

Práctica 4.

Regresión lineal

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Montse Llos i Bombardó*

*Práctica 4.1*

El objetivo de esta práctica es recrear el funcionamiento de una regresión lineal.

Para ello, cargaremos en Jupyter Notebook el fichero 0. Linear Regression - How works.ipynb que encontraremos en la carpeta de Práctica 3 de Recursos del Campus.

Disponemos de un fichero con el peso y altura de varias personas. El objetivo es crear una función que permita responder al peso de una persona dada la altura.

El fichero sobre el que haremos el proceso de aprendizaje es el fichero en formato CSV weight-height.csv

*Práctica 4.2*

Usaremos la función de linear regression de la librería sklearn.linear\_model para aprender a predecir el peso de una persona dada la altura y el sexo.

*Ejercicio 4.3*

*Opcional*

Probar otras librerías de Python con la función de regresión lineal

Número límite de personas: 2

Entregables:

* Un fichero con el código Python
  + Nombre del fichero: <Apellidos+InicalNombres>\_Ejercicio\_4.3.py o .ipynb
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_4.3.ipynb*
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio y conclusiones de comparación con la regresión lineal de sklearn
  + Nombre del fichero: <Apellidos+Nombres>\_Ejercicio\_4.3.pdf
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_4.3.pdf*

Fecha límite entrega: 17 febrero 2019 a las 24h

Enviar a: [montse.llos@campus.eae.es](mailto:montse.llos@campus.eae.es)

*Ejercicio 4.4*

*Opcional*

Probar los mismos datos con una regresión lineal dentro de RapidMiner

Número límite de personas: 2

Entregables:

* Un fichero .zip que contenga la carpeta del repositorio de RapidMiner con todos los ejercicios realizados.
  + Nombre del fichero: <Apellido+InicialNombre>\_Ejercicio\_4.4.zip
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio y conclusiones de comparación con la regresión lineal de sklearn
  + Nombre del fichero: <Apellido+InicialNombre>\_Ejercicio\_4.4.pdf

Fecha límite entrega: 17 febrero 2019 a las 24h

Enviar a: [montse.llos@campus.eae.es](mailto:montse.llos@campus.eae.es)

*Ejercicio 4.5*

En la inmobiliaria VivirMejor han recopilado datos de precios de las casas que han vendido últimamente.

Les gustaría poder predecir el precio de una casa a partir de los datos que disponen sobre características de la casa y del barrio donde está situada. Así pueden dar ya un primer precio de venta sin tener que llevar un tasador.

Las variables que vamos a usar son:

* ID
  + Identificador de la casa
* crim
  + tasa de criminalidad per cápita por ciudad.
* zn
  + proporción de terreno residencial dividido en zonas para lotes de más de 25,000 pies cuadrados.
* indus
  + proporción de acres de negocios no minoristas por ciudad.
* chas
  + Variable ficticia de Charles River (= 1 si el tramo limita con el río; 0 en caso contrario).
* nox
  + concentración de óxidos de nitrógeno (partes por 10 millones).
* rm
  + número medio de habitaciones por vivienda.
* age
  + proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
* dis
  + media ponderada de las distancias a cinco centros de empleo de Boston.
* rad
  + índice de accesibilidad a carreteras radiales.
* tax
  + Impuesto sobre el valor total de la propiedad por 10.000 dólares.
* ptratio
  + proporción de alumnos por maestro por ciudad.
* black
  + 1000(Bk - 0.63)^2 donde Bk es la proporción de negros por ciudad.
* lstat
  + un estatus más bajo de la población (porcentaje).
* medv
  + valor medio de las casas ocupadas por sus dueños en \ 1000 dólares.
  + Será la variable target a predecir

Tenemos 4 ficheros:

* 2 train
* 2 predict

Primero tendremos que hacer un proceso de ETL ya que los datos de la casa están separados de los datos del barrio (juntar los dos train entre ellos y los dos test entre ellos)

En predict, no tenemos la variable medv, que es la que tenemos que dar como respuesta.

Para llevar a cabo la predicción, usaremos Python. Es necesario que todas las líneas escritas en Python estén comentadas explicando qué se hace en cada paso.

Se puede usar cualquier algoritmo supervisado de regresión. No es necesario que sea de regresión lineal.

Número límite de personas: 2

Entregables:

* Un fichero .zip que contenga:
  + fichero de código Python para la creación del modelo y predicción,
  + fichero resultado\_pred.csv con el ID de la casa y la predicción resultante de aplicar el modelo sobre los datos predict.
  + Nombre del fichero: <Apellidos+InicalNombres>\_Ejercicio\_4.5.zip
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio y conclusiones
  + Nombre del fichero: <Apellidos+Nombres>\_Ejercicio\_4.5.pdf

Fecha límite entrega: 24 febrero 2019 a las 24h

Enviar a: [montse.llos@campus.eae.es](mailto:montse.llos@campus.eae.es)