UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

CC-3074 Data Mining

EJERCICIO No. 4
Regression Analysis

Carga de dataset

Descargue el siguiente dataset en su computadora y defínalo dentro de su ambiente de desarrollo: Insurance prediction | Kaggle

Exploración de datos

Utilizando las diferentes librerías que ya conoce, realice una exploración de los datos guiándose (más no limitándose) por las siguientes tareas:

- Explore la información del dataset: estructura e información de los campos que contiene.
- Visualice los campos numéricos en una gráfica de pairwise donde se pueda apreciar la distribución de los mismos.
- Para los campos no numéricos puede graficar histogramas para conocer su distribución, o incluirlos en los pairwise con claves de color.

Preparación de los datos – datos faltantes

Antes de aplicar cualquier algoritmo debemos revisar si hay valores faltantes y definir alguna manera de manejarlos.

- Lo primero es hacer un conteo de cuántos nulos se encuentran en cada campo
- Luego debemos definir la acción que tomaremos para los casos encontrados. Podríamos elegir entre:
    - Ignorar el campo (hacer drop a la columna completa) donde exista un nulo

- Ignorar la observación (hacer drop a la fila completa) donde exista un nulo
- Llenar los valores nulos.

Vamos a proceder con la tercer opción utilizando el valor promedio en los campos numéricos y la moda en los no numéricos donde debamos aplicar las transformaciones.

Preparación de los datos – datos categóricos

Dado que estaremos aplicando una regresión lineal, es necesario que codifiquemos nuestros datos categóricos. Existen dos tipos de codificación que seguramente se incluyen en la librería que esté utilizando según el lenguaje y herramienta con el que esté trabajando. Estos son: Label Encoding y On Hot Encoding; el primero para codificaciones binarias y el segundo para cuando la variable puede tener tres o más valores.

Utilice un Label Encoder para codificar los campos de **sex** y **smoker**, y el On Hot Encoder para el del **region**.

Dividir en training y test

Vamos a dividir nuestro dataset en training y test. La variable dependiente que vamos a querer predecir es el campo **charges**.

Preparación de los datos – escala

Esta parte de la preparación la estamos haciendo posterior a la división de train y test debido a que se debe realizar un análisis sobre el total de datos y en este debería influir sólo la información del train.

Debemos tener todas nuestras variables independientes en una misma escala para conseguir mejores resultados y que el resultado de la regresión no esté sesgado por una diferencia de escalas. Esto lo podemos hacer con normalización y con estandarización.

Cuando usamos normalización hacemos un *rescale* a un rango de [0, 1] y cuando aplicamos estandarización hacemos el *rescale* colocando la media en 0 y una desviación estándar de 1. Cualquiera de los dos se encuentran en las librerías normalmente y se

pueden encontrar como MinMaxScaler y StandardScaler respectivamente. Para este ejercicio vamos a utilizar el segundo.

Modelación Lineal

Una vez tenemos nuestra preparación de datos completa, vamos a hacer tres modelos de regresión lineal diferentes sobre los mismos datos: el primero sin apoyarse en una librería y el segundo utilizándola. Para ambos modelos puede considerar únicamente la variable **bmi** entre los features. Para el tercero utilice más variables en su vector de features.

Para el primero será necesario que implemente el algoritmo de least squares manualmente y mediante un loop encontrar la ecuación lineal que minimice el cálculo de los *least squares*. No tiene que ser un programa muy complejo, simplemente es para que aplique su aprendizaje sobre el algoritmo.

Para el segundo y tercer modelos sí se puede apoyar en la librería que esté utilizando.

Evaluación de los modelos

Para completar nuestra evaluación de los tres modelos, primero interprete el resultado obtenido en los coeficientes y determine cuáles son los atributos más significativos revisando los p values de los mismos (esto tendrá más relevancia en el tercero pero de igual manera interprete los resultados de los primeros dos).

Grafique los puntos de test en conjunto con su línea de regresión para ver el resultado (para los primeros dos que consideran un único feature).

Calcule para ambos el Mean Absolute Error (MAE), Mean Squared Error (MSE) y el Root Mean Squared Error (RMSE) para emitir un juicio sobre ambos modelos y poderlos comparar.

Regresión Polinomial

Repita el proceso de los modelos dos y tres pero considerando una regresión polinomial de grado 2. Tome en consideración que es probable que la librería que utilice le solicite utilizar un vector distinto para los atributos (por ejemplo PolynomialFeaturesX). Grafique nuevamente los puntos del test con la curva e imprima los resultados de la evaluación.