Andrea Abril Palencia Gutiérrez, 18198
José Gabriel Block Staackmann, 18935

Universidad del Valle de Guatemala
22 de oct. de 22
Métodos Numéricos

Proyecto No.3

Regresión Lineal y Modelos de Población

Introducción

En este proyecto se realizaron modelos de regresion lineal utilizando ***machine learning***. Por lo cual, ya que es una regresion lineal el tipo de aprendizaje de nuestro modelo es un aprendizaje supervisado, que es una herramienta útil para predecir una respuesta cuantitativa.

En sentido amplio lo que hace una regresión lineal es obtener la relación entre unas variables independientes (*X*) y una variable dependiente (*Y*). Es decir, teniendo una serie de variables predictoras obtiene la relación con una variable cuantitativa a predecir. La regresión lineal explica la variable *Y* con las variables *X*, y obtiene la función lineal que mejor se ajusta o explica esta relación.

Sabemos que la ecuación de una línea recta es básicamente:

$$Y = mx + b$$

Donde *b* es el intercepto y *m* es la pendiente de la línea. Así que básicamente, el algoritmo de regresión lineal nos da el valor más óptimo para la intercepción y la pendiente (en dos dimensiones). Las variables *y* y *x* siguen siendo las mismas, ya que son las características de los datos y no pueden cambiarse. Los valores que podemos controlar son el intercepto(*b*) y la pendiente(*m*).

Por lo cual esta relacion puede ser escrita de esta forma:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{modelo} + \underbrace{\epsilon}_{error}$$

La parte β0+β1X es el modelo de regresión lineal, siendo β0 y β1 los coeficientes de la regresión lineal y ϵ el error cometido por el modelo.

Casos

Caso #1:

Este caso es un poco más sencillo, en este se utilizó un dataset que trae la librería que importamos. En el que nos proponemos ver la relación existente entre

el número medio de habitaciones de un conjunto de viviendas en Boston y su valor medio. Para este caso se hizo una carga y exploración de datos distinta al caso #1 debido al tipo de dataset a trabajar.

La información general del dataset es la siguiente:

```
.. _boston_dataset:

Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```

Caso #2:

Este caso utiliza un dataset que se puede conseguir en ***Kaggle*** llamado ***Breast Cancer Wisconsin (Diagnostic) Data Set***, que contienen las características y diagnóstico de los núcleos celulares de 569 estudios realizados. El dataset completo contiene 32 columnas pero paras este ocasión se exploraron solo 12 las cuales son:

1. ID
2. Diagnóstico (B = Benigno, M = Maligno)
3. Radio medio (media).
4. Textura (media).
5. Perímetro (media).
6. Área (media).
7. Suavidad (media).
8. Compacidad (media).
9. Concavidad (media).
10. Puntos Cóncavos (media).
11. Simetría (media).
12. Dimensión Fractal (media)

Para este caso se utilizaron 3 variables que son el diagnostico, para observar el tipo de tumor que es, luego el área de dicho tumor y por ultimo el perimetro del tumor. Con esto en mente se queria evaluar o analizar si habia una relacion entre el perimetro y el area de los tumores haciendo una diferenciación entre tumor benigno y maligno.

Caso #3:

En este caso se quiso comparar nuestro modelo de ***machine learning*** con la forma en que se vio en clase. Este caso fue un poco más dificil de lo esperado ya que tuvimos que sacar la data del csv y colocarlo en arreglos para poder utilizar el ejemplo de ***Ajuste Lineal*** visto en clase.

Para este caso se utilizó un dataset que se puede conseguir en ***Kaggle*** llamado ***Graduate Admission 2***, que consiste en parametros de admisión para programas de maestrías en India. El dataset contiene parámetros que son importantes en el proceso de postulación de los estudios de maestría que son:

1. GRE (Graduate Record Examination)
2. Puntaje en TOELF
3. Ranking Universidad
4. Carta de recomendación
5. Calificación Escolar
6. Experiencia en Investigación
7. Oportunidad de admisión

Se urilizo la columna de **GRE (Graduate Record Examination)** y ***Oportunidad de admisión*** para lograr ver la relación de estas dos variable, si es que al tener un buen puntaje ***GRE*** este tiene mayor oportunidad de admisión en estudios de maestrias.

Resultados

Caso #1:

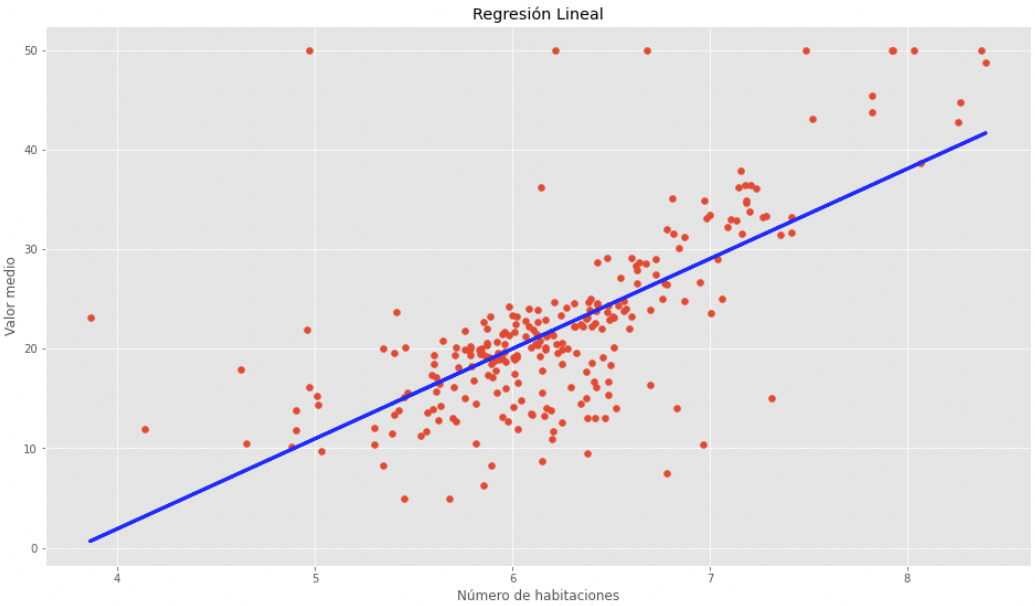

Ilustración 1: Regresión lineal de viviendas en Boston

```
Coeficiente:  [9.04200285]
valor donde corta el eje Y (en X=0):  -34.252438682975956
Error medio cuadrado:  353.10091416205535
Puntaje de Varianza: 0.47667422928407244
```

Caso #2:

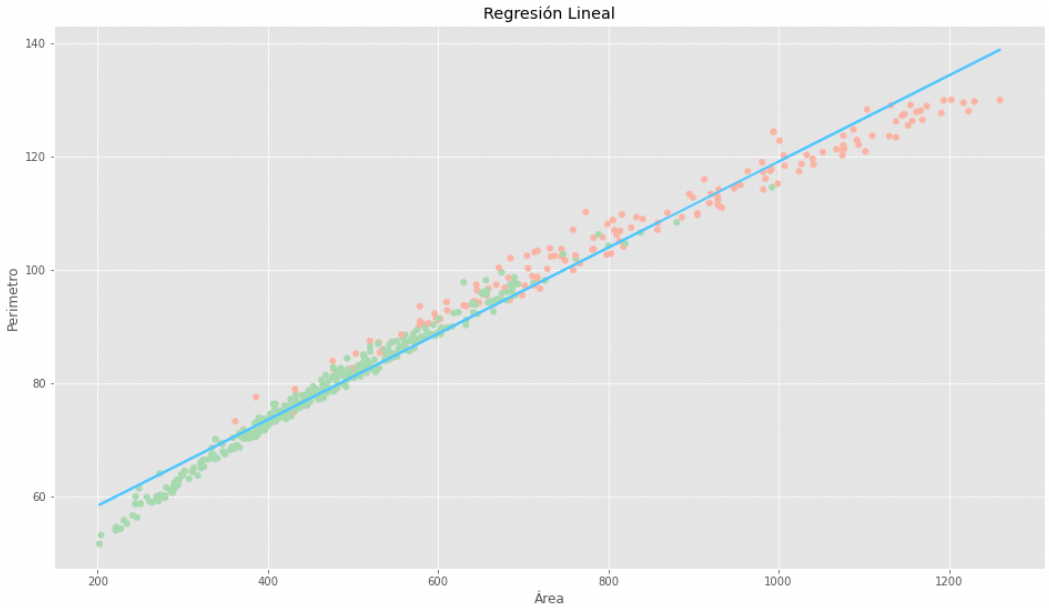

Ilustración2: Regresión lineal de cáncer de mama

```
Coeficiente:  [0.07586808]
valor donde corta el eje Y (en X=0):  43.261551514920036
Error medio cuadrado: 6.27
Puntaje de Varianza: 0.98
```

Caso #3:

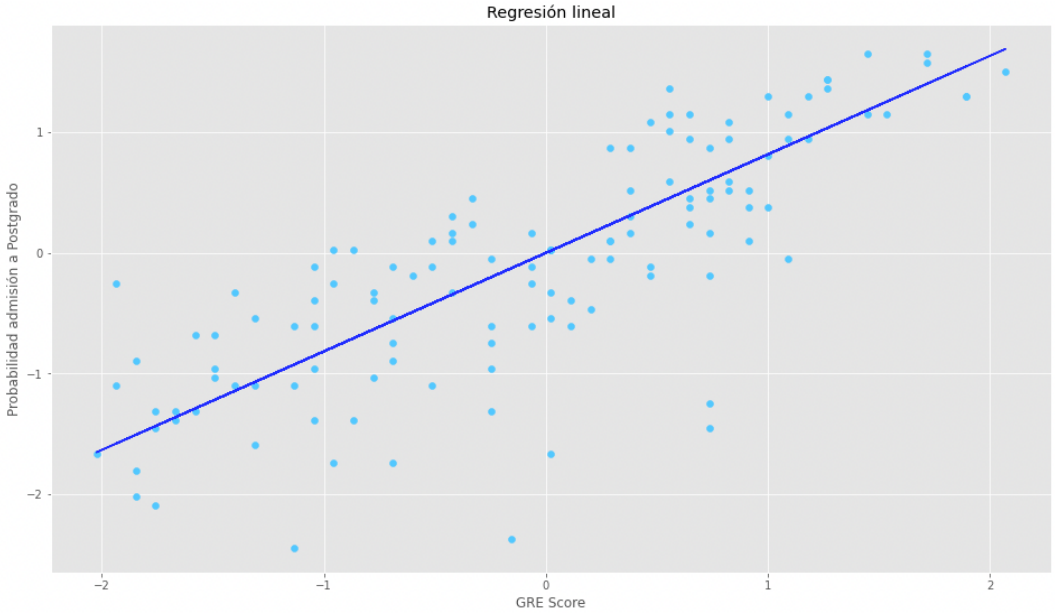

Ilustración 3: Regresión lineal de admisiones

Modelo	Ecuación
Visto en clase	$y = 0.81605 \cdot x + 2.97708 \cdot 10^{-13}$
Sklearn (***machine learning)***	$y = 0.81604 \cdot x + 1.46932 \cdot 10^{-15}$

Tabla 1: Comparación de modelos

Discusión de resultados

Caso #1:

En este caso nos habiamos planteado la hipotesis de que entre más habitaciones tiene la vivienda el valor iba a ser más alto, pero sin envargo como podemos observar en la ***Ilustracion 1*** esta nos muesta que hay muchos datos que estan fuera de la linea lo cual se podria deber a la naturaleza de los datos utilizados ya que estan muy dipersos entre si. Por lo cual no se podria decir la afirmación de nuestra hipotesis.

Otro punto a tomar en cuenta es nuestro error medio cuadrado es de 353.1 lo cual es demasiado elevado para el modelo y esto quiere decir que nuestro modelo fue mal entrenado debido al dataset. Aunque dando un vistazo a la ***Ilustración 1*** y metricas del caso si se obserba que dentro de lo que cabe el valor de la vivienda si tiende a subir al momento de tener mas habitaciones.

Una sugerencia o recomendación seria filtrar más el dataset para reducir la cantidad de datos que estan fuera de la media o dispersos y asi mejorando el rendimiento de nuestro modelo, pero tendriamos que tener cuidado para evitar el overfitting.

Caso #2:

Como se puede observar en la ***Ilustración 2*** las caracteristicas de área y perimetro son unas de las que más influyen en el valor del diagnostico (benigno = verde, maligno = rosa). Pero para lograr esto se tuvo que realizar una serie de ajustes en el dataset, ya que se delimito el conjunto de datos a analizar. Los datos que estaban más dispersos se decidieron eliminar para poder obtener un buen modelo de regresión lineal.

Pero por otro lado, si observamos el error medio cuadrado es de *6.27* lo cual es algo elevado para el modelo. Esto significa que nuestro algoritmo no fue muy preciso pero aún así puede hacer predicciones razonablemente buenas. Esto puede ser a que se deba a que el dataset no tiene una buena cantidad de datos para hacer una predicción más precisa o que realmente la relación entre el perimetro y el area no sea tan buena y el perimetro no dependa del area del tumor.

Aunque si observamos la concentraciopn de datos, los tumores malignos se encuentra más arriba y los benignos más debajo de la gráfica, tal vez se pueda

interpretar en que si el tumor tiene un perimetro y area más alta este tenga más posibilidad de ser un tumor maligno. Pero puede ser una conclusion erronea ya que no tenemos datos suficientes para hacer dicha afirmación, y también esto pueda depender en la etapa en que se diagnosticaron los tumores.

Caso #3:

Este caso fui algo complicado por el dataset que se llego a utilizar ya que se tuvo que examinar con cuidado para ver con que variables se iba a trabajar el modelo. Al final se obtuvo una buena regresion lineal como se puede observar en la Ilustración 3.

Como se observa en la Tabla 1 se compararon los Ajustes Lineales que se obtuvieron, en lo que podemos decir que fue mucho más facil y rapido implementar ***machine learning*** que como se vio en clase ya que el dataset tuvo que cambiarse y ajustarse al formato visto. Por otro lado, se puede ver que nuetro modelo de sklearn fue mucho más preciso que el otro, ya que el de visto en clase fue bastante tardado a comparación y no llego al nivel o grado de precision deseado. Esto se puede deber a que nosotros implementamos un aprendizaje automatizado de la maquina, teniendo que entrenar el modelo para que este sea mucho más preciso al hacer la prueba.

Como alguna recomendación para futuros proyectos podriamos decir que estos modelos pueden ser mejorados y aplicarse también para regresión lineal múltiple o regresión lineal polinómica, como para valorar muchas mas variables en los modelos haciendolos asi mucho más presisos e interesantes. También, que el modelo visto en clase es una buena herramienta para hacer regresiones lineales pero a nuestra consideración no a gran escala o gran dantidad de data porque se hace tedioso de manejar al momento de analizar los datos, si se utiliza para ejemplos un poco más pequeños esta forma es muy buena.

Conclusiones

- La implementación de ***machine learning*** hizo que las regresiones lineales fueran más precisas, como se puede observar en la ***Tabla 1***.
- Al tener un dataset no tan disperso se podra obtener un mejor modelo de regresion lineal siendo mucho más preciso, sin un error tan elevado.
- En el caso 3, si se puede observar que entre más puntaje de ***GRE*** la probabilidad de admisión es más alta, a excepción de algunos casos pocos comunes.
- En el caso 2, al tener más habitaciones una vivienda este sube su valor en Boston, a excepción de algunos casos pocos comunes.

Anexos

Dataset del caso No.3:
https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

Dataset del caso No.2:
https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data