Regresión Lineal IA ARLG

Abraham López

Marzo 2025

1 Introducción

1.1 ¿Qué es la regresión lineal?

La regresión lineal es un método estadístico que permite modelar la relación entre una variable dependiente Y y una o más variables independientes X. El objetivo es encontrar la línea recta (en el caso de una sola variable independiente) que mejor se ajuste a los datos observados. En términos más simples, busca modelar cómo cambia una variable (la dependiente) en función de otra variable (la independiente). La ecuación de la regresión lineal simple se puede expresar como:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde:

- Y es la variable dependiente.
- X es la variable independiente.
- β_0 es la intersección con el eje Y.
- β_1 es la pendiente de la recta.
- ϵ es el término de error.

1.2 ¿Qué es la regresión lineal en el machine learning?

En el machine learning, los programas de computación denominados algoritmos analizan grandes conjuntos de datos y trabajan hacia atrás a partir de esos datos para calcular la ecuación de regresión lineal. Los científicos de datos primero entrenan el algoritmo en conjuntos de datos conocidos o etiquetados y, a continuación, utilizan el algoritmo para predecir valores desconocidos. Los datos de la vida real son más complicados que el ejemplo anterior. Es por eso que el análisis de regresión lineal debe modificar o transformar matemáticamente los valores de los datos para cumplir con los siguientes cuatro supuestos:

1.2.1 Relación lineal

Debe existir una relación lineal entre las variables independientes y las dependientes. Para determinar esta relación, los científicos de datos crean una gráfica de dispersión (una colección aleatoria de valores x e y) para ver si caen a lo largo de una línea recta. De lo contrario, puede aplicar funciones no lineales, como la raíz cuadrada o el logaritmo, para crear matemáticamente la relación lineal entre las dos variables.

1.2.2 Independencia residual

Los científicos de datos utilizan residuos para medir la precisión de la predicción. Un residuo es la diferencia entre los datos observados y el valor previsto. Los residuos no deben tener un patrón identificable entre ellos. Por ejemplo, no querrá que los residuos crezcan con el tiempo. Puede utilizar diferentes pruebas matemáticas, como la prueba de Durbin-Watson, para determinar la independencia residual. Puede usar datos ficticios para reemplazar cualquier variación de datos, como los datos estacionales.

1.2.3 Normalidad

Las técnicas de representación gráfica, como las gráficas Q-Q, determinan si los residuos se distribuyen normalmente. Los residuos deben caer a lo largo de una línea diagonal en el centro de la gráfica. Si los residuos no están normalizados, puede probar los datos para detectar valores atípicos aleatorios o valores que no sean típicos. Eliminar los valores atípicos o realizar transformaciones no lineales puede solucionar el problema.

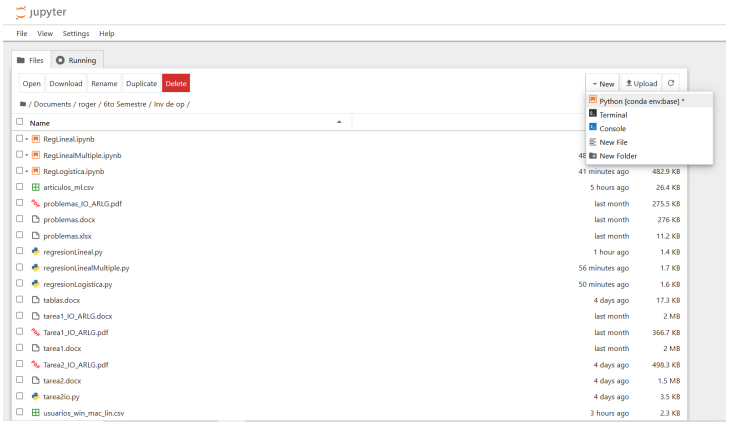
1.2.4 Homocedasticidad

La homocedasticidad supone que los residuos tienen una variación constante o desviación estándar de la media para cada valor de x. De lo contrario, es posible que los resultados del análisis no sean precisos. Si no se cumple esta suposición, es posible que tenga que cambiar la variable dependiente. Dado que la variación se produce de forma natural en grandes conjuntos de datos, tiene sentido cambiar la escala de la variable dependiente. Por ejemplo, en lugar de usar el tamaño de la población para predecir la cantidad de estaciones de bomberos en una ciudad, podría usar el tamaño de la población para predecir la cantidad de estaciones de bomberos por persona.

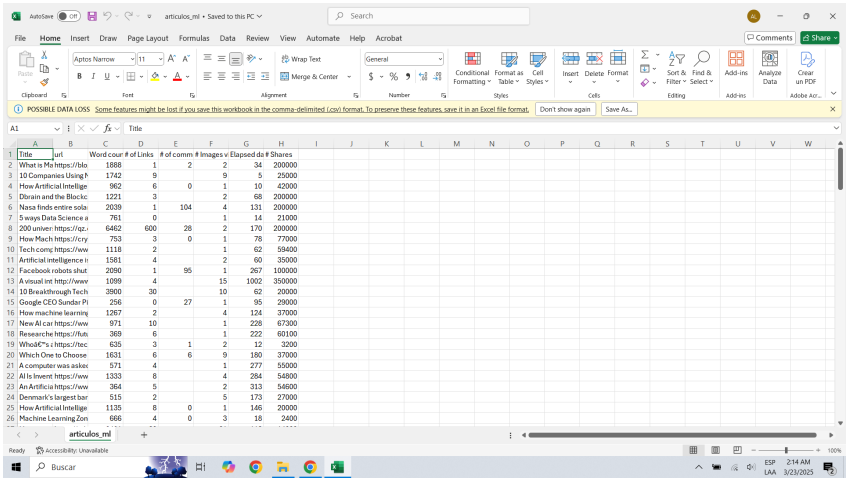
2 Metodología

En este ejemplo, se realizó un análisis de regresión lineal simple para explorar la relación entre el número de palabras (Word count) y el número de compartidos (Shares) en artículos de Inteligencia Artificial. El objetivo es determinar si el número de palabras en un artículo podía predecir su popularidad, medida en términos de compartidos.

Primeramente, creamos un archivo .ipynb en Jupyter para poder hacer el código y correrlo dentro de esta plataforma.



Descargamos el archivo .csv necesario para realizar las operaciones.



Importamos lo necesario y así poder comenzar el código.

jupyter RegLineal Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python [conda env:base] *

[5]:
```
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
%matplotlib inline
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score

data = pd.read_csv("./articulos_ml.csv")
data.shape
```

[5]: (161, 8)

[7]:
```
data.head()
```

[7]:

	Title	url	Word count	# of Links	# of comments	# Images video	Elapsed days	# Shares
0	What is Machine Learning and how do we use it ...	https://blog.signals.network/what-is-machine-l...	1888	1	2.0	2	34	200000
1	10 Companies Using Machine Learning in Cool Ways	NaN	1742	9	NaN	9	5	25000
2	How Artificial Intelligence Is Revolutionizing...	NaN	962	6	0.0	1	10	42000
3	Dbrain and the Blockchain of Artificial Intell...	NaN	1221	3	NaN	2	68	200000
4	Nasa finds entire solar system filled with eig...	NaN	2039	1	104.0	4	131	200000

El primer paso consiste en cargar los datos y realizar una exploración inicial para entender su estructura y contenido.

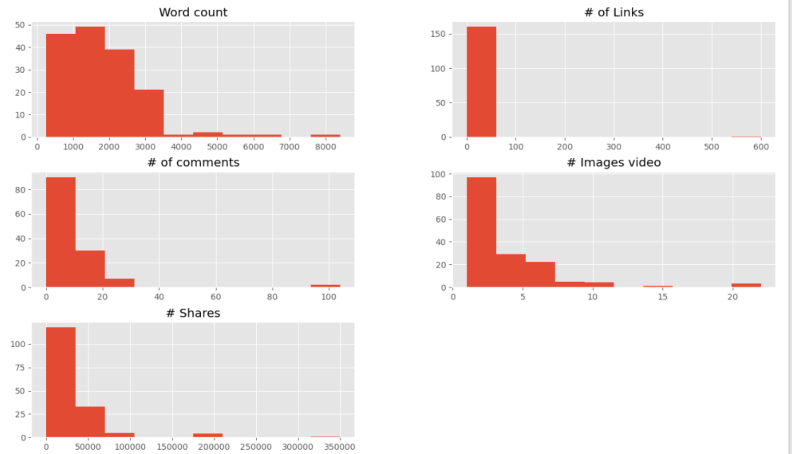
[9]:
```
data.describe()
```

[9]:

	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	1808.260870	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

Se realizan gráficos para visualizar la distribución de los datos y detectar patrones o valores atípicos.

[11]:
```
data.drop(['Title', 'url', 'Elapsed days'],axis=1).hist()
plt.show()
```

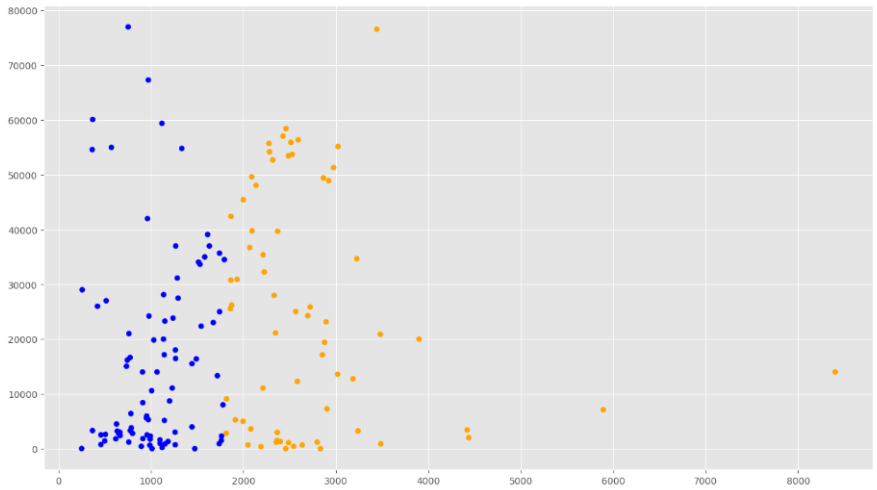


Se preparan los datos para crear un gráfico de dispersión que muestre la relación entre las variables.

```
[13]: filtered_data = data[(data['Word count'] <= 35000) & (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()
```



Se entrena un modelo de regresión lineal utilizando los datos filtrados.

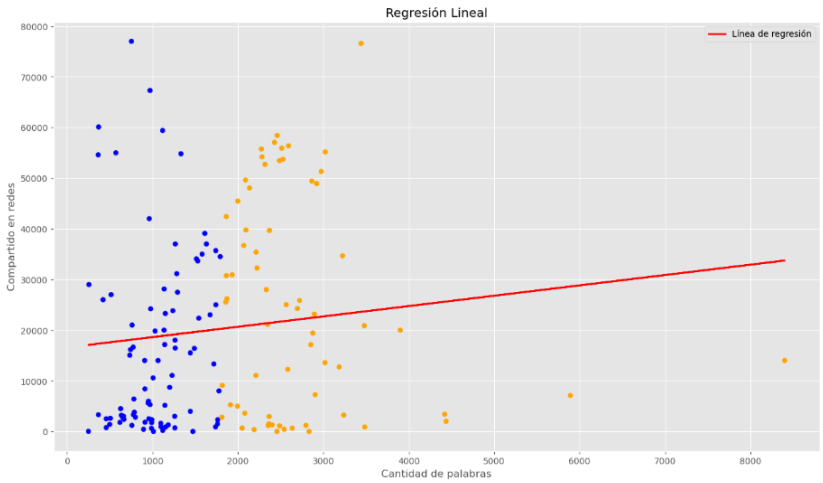
```
[15]: datax=filtered_data[["Word count"]]
X_train=np.array(datax)
y_train=filtered_data['# Shares'].values
regr=linear_model.LinearRegression()
regr.fit(X_train, y_train)
y_pred=regr.predict(X_train)
print('Coefficients: \n', regr.coef_)
print('Independent term: \n', regr.intercept_)
print("Mean squared error: %.2f" % mean_squared_error(y_train, y_pred))
print('Variance score:%2f' % r2_score(y_train, y_pred))
```

```
Coefficients:
 [2.04272716]
Independent term:
 16575.9880343537
Mean squared error: 382325815.45
Variance score:0.012777
```

```
[21]: plt.scatter(f1, f2, c=asignar, s=tamanios[0])
plt.plot(f1, y_pred, color='red', linewidth=2, label="Línea de regresión")

plt.xlabel("Cantidad de palabras")
plt.ylabel("Compartido en redes")
plt.title("Regresión Lineal")
plt.legend()

plt.show()
```



Por último, predecimos el número de shares para un artículo con 2000 palabras.

3 Resultados

Tras realizar el análisis de regresión lineal, se obtuvieron los siguientes resultados:

- **Coeficiente (pendiente)**: 2.04
  Esto indica que, por cada palabra adicional en un artículo, se espera que el número de shares aumente en aproximadamente 2.04.
- **Término independiente (intercept)**: 16575.98
  Este valor representa el número de shares esperado cuando el número de palabras es 0. En este caso, si un artículo tiene 0 palabras, se esperarían aproximadamente 16,576 shares.
- **Mean Squared Error (MSE)**: 382,325,815.45
  El MSE es una medida del error promedio entre los valores reales y los predichos. Un valor alto como este sugiere que el modelo tiene un error significativo en sus predicciones, lo que indica que no se ajusta bien a los datos.
- **Variance Score (R^2)**: 0.012777
  El coeficiente de determinación (R^2) indica que solo el 1.28% de la variabilidad en el número de shares es explicada por el número de palabras. Esto sugiere que hay otros factores no incluidos en el modelo que influyen en el número de shares.

4 Conclusión

El análisis de regresión lineal que se llevó a cabo nos permitió investigar la relación entre el número de palabras en los artículos y la cantidad de shares que reciben, específicamente en el ámbito del machine learning y la inteligencia artificial. Los resultados que obtuvimos muestran que el número de palabras tiene un efecto bastante limitado en la predicción de los shares, como lo indica el bajo coeficiente de determinación ($R^2 = 0.012777$), que solo explica el 1.28% de la variabilidad en los datos. Además, el elevado error cuadrático medio (MSE = 382,325,815.45) sugiere que el modelo no es muy preciso en sus predicciones.

Estos hallazgos indican que el número de palabras por sí solo no es un buen predictor del número de shares. Es probable que otros factores, como la cantidad de enlaces, imágenes, comentarios o incluso el tema del artículo, tengan un impacto considerable en la popularidad de los mismos. Por lo tanto, sería recomendable ampliar el análisis para incluir estas variables adicionales en un modelo de regresión múltiple o explorar técnicas más avanzadas de machine learning que nos ayuden a captar mejor la complejidad de los datos.

En resumen, aunque el modelo de regresión lineal nos da una primera idea, es fundamental adoptar un enfoque más robusto y multidimensional para mejorar la precisión de las predicciones y entender mejor los factores que influyen en el éxito de los artículos en términos de shares.