Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Práctica 08. Preprocesamiento de Datos.

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495
Luis Enrique García Gomez	315063880

1. Contesta las siguientes preguntas:

- ¿Es posible tener una $X^2 \leq 0$? ¿En que caso podría presentarse?

  Si es posible, ya que puede presentarse el caso cuando las frecuencias observadas todas *iguales o menores* que las frecuencias esperadas en cada celda de la **tabla de contingencia**; por lo que en este caso, el valor de X^2 sería negativo o cero, indicando de esta forma que no hay *desviación* entre las observaciones y las expectativas.

- ¿En que consiste una regresión? ¿Que tipos de regresión hay?

  Consiste en una una técnica utilizada para modelar la relación entre una variable *dependiente* y una o más variables *independientes*. Su objetivo es encontrar una *función matemática* que pueda predecir el valor de la *variable dependiente en función de las variables independientes*.
  Y tenemos los siguientes tipos de regresión:

  - **Regresión lineal simple**: Es el tipo más utilizado y sencillo de regresión, el se usa cuando hay una única *variable independiente* que se utiliza para predecir una variable dependiente.
  - **Regresión lineal múltiple**: Se utiliza cuando hay dos o más *variables independientes* para poder predecir una *variable dependiente*; donde la relación entre las variables se modela como un plano en un espacio de mayor dimensión.
  - **Regresión no lineal**: Se utiliza cuando la relación entre las variables no puede ser modelada de manera *lineal*. Y en este caso, se utilizan *funciones no lineales* para modelar la relación entre las variables.
  - **Regresión logística**: Se utiliza cuando la variable dependiente es *categórica* y se desea predecir la probabilidad de que una observación pertenece a una categoría específica. Aquí se usa una función logística para poder modelar la relación entre las *variables independientes* y la probabilidad de pertenecer a una categoría.
  - **Regresión polinómica**: Se utiliza cuando la relación entre las *variables independientes* y las *variables dependientes* no es lineal, sino que sigue una forma *polinómica*. Aquí se ajusta un polinomio de grado superior a los datos para capturar la relación no lineal.
  - **Regresión de Poisson**: Se utiliza cuando la *variable dependiente* es una frecuencia y se desea modelar la *tasa de ocurrencia* de un evento.
  - **Regresión de Cox**: Se utiliza en el *análisis de supervivencia* para modelar el tiempo hasta que ocurra un evento, como la muerte o el fracaso de un dispositivo.

2. Deberán hacer un resumen sobre el pdf, lectura01, sobre alguno de los ejemplos que se presentan.

Este artículo se enfoca en el preprocesamiento de datos para Big Data, y describe las tecnologías, herramientas y técnicas existentes para llevar a cabo esta tarea, destacando que el preprocesamiento de datos es crucial para garantizar la calidad de los datos y mejorar la eficacia de los algoritmos de análisis de datos.

El artículo se divide en varias secciones, comenzando con una introducción a la importancia del preprocesamiento de datos en el contexto del Big Data, luego, se describen las tecnologías y herramientas disponibles para el procesamiento de datos masivos, y se presentan algunas técnicas de preprocesamiento de datos, como la selección de atributos y la reducción de dimensionalidad.

Uno de los ejemplos presentados en el artículo es un caso de uso en el que se aplica un algoritmo de selección de atributos escalable sobre un problema de alta dimensionalidad, el objetivo de este ejemplo es demostrar cómo los algoritmos de preprocesamiento pueden mejorar la eficacia de los algoritmos de análisis de datos al reducir la complejidad del problema.

En resumen, este artículo es una guía completa sobre el preprocesamiento de datos para Big Data, y proporciona información valiosa sobre las tecnologías, herramientas y técnicas disponibles para llevar a cabo esta tarea, los ejemplos presentados en el artículo ilustran cómo los algoritmos de preprocesamiento pueden mejorar la eficacia de los algoritmos de análisis de datos y ayudar a los científicos de datos a extraer información valiosa de los datos masivos.

3. Las puntuaciones obtenidas por un grupo de alumnos en una batería de test que mide la habilidad verbal (X) y el razonamiento abstracto (Y) son los siguientes:

22 > Y/X	22>20	22>30	22>40	22>50
22 > (25 - 35]	6	4	0	0
22 > (35 - 45]	3	6	1	0
22 > (45 - 55]	0	2	5	3
22 > (55 - 65]	0	1	2	7

- ¿Existe correlación entre ambas variables?

Para responder esta pregunta obtengamos la matriz de correlación de nuestros datos, esto es

	vector1	vector2	vector3	vector4
vector1	1.0000000	0.7139329	-0.9073929	-0.7504788
vector2	0.7139329	1.0000000	-0.4841820	-0.6888748
vector3	-0.9073929	-0.4841820	1.0000000	0.4120210
vector4	-0.7504788	-0.6888748	0.4120210	1.0000000

Ahora, observemos que nuestros valores resultantes oscilan entre -0.907 y 0.714. Dado que ninguno de los valores se acerca a 1 o -1[1], podríamos decir que la magnitud de la correlación es moderada.

- Según los datos de la tabla, si uno de los alumnos obtiene una puntuación de 70 puntos en razonamiento abstracto, ¿En cuanto se estimara su habilidad verbal?

La regresión lineal utilizada para predecir el valor de la habilidad verbal ha sido de entre 1 y -200, por lo que podemos concluir que los datos son insuficientes para realizar la predicción.

[1]Recordar que si nuestros valores se acercan a 1 tenemos una correlación fuerte, en otro caso hay una correlación débil.