



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores
Acatlán

Practica 1:

Regresión lineal simple

Aprendizaje de maquina

Integrantes:

Gustavo Adolfo Alvarez Hernández

Profesor:

Eduardo Eloy Loza Pacheco



FES Acatlán, 05 de septiembre 2022

Objetivo

Aprender a modelar datos con comportamientos lineales con el fin de dar predicciones sobre valores ajenos a los datos que conocemos, así como notar el impacto de la variable explicativa en la explicada; todo esto por medio de la implementación de la regresión en Python.

Materiales y métodos

- Google colab: Utilizando Sklearn, pandas y matplotlib.
- Regresión lineal simple.

Resumen

Los modelos de regresión lineal simple buscan modelar la relación que tienen dos variables entre sí, por medio de ajustar los datos que se tienen por medio de una recta. En este caso una variable se denomina explicativa (llamémosla x), la cual explica a la variable dependiente (en su caso y).

De este modelo podemos obtener que tan relacionadas están las variables y además su término independiente α , que nos indica desde donde empieza y . Dando como resultado una recta de la forma $y = \alpha + \beta x$.

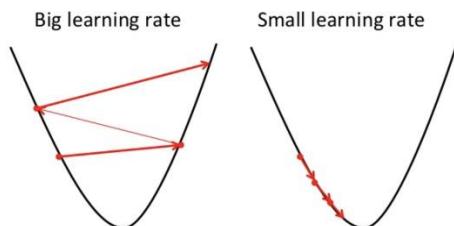
Para poder realizar este modelo nuestros datos deben cumplir los siguientes supuestos:

1. Homogeneidad en la varianza
 - a. Los errores de nuestra estimación no cambian significativamente a través de la variable independiente.
2. Independencia de las observaciones
 - a. Las observaciones se recabaron con métodos de muestreo y no hay relaciones ocultas entre ellas.
3. Normalidad
 - a. Nuestros datos se distribuyen

El cálculo de estos coeficientes se obtiene por medio de minimizar el error cuadrático medio, siendo esta nuestra función de pérdida y encontrando el mínimo por medio del gradiente descendiente.

$$\text{minimize} \frac{1}{n} \sum_{i=1}^n (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^n (pred_i - y_i)^2$$



Otra opción es el uso del cálculo para encontrar los mínimos, el cual para el caso de una regresión lineal tenemos que:

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

$$\hat{\beta} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

Desarrollo

1. Cargamos todos las bibliotecas que requerimos para la práctica:

Modules

```
▶ import numpy as np
  import matplotlib.pyplot as plt
  import pandas as pd
  from sklearn.model_selection import train_test_split
  from sklearn.linear_model import LinearRegression
```

2. Cargamos nuestros datos y les damos un vistazo:

▼ Load data

```
[ ] df_sal = pd.read_csv("/content/Salario.csv")
```

```
[ ] df_sal.head()
```

	Experiencia	Salario
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

3. Para proseguir a tomar a X como nuestra variable explicativa, que será la experiencia y en su caso y, será el salario. Después dividimos los datos en entrenamiento para el modelo y test para ver si el resultado fue satisfactorio.

```
▶ X= df_sal.iloc[:,1].values
  y= df_sal.iloc[:,1:2].values
```

▼ Split data

```
[ ] x_train,x_test,y_train,y_test = train_test_split(X,y,test_size =1/3, random_state= 0)
```

4. Entrenamos el modelo con los datos de entrenamiento y predecimos los valores de test, puesto que nunca ha visto estos valores:

▼ Linear model

```
[ ] regresion = LinearRegression()  
regresion.fit(x_train,y_train)  
y_pred = regresion.predict(x_test)
```

5. Por último, graficamos los datos de train y test, además de la recta de regresión para ver su performance a la hora de ajustarse.



Conclusiones

Podemos observar que la recta en ambos casos no pasa exactamente por todos los puntos, puesto que no sería posible una recta con esa cualidad, no obstante, esta bastante cerca de los datos; este comportamiento al conservarse en ambos conjuntos de datos es una buena señal, puesto que podemos decir que el modelo generaliza para los datos (o cuando menos para los salarios de esta empresa).

Por otra parte, se ve una estrecha relación entre el salario y los años trabajados, siendo a mayor experiencia la persona puede acceder a un salario mayor. Por lo cual es directamente proporcional y tiene sentido en el mundo real.

Otro punto para denotar es que se obtuvieron resultados satisfactorios debido a la naturaleza de los datos, esto es se comportan linealmente, puesto que al no ser el caso (comportamiento exponencial, por decir uno) nuestra regresión daría resultados malos.

Referencias.

- [1] R. Bevans, «Scribbr,» 19 02 2020. [En línea]. Available:
<https://www.scribbr.com/statistics/simple-linear-regression/>. [Último acceso: 05 09 2022].
- [2] R. Gandhi, «Towards Data Science,» 10 05 2018. [En línea]. Available:
<https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a>. [Último acceso: 05 09 2022].