Reporte de Laboratorio Nro. 1

Vinicio Borja[L00393007]

Universidad de las Fuerzas Armadas
vlborja@espe.edu.ec

Tema: Regresión Lineal

Resumen

Este laboratorio tiene como objetivo documentar el código aplicando distintas técnicas de procesamiento de datos en el conjunto de datos para predicción de precios inmobiliarios, a su vez, haciendo uso de la ingeniería de características. Para cumplir este objetivo, se ha utilizando el lenguaje de programación Python. Para este propósito, se utilizaron funciones para técnicas de procesamiento para mostrar los resultados estadísticos por medio de acceso a bibliotecas y marcos de inteligencia artificial y aprendizaje automático (ML). Ya en la etapa de prueba se mostró los datos de test y los datos predichos. Por lo tanto, la relación entre dos variables del conjunto de datos de las viviendas. Representa puntos de datos en un plano 2D o sistema cartesiano. En el gráfico, se muestra una fuerte correlación positiva general entre el diámetro de un árbol y su altura. También podemos observar un punto atípico, un árbol con un diámetro mucho mayor que los demás. En conclusión, se puede afirmar que un modelo de regresión lineal simple establece una relación entre dos variables, donde la variable dependiente es una función de las variables independientes, y el objetivo es calcular la constante o los coeficientes de los parámetros de intersección y pendiente para determinar la ecuación de regresión lineal.

1. Introducción

El aprendizaje automático (ML) se considera un subconjunto de la inteligencia artificial y una rama de la ciencia computacional que estudia y analiza e interpreta datos de reorganización de patrones y rastreo de datos. Para ayudar al sistema sin intervención humana se desarrollan algoritmos específicos en ML [2]. Por ello, al desarrollar un algoritmo en ML, este se basa en recomendaciones y decisiones basados en datos de entrada para el entrenamiento, si se detecta alguna modificación, el modelo desarrollado debería poder ajustarse para tomar mejores decisiones hasta que el algoritmo obtenga los resultados satisfactorios conocidos [5].

La técnica de modelado predictivo que se utiliza para determinar la relación entre dos variables, como el pronóstico de tendencias, el pronóstico de efectos, el valor crecimiento económico, el valor del precio del producto, las ventas de viviendas y del clima, el pronóstico de resultados, entre otros [1]. La regresión lineal en el modelo de cambio es regresión lineal (simple, múltiple, polinomial o no lineal). La regresión lineal (LR) depende principalmente de dos factores: qué variables predicen el resultado y cuán precisas son todas las predicciones [3].

El enfoque del aprendizaje supervisado es usar las variables de entrada y salida de un algoritmo para predecir el resultado. Si aparece una nueva variable de entrada. Un algoritmo de

regresión lineal en el ML es una técnica de aprendizaje supervisado que se aproxima a una función de mapeo para producir el mejor resultado [4].

El presente laboratorio tiene como objetivo documentar el código aplicando distintas técnicas de procesamiento de datos en el conjunto de datos para predicción de precios inmobiliarios, a su vez, haciendo uso de la ingeniería de características. Por ello, se identificará si el modelo está funcionando o no correctamente por medio del resultado de la precisión al entrenar el modelo, vinculado a esto, se espera que el código de aprendizaje automático falle en un cierto número de muestras.

2. Método

2.1. Regresión lineal

El objetivo principal de la regresión es construir un modelo eficiente para predecir atributos relacionados a partir de un conjunto de variables de atributos. Los problemas de regresión surgen cuando las variables de salida son valores reales o continuos (como salario, peso, área, etc.). En este caso se puede definir la regresión como una herramienta estadística para aplicaciones como vivienda, inversión, entre otros. Se utiliza para predecir la relación entre la variable dependiente y un conjunto de variables independientes.

2.1.1. Librerías

Para implementar técnicas de procesamiento es necesario especificar las librerías que se van a utilizar, en este caso utilizando el lenguaje de programación Python. A diferencia de otros lenguajes programación Python tiene una gran cantidad de librerías para tener acceso a funciones específicas. Las librerías para manipular y analizar son de gran utilidad ahorrando tiempo de codificación y no indagando tanto en temas netamente estadísticos para su implementación. Es por este motivo que en el Listing 1 se dan uso de las librerías a utilizar en el algoritmo.

```
""" import numpy as np para trabajar con matrices y librerias de scipy
    para la regresi n lineal y matplotlib para graficar
    import matplotlib.pyplot as plt para graficar y se importar  la librer a
    pandas para trabajar con dataframes
"""
import pandas as pd#para trabajar con dataframes y librer a scipy para la
    regresi n lineal
import matplotlib.pyplot as plt#para graficar y se importar  la librer a numpy
    para trabajar con matrices
from sklearn import linear_model # para la regresi n lineal y librer a scipy
    para la regresi n lineal
from sklearn.metrics import r2_score # para calcular el coeficiente de
    determinaci n
from sklearn.preprocessing import StandardScaler#para estandarizar la data y
    librer a scipy para la regresi n lineal
```

Listing 1: Importación de librerías

2.1.2. Dataset

En el Listing 2 se puede ver la manera de cómo se implementa en el código para cargar el dataset “Realestate.csv” en un dataframe df con pandas y se lee como csv. También, se imprimen los primeros cisco registros del dataframe “df” para verificar que se cargó correctamente el dataset con la función “head()”.

```
#C digo para cargar el Dataset
"""Se cargar  el dataset de la siguiente forma: """
df = pd.read_csv('Realestate.csv')#se carga el dataset Realestate.csv
df.head() # se imprime el dataset para verificar que se cargo correctamente
```

Listing 2: Cargando el dataset

2.1.3. Caracterización del Dataset

Para saber el número de instancia del dataset se utiliza la función "count()". Por ello, el Listing 3 muestra la manera de cómo implementar para calcular el número de instancias en total del dataset "df" con pandas. Debido a esto, se muestra que el número de instancias en total es de 596 registros (596 instancias). Segundo, se utiliza la función "drop('Y house price of unit area', axis=1).info()" que imprime número de instancias en total del dataset para verificar que se cargó correctamente

```

#N mero de instancias en total
""" Se calcula el n mero de instancias en total del dataset"""
df.count() # se imprime el n mero de instancias en total del dataset df con
    pandas y se puede observar que el dataset tiene una columna de tipo
    categ rica y otra de tipo num rica

#Segundo

"""Se imprime el n mero de instancias en total del dataset para verificar que se
    cargo correctamente """
df.drop('Y house price of unit area', axis=1).info()
```

Listing 3: Número de instancias en total del dataset

2.1.4. Estadísticas de la variable objetivo

A continuación, en el Listing 4 se puede ver la manera de cómo se implementa en el código para mostrar la estadística descriptiva de la columna "Y house price of unit area" del dataframe df con pandas. La estadística muestra los resultados de count, mean, min, entre otros. Dando así una perspectiva de cómo está conformado la variable objetivo.

```

""" Estadistica descriptiva del dataset de Y house price of unit area"""
df[['Y house price of unit area']].describe() #se imprime la estadistica
    descriptiva de la columna Y house price of unit area del dataframe df con
    pandas
```

Listing 4: Estadistica descriptiva del dataset de Y

3. Results and Analysis

Evaluación del modelo

La Figura 1 muestra los resultados del modelo en el gráfico de dispersión de los datos de test y los datos predichos. Por lo tanto, la relación entre dos variables del conjunto de datos de las viviendas. Representa puntos de datos en un plano 2D o sistema cartesiano. En el gráfico, se muestra una fuerte correlación positiva general entre el diámetro de un árbol y su altura. También podemos observar un punto atípico, un árbol con un diámetro mucho mayor que los demás. La circunferencia de este árbol parece bastante corta.

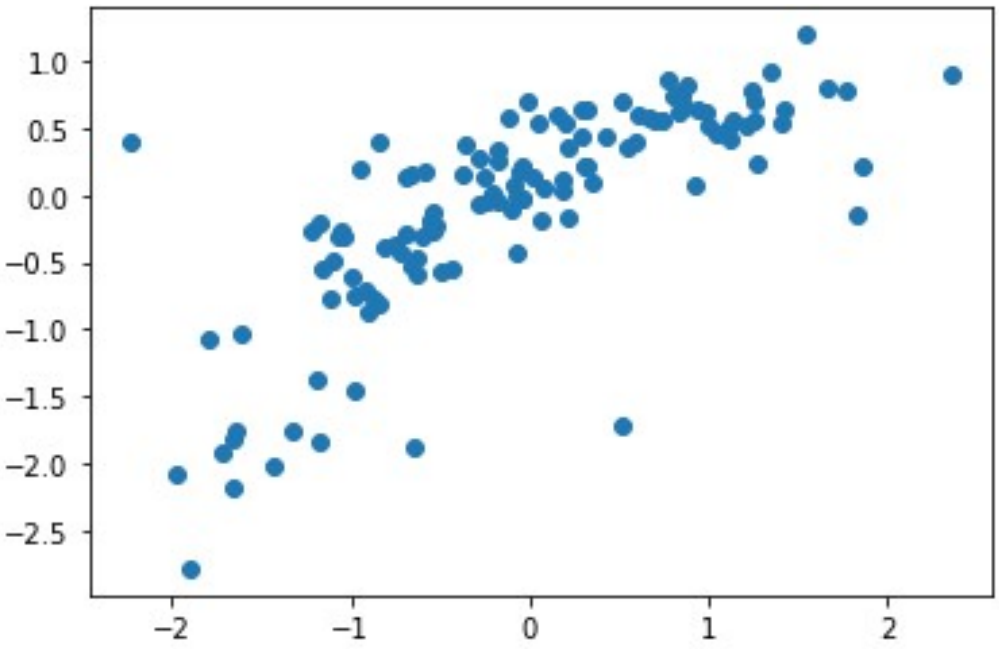

Figura 1: Visualización de resultados del modelo.

Precisión del modelo

Evaluación del modelo con el conjunto de datos de prueba y el resultado de la predicción del test para verificar que se cargó correctamente el modelo. Por ello, es bueno para predecir el valor de la variable objetivo y se importa la librería pandas para trabajar con data frames, a su vez, importa la librería numpy para trabajar con matrices y se importa la librería matplotlib para graficar y se importa la librería scipy para la regresión lineal y se importa la librería sklearn para el entrenamiento y prueba de los datos con el conjunto de datos de entrenamiento y la librería sklearn.modelSelection para la división de los datos en entrenamiento y prueba. Lo anteriormente expuesto, el resultado de la presión de la variable objetivo es de "0.56006". Hay que tener en cuenta que el dataset ocupado fue predicción de precios inmobiliarios.

4. Discusión

En este trabajo al determinar la relación entre las variables test y los datos predichos por el modelo en el gráfico de dispersión, se puede encontrar que el valor de precisión (0.560064) a partir del conjunto de datos de entrenamiento. Esto quiere decir que en caso de trazar una línea recta los variable estarán en su mayor parte cerca, lo cual quiere decir que hay una fuerte relación. Frente a lo mencionado se puede decir con certeza que si existe una fuerte relación entre las variables. En tal sentido, bajo lo referido anteriormente y al analizar estos resultados, confirmamos que mientras las variables estén más alejadas de la recta tendrían una menor relación lo cual no es es el caso para este trabajo.

5. Conclusión

En este trabajo se documentó el código aplicando distintas técnicas de procesamiento de datos en el conjunto de datos para predicción de precios inmobiliarios, a su vez, haciendo uso de la ingeniería de características. Por ello, se identificará si el modelo está funcionando o no correctamente por medio del resultado de la precisión al entrenar el modelo, vinculado a esto, se espera que el código de aprendizaje automático falle en un cierto número de muestras.

Un modelo de regresión lineal simple establece una relación entre dos variables, donde la variable dependiente es una función de las variables independientes, y el objetivo es calcular la constante o los coeficientes de los parámetros de intersección y pendiente para determinar la ecuación de regresión lineal. La ecuación encontrada se usa para estimar el valor de la variable dependiente en condiciones de un posible cambio en la variable independiente, es decir, la ecuación se usa para hacer una predicción o pronóstico.

Referencias

[1] Odd O Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

[2] R Montero Granados. Modelos de regresión lineal múltiple. *Granada, España: Departamento de Economía Aplicada, Universidad de Granada*, 2016.

[3] Jurgen Gross and Jürgen Groß. *Linear regression*, volume 175. Springer Science & Business Media, 2003.

[4] Takashi Isobe, Eric D Feigelson, Michael G Akritas, and Gutti Jogesh Babu. Linear regression in astronomy. *The astrophysical journal*, 364:104–113, 1990.

[5] George AF Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2012.