Reporte de Laboratorio Nro. 4

Victoria Mendoza[L00379393] and Yandry Montaño[L00382537]

Universidad de las Fuerzas Armadas
mvmendoza5@espe.edu.ec
ydmontano@espe.edu.ec

Tema: Regresión Lineal

Resumen

Tanto en minería de datos, como en cualquier tema, la manera de resolver un problema siempre va a depender de las características que este aporte. Existen varios modelos matemáticos que se usan para predicción y modelado de datos. En este documento, se describe uno de los modelos más básicos, como lo es la regresión lineal, el cual permite hacer predicciones sobre la relación entre variables dependientes e independientes. Así pues, se analiza un ejemplo de una función básica para los valores de la variable dependiente y los valores aleatorios para la variable independiente. Como resultado, se tiene el análisis de la figura con los puntos ubicados en el plano, conocido como "diagrama de dispersión" y se encuentra la función de relación lineal que encuentra la menor distancia entre la recta y los puntos.

1. Introducción

La regresión lineal es uno de los modelos matemáticos más básicos en minería de datos. Este análisis predictivo permite hacer estimaciones entre la relación de una variable dependiente y una o algunas variables independientes. La forma más básica de la regresión lineal se define en una función con solo una variable independiente y una dependiente. En el presente documento, se explica un ejemplo básico de un modelo de regresión lineal. Para ello, se usa una función sencilla para la variable dependiente y números aleatorios para la variable independiente.

La regresión lineal, al ser un modelo sencillo, permite que su implementación fácil. Aún así, no siempre los resultados se asemejan tanto a la realidad y generalmente tienen baja precisión. A pesar que se considera un modelo sencillo, no es conveniente resolver todos los casos con regresión lineal. A veces, se necesita de funciones mucho más complejas, debido a que, al ser una relación lineal entre variables, una función lineal no siempre satisface los datos más complejos.

2. Método

Para el desarrollo de este laboratorio, se hace uso de las librerías matplotlib, scipy y numpy. La librería Matplotlib, se usa en el último comando, que permite realizar la gráfica del modelo que se está utilizando. La librería scipy, se la usa para llamar a la función de regresión lineal y por último, NumPy sirve para realizar las operaciones matemáticas. En la figura 1, se hace la importación de todos los módulos necesarios para la ejecución del código. Posteriormente, se

crea una variable que contiene el generador randómico. De esta manera, se procede a crear las variables “x” e “y” que representan respectivamente a la variable dependiente e independiente. La variable “x” por una parte, solo contiene un array de 35 números randómicos. Mientras que la variable “Y” tendrá el valor que resulte de la función que se muestra en la figura 1. Se debe tomar en cuenta que, para este tipo de funciones, al crear valores randómicos, se debe tener la misma cantidad de elementos tanto en “X” como en “Y”.

Linear Regression

```
# Import the necessary modules
%matplotlib notebook
%matplotlib inline
import matplotlib.pyplot as plt
from scipy import stats
import numpy as np
```

```
# random generator
rng = np.random.default_rng()

# generate random data
x = rng.random(35)
print(x)
y = 1.9*x + rng.random(35)
print(y)
```

```
[0.13578009 0.43232932 0.51472447 0.15815462 0.14282654 0.99953173
 0.77244137 0.23156465 0.36186891 0.04546291 0.39491043 0.33419526
 0.59550617 0.05164021 0.82569485 0.34328547 0.34142207 0.81491863
 0.53507089 0.00572241 0.1438534  0.13545758 0.55939683 0.15054721
 0.56078724 0.64874405 0.04047981 0.30302428 0.78233301 0.07951308
 0.13289808 0.59007615 0.6476858  0.2189592  0.24870053]
[1.17602754 0.85449602 1.1102933  0.45081339 0.53295741 2.2096856
 2.27707579 0.90977752 1.31792218 1.02290798 0.95845107 0.92044243
 1.73715267 0.79999446 1.57578878 1.02075578 0.98626804 2.40944358
 1.28383456 0.06101007 0.77603827 0.6498208  1.94495274 0.30993467
 1.45867814 1.62807929 0.58376291 1.04143343 1.90746435 0.73008168
 0.92968163 1.30545649 1.85823384 0.58422961 0.7670839 ]
```

Figura 1: Importación de módulos y creación de las variables.

En la figura 2, se muestran 4 cuadros con diferentes códigos. En el primer cuadro, se calcula la regresión lineal de los cuadrados tanto para “x” como para “y”. En el siguiente cuadro, se utiliza una función extra, la cual, con la intersección y la pendiente calculadas, devuelve un nuevo valor. Se dice pues, que la pendiente y la intersección determina la relación lineal que existe entre dos variables, por lo tanto, sirve para dar un resultado aproximado sobre la tasa de cambio. En el tercer cuadro, se utiliza la función extra creada anteriormente para calcular los valores que se ajustan al eje “Y” y se los guarda en una variable. En los parámetros que se usan para calcular dichos valores, son la función extra y los valores que almacena ”x”.

```python
# Calculate a linear least-squares regression for two sets of data
slope, intercept, r, p, std_err = stats.linregress(x, y)
```

```python
# Helper function that uses the calculated slope and intercept to return a new value
def fitted_value(x):
    return slope * x + intercept
```

```python
# Calculate the fitted values for the y-axis using the helper function
new_values = list(map(fitted_value, x))
```

Figura 2: Cálculo de los cuadrados y de los valores que se ajustan al eje "Y".

3. Results and Analysis

Finalmente, para mostrar la gráfica, se definen los puntos iniciales que contenían "x" e "y" y se coloca la etiqueta de "datos originales". Y para realizar la gráfica de la regresión lineal, se deben usar los valores que se estimaron de y de la misma manera, se le cola una etiqueta para identificar la línea. Para efecto de la práctica, y que el modelo de regresión lineal sea un poco más gráfico, se definieron 35 puntos en el plano. Con ello, se puede observar en la figura 3, que no todos los puntos, coinciden con la recta, sino más bien, son valores aproximados. Esto sucede cuando se quiere hacer el análisis de una predicción con este modelo básico. Se puede hacer un pronóstico sobre un efecto determinado o la tendencia que tendrá. En este caso, al ser valores randómicos y una función que no se basa en una función específica para determinados datos, el análisis que se le da es muy general. Según los puntos dados, se acercan considerablemente a la recta, aunque la mayoría ni siquiera la toca.

```python
# Draw the original scatter plot
plt.plot(x, y, '*', label='original data')
# Draw the line or linear regression
plt.plot(x, new_values, 'pink', label='fitted line')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1ba33f90070>
```

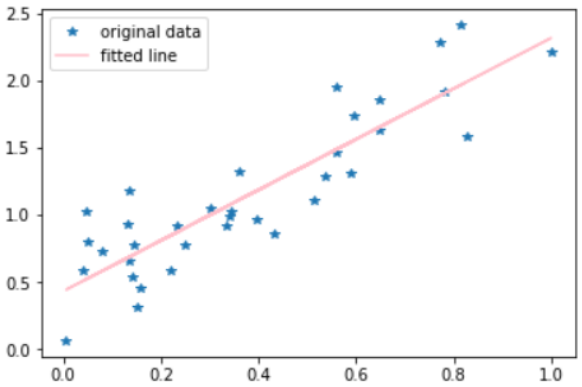

Figura 3: Gráfica del diagrama de disperción y la recta.

4. Discusión

Por la posición en la que está dibujada la función de regresión lineal, se puede decir que es una relación lineal positiva y que se ha encontrado una función que cumple con los requerimientos de los puntos o al menos la mayoría de ellos. Para esto, se tuvo que utilizar el método de mínimos cuadrados, el cual, es el que permite hallar la relación entre las variables dependientes e independientes. Todos los puntos que se graficaron en el plano se conocen como "diagrama de dispersión" y la recta de regresión, es la línea que se encontró en el desarrollo, para hallar la relación entre todos los puntos. Con dicha función, se puede obtener una recta que determina la menor distancia entre ella y cada uno de los puntos.

Se puede determinar entonces, que la regresión lineal permite la predicción de la relación entre las variables dependientes e independientes. Con la recta que se encuentra, se puede determinar la línea que cruza por en medio de los puntos y la menor distancia entre ellos y la recta. Las ventajas que ofrece la regresión lineal, es que es uno de los algoritmos más simples y por lo tanto, su implementación es sencilla. Con la gráfica que se obtiene, se puede analizar la relación que existe entre variables y es que se ajusta muy bien a cada uno de los datos que se encuentran separados linealmente. Otra de las ventajas, es que se puede usar la regularización, que es una técnica para reducir la complejidad de la función, de manera que se reduzca el sobreajuste. Este último término, hace referencia a la situación, en que los datos están muy cerca de la recta lineal, causando ruido y afectando negativamente el rendimiento del modelo.

Por otro lado, también se presentan desventajas respecto a este modelo. Como, por ejemplo, el desajuste. Se da cuando el modelo no captura correctamente los datos, debido a que la relación lineal entre variables, no se ajusta a conjuntos de datos complejos. Para ello, se debe usar una función más compleja, pero aún así, tiene un bajo nivel de precisión. Otra de las ventajas es que es sensible ante valores anómalas, los cuáles afectan significativamente el rendimiento del modelo. Además, la regresión lineal, asume que los datos son independientes, aún cuando no lo son, lo que obliga a aplicar otros procesos como la multicolinealidad, antes de aplicar regresión lineal.

5. Conclusión

La regresión lineal es uno de los modelos más básicos que permite la predicción de relación entre variables, pero no puede asegurar un alto nivel de confianza en el resultado. En este laboratorio, se usaron datos aleatorios, por lo que no se puede dar un análisis en específico sobre un determinado ejercicio. Aún así, en las ventajas y desventajas que se mencionan a lo largo de este documento, se puede deducir que si bien, la regresión lineal es muy sencilla y fácil de implementar, existen ejercicios con datos más complejos, que necesitarán funcionas mucho más complejas. Esto, con el fin de evitar que se generen ruidos en la recta de relación lineal, con datos que se estén sobre la línea y afecten negativamente al resultado de la predicción.

Referencias

Guía entregada por el docente