Practica 4: Regresión Logística y kNN

El ente de salud pública de la ciudad de Framingham, MA en Estados Unidos se propuso realizar una investigación que permita predecir la probabilidad de que una persona sufra una enfermedad cardiovascular en 10 años. Para ello, recopiló datos de 4 mil pacientes. Se les solicita que formulen un modelo que a partir del dataset provisto infiera si una persona sufrirá este tipo de enfermedad con una probabilidad mayor o igual al 75%.

Se nos provee la siguiente descripción del problema:

Demographic:

• **Sex**: male or female (Nominal)

• **Age**: Age of the patient; (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral

• **Current Smoker**: whether or not the patient is a current smoker (Nominal)

• **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical (Continuous)

• **BP Meds**: whether or not the patient was on blood pressure medication (Nominal)

• **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)

• **Prevalent Hyp**: whether or not the patient was hypertensive (Nominal)

• **Diabetes**: whether or not the patient had diabetes (Nominal)

Medical (current)

• **Tot Chol**: total cholesterol level (Continuous)

• **Sys BP**: systolic blood pressure (Continuous)

• **Dia BP**: diastolic blood pressure (Continuous)

• **BMI**: Body Mass Index (Continuous)

• **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

• **Glucose**: glucose level (Continuous)

Predict variable (desired target)

• 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

Los datos presentados son reales y fueron obtenidos de: https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression. El dataset es real, la situación problemática es ficticia y puede ser errónea / incompleta.

data_analysis.py

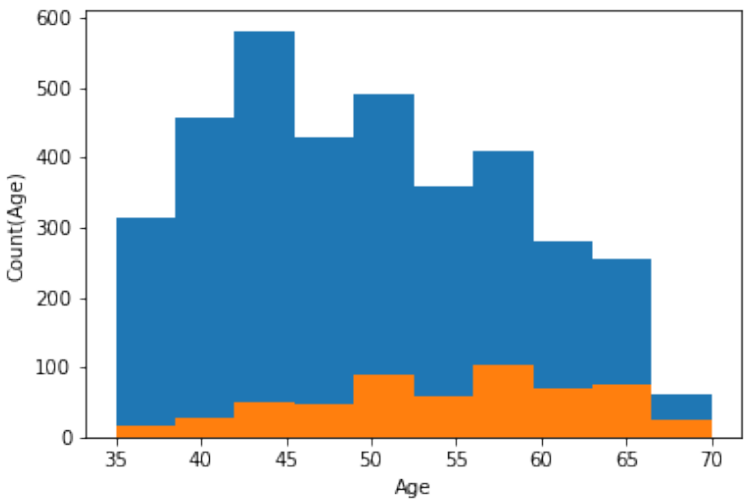
Modificaremos nuestro programa realizado en trabajos anteriores para visualizar nuestra data de manera más eficiente para el caso de las variables clasificadores

`plot_histogram(dataframe, ind_variable, target, bins)`

La función `plt.hist(x_values, bins=bins, stacked=True, range=(start, end))` nos permitirá plotear un histograma que, para un valor X de nuestro dataframe, nos muestre dividido en n `bins` (o columnas) la cantidad de valores en un rango de ancho fijo. Con el parámetro `stacked` en `True` podremos ver histogramas superpuestos (muy útil para comparar) y con `range` fijamos valores de inicio y fin para demarcar como agruparemos la data. Es muy importante esto último ya que para apilar histogramas debemos tener en todos los casos el mismo valor de `start`, `end`.

El objetivo es que *ploteen* (disculpen el spanglish) para una `ind_variable` la dispersión total de los valores agrupados en n columnas y superpuesto a éste la misma ind_variable pero solo en los casos que TenYearCHD sea igual a 1 (Recordamos la función `dataframe.loc[ condición ]`).

El output esperado es:



chd-prediction.py

`train(dataframe, target)`

Utilizando la librería sklearn entrenar un modelo de Regresion Logística, emplear las variables que consideren como variables predictoras. El parámetro objetivo es dato. Retornar el modelo entrenado.

`predict(x, model, threshold=0.5)`

Dado un X que contiene valores que queremos predecir (toma la forma de un arreglo, incluso siendo un solo valor) y dado el modelo de regresión lineal más el umbral donde queremos establecer que un valor sea 0 ó 1 retornar un arreglo que devuelva las predicciones para cada input (aunque sea un solo valor)

`report(x, predictions)`

Dado un X que contiene valores que utilizamos para la predicción, más el valor que retorna la predicción. Crear el archivo CSV que se utilice como reporte.

Nuestro programa en acción

Crear un reporte que prediga para los siguientes pacientes si pueden sufrir o no en los próximos 10 años una enfermedad cardiovascular

Paciente: Aldo Bueno (id 1)

Edad: 45
Nivel educativo: Universitario (4)
Fuma? No
Consume medicamentos de la presión? No
Sufrió algún ataque cardíaco? Si
Hipertensión? Si
Diabetes? No
Colesterol: 205
Presion Arterial: 125 / 110
Indice de Masa Corporal: 26
Ritmo Cardiaco medido: 85
Glucosa en Sangre: 85

Paciente: Bianca Bueno (id 2)

Edad: 43
Nivel educativo: Universitario (4)
Fuma? Si
Consume medicamentos de la presión? No
Sufrió algún ataque cardíaco? No
Hipertensión? Si
Diabetes? Si
Colesterol: 250
Presion Arterial: 155 / 130
Indice de Masa Corporal: 28
Ritmo Cardiaco medido: 90
Glucosa en Sangre: 88

Enviá tu archivo CSV con las predicciones para cada id de paciente: https://forms.gle/ifddCsxJEhNoa5HL9

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 27/9 antes de clase.*

Autoría de esta práctica: Tomas Nozica