Laboratorio: Resolver un problema de regresión

**Objetivos de la actividad**

En esta actividad vas a profundizar en las distintas técnicas que se pueden aplicar para abordar un problema de regresión. Además, profundizarás en tus conocimientos sobre las librerías *statsmodels* y *scikit-learn* de Python.

**Descripción de la actividad y pautas de elaboración**

El primer paso consiste en crear un conjunto de datos ficticio. Para garantizar que cada alumno obtiene uno distinto se va a emplear el documento de identidad de cada uno para crear el conjunto de datos. Para que sean comparables entre todos, si el número de identidad tiene menos de 8 cifras replicaremos las primeras hasta obtener exactamente 8. Además, para evitar los dígitos 0 y 1, si alguna de las cifras es menor que 2 la sustituiremos por ese número. Aplicando estos cambios tendremos el número del documento de identidad preparado para la resolución de la actividad. Veamos algunos ejemplos:

* Ejemplo 1:

12345678 -> 22345678

* Ejemplo 2:

304156 -> 304145630 -> 32425632

Una vez tenemos el número del documento de identidad preparado vamos a crear el conjunto de datos para el problema de clasificación con la función *sklearn.datasets.make\_regression* de la librería *scikit-learn* (más información en: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_regression.html>).

* Crea el conjunto de datos empleando los siguientes argumentos:
  + n\_muestra = 200 + 10 · primer dígito dni
  + n\_predictores = 10 + segundo dígito dni + tercer dígito dni
  + n\_informativas = 10 + segundo dígito dni
  + sesgo = 2
  + ruido = 10 \* cuarto dígito dni
  + semilla = numero\_dni
  + shuffle = False

Este conjunto de datos tendrá al menos 220 observaciones y 14 variables predictoras, de las cuales al menos 2 no estarán relacionadas con la variable respuesta. Se añade el sesgo y el ruido para que la variable respuesta no sea una combinación lineal exacta de las variables informativas. Al definir shuffle = False estamos forzando a que las variables no informativas del conjunto de datos aparezcan al final.

* Divide el conjunto de datos en 200 observaciones para el entrenamiento y el resto para realizar la validación de los distintos métodos de regresión aplicados.
* Describe tu conjunto de datos (transformalo en un data.frame, aplica los métodos .info(), .describe() y obtén el histograma de todas las variables (predictoras y la variable respuesta).
* Obtén un modelo de regresión lineal múltiple. ¿Son todas las variables predictoras significativas? Utiliza la librería *statsmodels*.
* Realiza una selección de variables mediante un algoritmo de tipo *step-wise*, donde en cada paso elimines la variable predictora menos significativa atendiendo a su p.valor hasta que todas las variables del modelo sean significativas (p.valor < 0.05).
* Realiza una regresión mediante la red elástica. Prueba distintos valores de r y obtén el valor óptimo de r y de α mediante validación cruzada.
* Comprueba con la muestra de validación con cuál de los tres modelos se obtiene un menor error cuadrático medio.
* No olvides añadir las referencias utilizadas para la elaboración del trabajo al final de la memoria-

**Extensión** **máxima**: 20 páginas, con fuente Calibri 12 e interlineado 1,5.

**Rúbrica:**

|  |  |  |  |
| --- | --- | --- | --- |
| Resolver un problema de regresión | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | El conjunto de datos y su descriptivo son correctos. | 2 | 20% |
| Criterio 2 | El modelo de regresión lineal múltiple y la selección de variables se realizan de manera adecuada. | 3 | 30% |
| Criterio 3 | Se aplica correctamente el modelo de regresión de red elástica. | 3 | 30% |
| Criterio 4 | La validación se ha realizado correctamente. | 1 | 10% |
| Criterio 5 | Presentación y formato de la memoria de la práctica. Bibliografía. | 1 | 10% |
|  |  | **10** | **100 %** |