Practica 2: Selección y Regresión

Una empresa de seguros médicos ubicada en Estados Unidos decide crear una herramienta en su nuevo sitio web que le permita a potenciales usuarios estimar cuál será el costo de la prima. Dicha información sera estimada por técnicas de Machine Learning y se busca que el mismo sea generado con la mayor precisión posible.
Se nos provee un dataset con aproximadamente 1300 datos de clientes reales con datos sobre su salud y ubicación, además del monto que pagó ese asegurado:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Los datos presentados son reales y fueron obtenidos de: https://www.kaggle.com/mirichoi0218/insurance

data_preparation.py

Dado que tenemos variables categóricas como sex, smoker o region debemos transformarlas en variables dummy. Reutilizaremos funciones que trabajamos en data_analysis.py de la practica 1. Les proveemos un `main()` para que visualicen la información.

`set_dummy_variable(dataframe, column, label)`

Dado un data frame, utilizando la función `pd.get_dummies(column, prefix=label)` generar las columnas del tipo dummy. Luego insertarlas en el data frame original manteniendo el indice que poseía la columna que fue remplazada (eliminar la columna original que generó las dummy). Retornar el dataframe.
Para insertarlas usar `dataframe.insert(index, name, values, True)`

(hint: `dataframe.columns.get_loc(label)` devuelve el indice como numérico).

`add_dummies(dataframe, columns)`

Generar un programa que empleando la función anterior, aplicar para todo un array de columnas que contenga aquellas que queremos generar las dummy

insurance-estimator.py

Es hora de desarrollar el programa que estimará el costo de las primas. Para ello debemos seleccionar qué variables incluirá nuestro modelo. Usaremos los tres algoritmos provistos en clase y será decision de ustedes, los científicos de datos, qué esquema será el que defina el modelo.

`forward_stepwise_selection(dataframe)`

Dado un Data frame válido, esta función debe emplear el algoritmo de Forward Stepwise Selection provisto en clase y retornar un arreglo de los R^2 ajustado que se generaron y las variables seleccionadas

`backward_stepwise_selection(dataframe)`

Dado un Data frame válido, esta función debe emplear el algoritmo de Backward Stepwise Selection provisto en clase y retornar un arreglo de los R^2 ajustado que se generaron y las variables seleccionadas

`backward_stepwise_selection_pvalues(dataframe)`

Dado un Data frame válido, esta función debe emplear el algoritmo de Backward Stepwise Selection with p values provisto en clase y retornar un arreglo de los R^2 ajustado.

`r2_variation(vars_size, r2_adj, title, x_label, y_label)`

Implementar la función vista en clase.

`create_model(r2_adj, var_model, dataframe, target, mode)`

Dados los valores que entrega el algoritmo de selección, más la información sobre el data frame crear el modelo de regresión lineal correspondiente. Se indica el modo de algoritmo que se empleó para poder definir el accionar de eleccion de variables.

Nuestro programa en acción

¿Cuál sería el costo estimado de seguro para Helen (residente en NY -noreste de EEUU-) de 25 años de edad, sin hijos, fumadora con un BMI/IMC de 26.3?

Tu respuesta: https://forms.gle/2ziHQcEPeQViCafb9

Si tu solución es destacable por su estilo, ingeniosidad o forma de resolver el problema te invito a que la presentes en los primeros 10 minutos de clase.

** La solución creada por mí será lanzada el lunes 6/9 antes de clase.*

Autoría de esta práctica: Tomas Nozica