

Regresión Lineal

Valeria Rodríguez

22 de marzo de 2025

1. Introducción

La regresión lineal es un modelo dentro de la estadística cuyo objetivo es relacionar una o varias variables independientes con una variable dependiente. Esta técnica sirve para encontrar una ecuación que aproxime la relación entre una o varias variables explicativas y una respuesta.

Existen dos tipos de regresiones lineales:

- **Regresión lineal simple:** Se relaciona una única variable independiente con una variable dependiente.
- **Regresión lineal múltiple:** El modelo de regresión tiene varias variables explicativas y una variable respuesta.

2. Metodología

Para la realización de esta actividad se llevaron a cabo una serie de pasos encaminados a la recreación del problema de regresión lineal presentado en el libro *Aprende Machine Learning* del autor Juan Ignacio Bagnato, páginas 35-41.

2.1. Descarga de archivo .csv

Como primer paso, se descargó el archivo 'articulos_ml.csv' proporcionado por el libro de texto para la realización de esta actividad. Para ello se hizo click en el hipervínculo del archivo, el cual redirigió a una página con el link al archivo descargable. Se hizo click y éste se descargó automáticamente.

2.2. Creación de carpeta de trabajo

Seguido de lo anterior, se creó una carpeta llamada Regresion Lineal, en la cual se copió el archivo csv descargado. Posteriormente se creó dentro de esta misma carpeta un archivo Python llamado reg lineal para la codificación de la actividad.

2.3. Desarrollo del código

Con ayuda del IDE Visual Studio Code se abrió el archivo .py anteriormente creado y se codificó según las especificaciones del libro. Primeramente, se hicieron los imports al archivo

```
# Imports necesarios
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import linear_model
```

```
from sklearn.metrics import mean_squared_error, r2_score
```

Para esto, fue necesario instalar algunas de las librerías desde la consola con ayuda de 'pip install'.

Después, hemos estudiado el contenido del archivo csv con ayuda de algunas funciones de Data Frames:

```
#cargamos los datos de entrada
data = pd.read_csv('articulos_ml.csv')
#veamos cuantas dimensiones y registros contiene
print(data.shape)
print(data.head())
print(data.describe())
```

Con estos comandos se ha obtenido la dimensión del archivo, las primeras cinco líneas, así como una breve descripción matemática de su contenido.

Después, se ha creado una regresión lineal con el objetivo de encontrar una correlación entre el número de palabras del texto y la cantidad de compartidos que obtuvo el mismo.

Primeramente, se visualizan los datos en diferentes gráficas de barras, ignorando las columnas correspondientes al Título, URL y el tiempo transcurrido de cada texto.

```
# gráficas de barras del contenido del archivo csv
data.drop(['Title', 'url', 'Elapsed days'], axis=1).hist()
plt.show()
```

Con esto, se han obtenido diversas gráficas que muestran entre qué valores se concentran la mayoría de los datos.

Seguido, se han filtrado los datos de cantidad de palabras para obtener solo aquellos textos que contengan entre 3500 y 80000 palabras para graficar los que estén por debajo de la media de color azul y por encima de la media de naranja.

```
filtered_data=data[(data['Word count'] <=3500)&(data['# Shares'] <=80000)]
colores=['orange','blue']
tamanios=[30,60]
```

```
f1=filtered_data['Word count'].values
f2=filtered_data['# Shares'].values
```

```
asignar=[]
for index, row in filtered_data.iterrows():
    if(row['Word count']>1808):
        asignar.append(colores[0])
    else:
        asignar.append(colores[1])
```

```
plt.scatter(f1,f2,c=asignar,s=tamanios[0])
plt.show()
```

Finalmente, creamos una regresión lineal con los datos correspondientes al recuento de palabras y el número de compartidos de cada texto. Para ello, se entrenó al modelo con el método fit y se predijo el valor de compartidos de un texto de 2000 palabras.

```
#Asignamos nuestra variable de entrada X para entrenamiento y las etiquetas Y.
dataX=filtered_data[["Word count"]]
X_train=np.array(dataX)
y_train=filtered_data['# Shares'].values
```

```

#Creamos el objeto de Regresión Lineal
regr=linear_model.LinearRegression()

#Entrenamos nuestro modelo
regr.fit(X_train,y_train)

#Hacemos las predicciones que en definitiva una línea (en este caso, al ser 2D)
y_pred=regr.predict(X_train)

#Veamos los coeficientes obtenidos, en nuestro caso, serán la Tangente
print('Coeficients: ', regr.coef_)
#Este es el valor donde corta el ejeY (enX=0)
print('Independentterm: ', regr.intercept_)
#Error Cuadrado Medio
print("Mean squared error: %.2f" %mean_squared_error (y_train,y_pred))
#Puntaje de Varianza. El mejor puntaje es un 1.0
print('Variance score: %.2f' %r2_score (y_train,y_pred))
y_Dosmil = regr.predict([[2000]])
print('Predicción de Shres en un artículo de 2000 palabras: ', int(y_Dosmil))

```

3. Resultados

A continuación, se muestran los resultados obtenidos en las diversas fases de codificación mostradas en la sección anterior.

Primeramente, se ejecutaron una serie de comandos para obtener información del archivo de datos.

(161, 8)

Figura 1: Resultado obtenido del data.shape.

	Title	url	Word count	...	# Images video	Elapsed days	# Share
0	What is Machine Learning and how do we use it ...	https://blog.signals.network/what-is-machine-l...	1888	...	2	34	20000
1	10 Companies Using Machine Learning in Cool Ways	NaN	1742	...	9	5	2500
2	How Artificial Intelligence Is Revolutionizing...	NaN	962	...	1	10	4200
3	Obtain and the Blockchain of Artificial Intell...	NaN	1221	...	2	68	20000
4	Nasa finds entire solar system filled with eig...	NaN	2039	...	4	131	20000

Figura 2: Resultado obtenido del data.head.

[5 rows x 8 columns]							
	Word count	# of Links	# of comments	# Images video	Elapsed days	# Shares	
count	161.000000	161.000000	129.000000	161.000000	161.000000	161.000000	
mean	1888.260870	9.739130	8.782946	3.670807	98.124224	27948.347826	
std	1141.919385	47.271625	13.142822	3.418290	114.337535	43408.006839	
min	250.000000	0.000000	0.000000	1.000000	1.000000	0.000000	
25%	990.000000	3.000000	2.000000	1.000000	31.000000	2800.000000	
50%	1674.000000	5.000000	6.000000	3.000000	62.000000	16458.000000	
75%	2369.000000	7.000000	12.000000	5.000000	124.000000	35691.000000	
max	8401.000000	600.000000	104.000000	22.000000	1002.000000	350000.000000	

Figura 3: Resultado obtenido del data.describe.

Como se observa, se obtuvo un archivo de 161 filas y 8 columnas, las primeras 5 filas del archivo, así como un resumen numérico de su contenido.

Después, se creó una serie de gráficas de barras para visualizar la concentración de los datos.

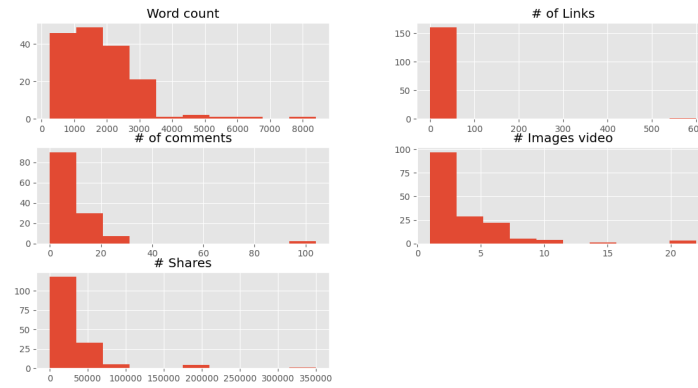


Figura 4: Gráficas de barras del contenido del archivo.

Seguido, se creó una gráfica de dispersión con los textos entre 3500 y 80000 palabras, distinguidos por color entre los que estuvieran por debajo (azul) y por encima (naranja) de la media con respecto a la cantidad de compartidos de cada texto.

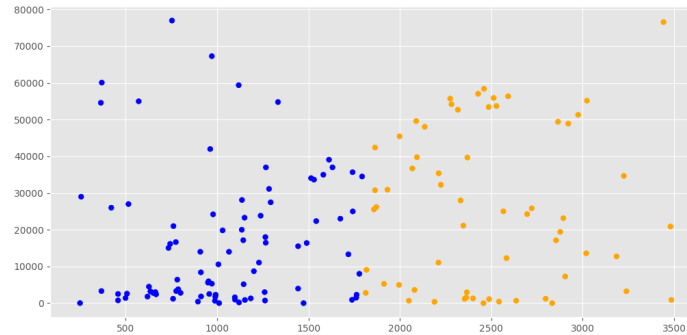
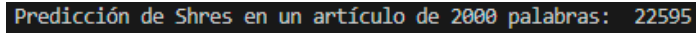


Figura 5: Gráfica de dispersión. Cantidad de palabras vs compartidos.

Finalmente, se creó la regresión lineal de estos datos y se predijo el valor de compartidos para un texto de 2000 palabras, obteniendo los siguientes resultados:

```
Coefficients: [5.69765366]
Independentterm: 11200.30322307416
Mean squared error: 372888728.34
Variance score: 0.06
```

Figura 6: Valores del entrenamiento del modelo.



```
Predicción de Shres en un artículo de 2000 palabras: 22595
```

Figura 7: Predicción para un texto de 2000 palabras

Como se observa, el error cuadrado fue significativamente alto, lo que significa que este modelo no es muy apto para los datos y será necesario una adaptación para un mejor ajuste.

Aún así, con este modelo se ha predicho el número de compartidos para un texto de 2000 palabras, lo que arrojó un número de 22595 compartidos.

4. Conclusión

En conclusión, considero que una regresión lineal puede ofrecer una estimación bastante acercada a la realidad respecto a un conjunto de datos que modele una realidad. Si bien en este caso no se obtuvo un modelo que se ajustara de la mejor forma, considero que el número de datos era un poco escaso para la estimación. Probablemente, con un número mayor de registros se podría haber realizado un mejor ajuste, o bien, con un tipo de regresión diferente, como se abordará en otras actividades.

Referencias

Material de clase. (2025). UANL