte de la regla y soporte del antecedente solamente. Confianza (A=>B) = Soporte (A=>B)/Soporte (A) Si el soporte mide la frecuencia, Confianza mide la fortaleza de la regla. Confianza (A=>B) = P(B/A).

Lift (A→B) = Soporte(A→B)/(Soporte (A) * Soporte(B)) , si Lift = 1 o muy cerca a 1, indica que la relación es producto del azar de lo contrario, indica que la relación es realmente fuerte.

Apriori: uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación 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.

Outliers

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

Los métodos de detección: Se pueden dividir en univariados y multivariados. Los multivariantes son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Son más difíciles de identificar que los outliers unidimensionales, dado que no pueden considerarse “valores extremos”, como sucede cuando se tiene una única variable bajo estudio.

Método de detección - desviación estándar: Si se tiene algún punto de datos que sea más de 3 veces la desviación estándar, es muy probable que esos puntos sean anómalos o atípicos.

Método de detección - boxplots: Los diagramas de caja son una representación gráfica de datos numéricos a través de cuantiles. 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: Core Points , min_samples es simplemente el número mínimo de puntos centrales necesarios para formar un grupo, eps es la distancia máxima entre dos muestras para que se consideren como en el mismo grupo. Border Points , se encuentran en el mismo grupo que los puntos centrales, pero mucho más lejos del centro del grupo.

Predicción

Consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento.

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.

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.

Distancia euclídea

Este tipo de distancia es usada principalmente para calcular distancias. La distancia entre dos puntos en el plano con coordenadas (x, y) y (a, b).

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.

Distancia de Manhattan

Este tipo de distancia es definida como la suma de las longitudes de las proyecciones del segmento de línea entre los dos puntos en los ejes de coordenadas.

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.

Clustering jerárquico

El clustering jerárquico puede realizarse tanto en forma divisiva o aglomerativa, y permite analizar alternativas para distintos números de grupos.

Clustering jerárquico aglomerativo

Se comienza con tantos clusters como individuos y consiste en ir formando grupos según su similitud.

VISUALIZACIÓN DE DATOS

Tipos de visualización de datos:

- **Gráficos:** Para representar datos de manera sencilla, como gráficos circulares, líneas, columnas, barras, burbujas, diagramas de dispersión y mapas de tipo árbol.
- **Mapas:** La herramienta más conocida para la visualización de mapas es Google maps.
- **Infografías:** ayudan a procesar más fácil la información compleja.
- **Cuadros de Mando (Dashboards):** es una herramienta que permite saber en todo momento el estado de los indicadores del negocio.

Las aplicaciones que podemos entender de la visualización de datos son:

- Comprender la información con rapidez Mediante el uso de representaciones gráficas de información de negocios, las empresas pueden ver grandes cantidades de datos de formas claras y cohesivas y sacar conclusiones a partir de esa información.

- **Identificar relaciones y patrones.** Incluso muy grandes cantidades de datos complicados comienzan a tener sentido cuando se presentan de manera gráfica; las empresas pueden reconocer parámetros con una correlación muy estrecha.
- **Identifique tendencias emergentes.** El uso de la visualización de datos para descubrir tendencias en los negocios y en el mercado puede dar a las empresas una ventaja sobre la competencia, y eventualmente tener un impacto en la base de operación.

PATRONES SECUENCIALES

En minería de datos secuenciales, los cuales es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el tiempo. Reglas de asociación secuencial representan patrones en distintos lapsos del tiempo.

El objetivo de los patrones secuenciales es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

Características:

▫ El orden importa

▫ Objetivo: 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 items

▫ 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

Los patrones que se suelen ver para entender al hacer ejercicios es la siguiente:

- lsl es el número de elementos de una secuencia.
- Una k-secuencia es una secuencia de k eventos
- Una subsecuencia es una secuencia que está dentro de otra. Pero se deben de cumplir ciertas normas el cual es el “orden”.

Clasificación

La clasificación se encarga de predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos, lo cuales "la clasificación" esta dentro de 4 grandes subramas de predictivo, los cuales los otros 3 son.

- Predicción
- Regresión
- Patrones secuenciales

Algunos métodos de clasificación son:

- Reglas de clasificación: buscan términos no clasificados de forma periódica. Un ejemplo de este método son los algoritmos usados en YouTube donde aparece contenido relacionado dado las búsquedas previas.
- Analisis discriminante: método usado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos. Es uno de los métodos más sencillos y fáciles de comprender. Un ejemplo de este sería la separación de colores o calificaciones.
- Redes neuronales artificiales: es un modelo de unidades conectadas para transmitir señales. Este método es parecido a los árboles de decisión solo que a diferencia este tiene más respuestas o caminos de decisiones a tomar, es decir, que una sola decisión puede tener múltiples caminos.
- Árboles de decisión: método analítico que facilita la forma de decisiones. Los árboles de decisiones solo tienen una respuesta o un solo camino a seguir.

Las características en clasificación son:

1. Precisión en la predicción
2. Eficiencia
3. Robustez
4. Escalabilidad
5. Interpretabilidad