Regresión Lineal

Alejandra Verónica López Chiquito[a]

[a]*Universidad Autónoma de Querétaro, Querétraro, México*

Abstract

Durante estos ejercicios se aplicará el conocimiento aprendido sobre regresión lineal en la clase de Deep Learning, así como la presentación de sus resultados y conclusiones.

Keywords: Python, Regresión lineal, Estadística, SME

1. Introducción

En estos ejercicios se aplica todos los conocimientos aprendidos en sesiones de la clase de deep learning, como son, corarianza, correlación de Pearson, regresión lineal simple y gradiente descendente. Además, en el dataset de articulos de maching learning articles se aplicaron técnicas de imputación como parte del pre-procesamiento de los datos y esto debido a que el dataset presenta instancias con datos faltantes.

2. Metodología.

En ambos casos se siguió la idea de poder encontrar la relación más solida entre una variable independiente y la variable dependiente. Posteriormente, se aplicará el algoritmo de optimización, gradiente descendente, para obtener los mejores pesos para la hipotesis (h). Una vez hecho esto, se comentarán las conclusiones al final de este documento.

3. Definiciones

3.1. Covarianza

La covarianza es una medida estadística que se utiliza para evaluar la relación entre dos variables aleatorias. Mide cómo estas dos variables cambian juntas. En otras palabras, la covarianza indica si hay una tendencia a que ambas variables aumenten o disminuyan simultáneamente, o si una aumenta mientras la otra disminuye.

3.2. Correlación de Pearson

Es una medida estadística que cuantifica la fuerza y la dirección de la relación lineal entre dos variables continuas. El coeficiente de correlación de Pearson varía en un rango entre -1 y 1.

3.3. Regresión lineal

Es un método que se utiliza para representar la relación entre variables independents con una variable dependiente, esto mediante el calculo de los valores corrector para satisfacer la hipotesis (h), la cual tiene la format: Y = mx + b.

3.4. Gradiente Descendente

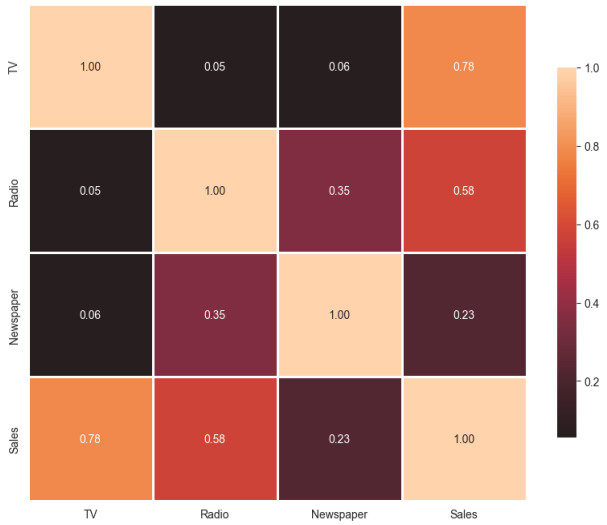
Es un algortimo de optimización que permite encontrar los valores mínimos globales de una función, esto mediante un proceso iterativo en donde se agrega un valor de aprendizaje con el cual en cada iteración se actualizan los valores que representan el mínimo de la hipotesis hasta llegar al optimo o acabar el número de iteraciones o épocas designadas.

4. Ejercicios

4.1. Advertising.

Dentro del dataset usado para este ejercicios, se puede observar que se tienen 3 variables independientes (VI) las cuales son 'TV', 'Radio' y 'Newspaper'; Y una variable dependiente (VD) que es la columna 'Sales', es decir que en estos datos el objetivo es poder encontrar que medio de comunicación es la que se relaciona más con las ventas.

Para esto se desplegaran gráficas de cada una de las VI con la VD, es decir, TV y Sales, Radio y Sales, Newspaper y Sales, de esta manera de visualizará de manera más fácil si existira o no un a relación entre ellas. Al visualizar las gráficas se puede intuir que la variable TV tiene mayor relación con la VD.



Para validar esta hipoteis, se crean 2 funciones con métodos de estadistica descriptiva que son covarianza y correlación de Pearson. Se aplican a todas las VI con la VD y los resultados

De manera manual y usando funciones predeterminadas se obtienen los valores de la covarianza y correlación de Pearson para así validar la hipotesis planteada anteriormente, la columna de TV es la VI que tiene más influencia en la VD.

Una vez que se tiene esta información, se desarrorlla la función del gradiente descendente (GD) y además las funciones correspondientes para gráficar la recta que representa la regresión lineal como es el MSE (Mean Square Error) que es la métrica que nos dice que tan 'bueno' es nuestro modelo, o que tan bien está aprendiendo. Esta métrica es muy utilizada en la regresión lineal ya que en téminos generales, mide la distancia de cada punto de la VI con la recta obtenida del algoritmo del GD.

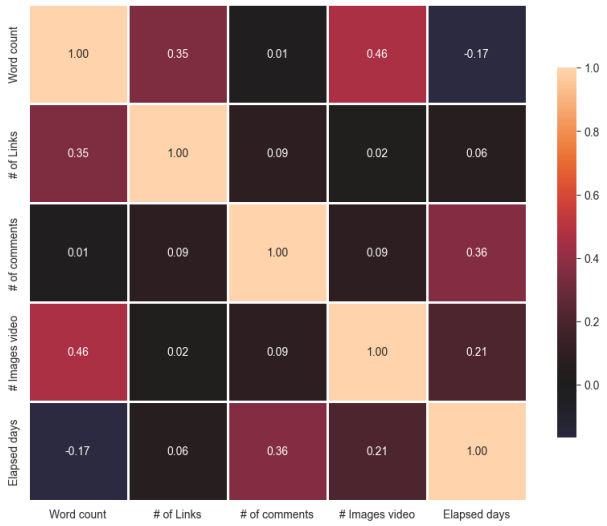
Para usar a la función del GD, se requiere proporcionar los valores de la columna de la VI que se está estudiando, la columna de la VD, los valores iniciales de las variables thetas, las cuales se van a actualizar en cada iteración, el valor de alfa o learning rate, es decir la distancia que recorrerá el algoritmo en cada iteración, y las épocas o el número de itraciones que se va a repetir el algoritmo. La función inicializa los valores de los gradientes en ceros y obtiene el tamaño de la la VI. Con toda esta información, el algoritmo empieza a actualizar los valores de los gradientes y a su vez las variables thetas hasta terminar las épocas. Una vez que termina imprimer los valores finales de las thetas y con la función de regression lineal, se puede ver gráficamente la inclicación de la recta obtenida. Es importante que los valores en los cuales iniciar los thethas sean los más adecuados para que el algoritmo pueda dar el mejor resultado por lo que se estuvo intentanto con muchos valores para llegar a los usados en el notebook.

Se repite este proceso con las 3 VI y al final del notebook se puede concluir que la VI que más influye en las ventas es la TV debido a que el valor del MSE más bajo que que se logró obtener fue con dicha variable.

4.2. Machine learning articles.

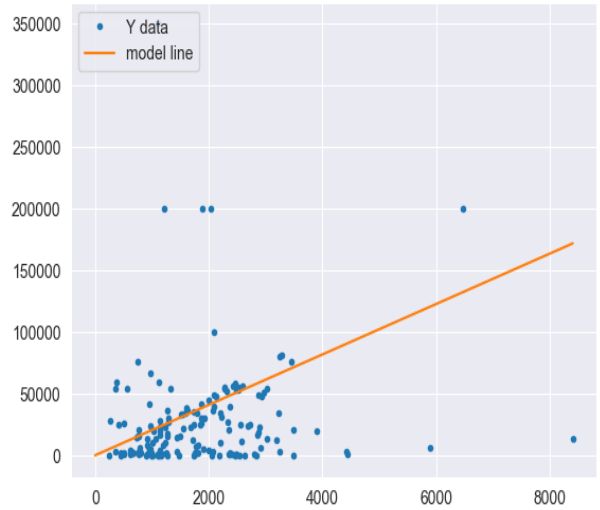
En este ejercicio de trabará con un documento que contiene información realcionada con articulos de machine learning. Dentro de este dataset se tienen 8 columnas de las cuales 5 de ellas actuarán como VI (Word count, Num of Links, Num of comments, Num Images video, Elapsed days) ya que son columnas de valores continuos que intentamos relacionar con la columma de la VI (Num Shares).

Como primer paso se aplican las mismas funciones de estadistica descriptiva del ejercicio anterior para obtener valores de covarianza y correlación de Pearson. Como complemento se muestran las gráficas de todas las VI con la VD.



A diferencia del primer ejemplo en este documento, no se puede observar que alguna VI tenga una relación significativa o por encima de las demás con la VD. Se usará el mismo algoritmo del GD en este ejercicio y se aplicará a todas las VI.

Ya que como se menciona, no hay relación con ninguna VI, se muestra una imagen de la regresión más óptima para este ejercicio, la cual es con la columma 'Num Comments'.



5. Conclusiones

El método de Regresión lineal es muy utilizado y eficiente para lograr predecir algún valos continuo a partir de una serie de entrenamiento y datos. Sin embargo, como se pudo apreciar en los resultados obtenidos de los ejercicios expuestos anteriormente, no es un método que se pueda utilizar en todos los casos reales que tengamos. Cuando el caso es apropiado, los resultados pueden ser muy útiles, por otro lado, cuando no es apropiado aplicar este método, los resultados son muy malos.

Algunas de las ecuaciones propuestas por este modelo, son únimanete para realciones lineales, y como se sabe, no todos los problemas reales las presentan. Con por ejemplo la correlación de Pearson, la cual, con ayuda de la librearía matplotlib, devuelve un resumen muy completo de la correclación entre todas las variables del dataset, únicamente es útil cuando la relación es lineal.

Es importante analizar le problema que tenemos para decidir y aplicar el método más correcto y garantizar el mejores resultados del modelo.