

Aprenentatge màquina (I). Máster en Ciencia de Datos. ETSE. Universitat de València

Tarea: Modelos de regresión lineal simple y múltiple

Ana Navarro Quiles

En la librería MASS puedes encontrar un famoso banco de datos llamado `Boston` que contiene información sobre 506 barrios de Boston, Massachusetts, en 1970. La base de datos contiene 14 variables relativas a 506 barrios. Para saber qué información está contenida en las variables puedes escribir `?Boston` (después de haber cargado la librería MASS).

En esta tarea trabajaremos con el conjunto de datos “`boston.xlsx`” que encontraréis en el aula virtual (con 199 datos y 11 variables).

Ejercicio 1: Considera la variable respuesta `crim` relacionándola con la variable `X` con la que tenga mayor relación lineal.

1. Evalúa el efecto de `X` sobre `crim`, gráficamente y numéricamente. Es decir, indica como es la relación (fuerza y tipo).
2. Obtén la recta de mínimos cuadrados. Interpreta los resultados obtenidos (coeficientes, significatividad, R^2 , contraste del modelo, etc...).
3. Dibuja el diagrama de dispersión, la recta de regresión y las bandas de confianza al 90 %.
4. Realiza un diagnóstico de los residuos. Si falla algunas de las condiciones, busca una (o varias) posible solución.

Ejercicio 2: Considera la variable respuesta `crim` relacionándola con el predictor `medv`.

1. Evalúa el efecto de `medv` sobre `crim`.
2. Obtén la recta de mínimos cuadrados. Interpreta los resultados obtenidos (coeficientes, significatividad, R^2 , contraste del modelo, etc...).
3. Dibuja el diagrama de dispersión, la recta de regresión y las bandas de predicción al 90 %.
4. Realiza un análisis de los residuos.
5. ¿Te parece adecuado haber realizado regresión lineal o es preferible otro tipo de regresión?. Ajusta el modelo que te parezca más adecuado.

6. ¿Qué tasa de criminalidad se espera para aquellos barrios con un precio mediano de la vivienda de 30000 dolares? ¿Y 10000? ¿Y 100000? Calcula e interpreta los intervalos de confianza y de predicción.

Ejercicio 3:

1. Encuentra el número óptimo de variables a incluir en un modelo predictivo de *crim*, según los criterios R^2 , BIC y CP, utilizando la metodología RegSubsets. Indica brevemente en que consiste esta metodología.
 - ¿Qué variables incluye el modelo obtenido? (Seleccionar el criterio que más os guste). Interpreta los coeficientes obtenidos. ¿Tienen todas sentido?. ¿Son significativos?.
 2. Selecciona el mejor modelo con el método stepwise. Indica brevemente en que consiste esta metodología y contesta a las siguientes preguntas:
 - ¿Qué modelo piensas que es mejor? (Entre este y el/los obtenido/s mediante Regsubsets).
 - ¿Qué % de la varianza de *crim* explica el modelo?
 - ¿Cuál es el efecto de la variable *chas* sobre *crim*?
 3. Con el modelo obtenido con stepwise, realiza el diagnóstico de tu modelo, sin emprender ninguna acción, e indica los problemas que presenta.
 4. Emprende ahora las acciones que te parezcan oportunas e indica los problemas que has conseguido solucionar o mejorar un poco.
 5. Obtén la predicción de la tasa de criminalidad para un barrio en la mediana de los predictores en el modelo escogido. *Notar que las variables categóricas se tratan de diferente manera, no hay mediana.*
-