

Práctica 1

Casas en Boston

Análisis de Datos

Grado en Ingeniería Informática

Curso 2020/21



Casas en Boston

- Queremos conocer el precio mediano de la vivienda en distintas zonas del área metropolitana de Boston (Massachusetts, EEUU)
- 506 instancias
- 13 variables
- Salida numérica (**problema de regresión**)

Variables



Se trata de un conjunto de datos antiguo que no refleja la realidad actual. Las variables están mejor explicadas en el artículo original:
Harrison Jr, David, and Daniel L. Rubinfeld. "Hedonic housing prices and the demand for clean air." *Journal of environmental economics and management* 5, no. 1 (1978): pp. 81–102

crim	tasa de criminalidad <i>per cápita</i>
zn	proporción de zona residencial en áreas de más de 25.000 pies cuadrados
indus	proporción de extensión (acres) de negocios (excl. comercio al por menor) en la ciudad
chas	presencia del Río Charles (binaria)
nox	concentración de óxidos de nitrógeno (partes por 10 millones)
rm	número medio de habitaciones por vivienda
age	proporción de viviendas ocupadas con anterioridad a 1940
dis	distancia ponderada a cinco centros de empleo en Boston
rad	índice de accesibilidad a autopistas radiales
tax	impuesto sobre el valor de la propiedad por 10.000\$
ptratio	ratio de alumno-profesor en la ciudad
black	$1000(Bk - 0,63)^2$, siendo Bk la proporción de población negra en la ciudad
lstat	porcentaje de la población que es de clase baja
medv	valor mediano de las viviendas habitadas, expresado en miles de dólares

Objetivo

Queremos entrenar un modelo de regresión que pueda estimar el precio de la vivienda basándose en los atributos disponibles.

Para ello, se realizará un proceso **sistemático** de análisis de datos.

A continuación se describen las tareas para alcanzar este objetivo de forma satisfactoria.

Tarea 1

Carga de los datos | 1 punto

En Aula Global se proporciona un fichero de texto con los datos.

Se debe examinar la estructura del fichero y determinar cómo realizar la carga en el cuaderno empleando Pandas.

Se debe comprobar que la carga es correcta, es decir, que el número de filas y columnas es el esperado.

Tarea 2

Análisis exploratorio | 3 puntos

Se deben examinar los datos de la forma que se considere más oportunas (resúmenes estadísticos, visualizaciones, etc.)

Este análisis exploratorio debe ir encaminado al problema que se quiere resolver (regresión de la variable **medv**).

En cualquier caso, las decisiones tomadas se deben justificar, y de los análisis realizados se deben obtener las conclusiones oportunas (*por ejemplo, no basta con dibujar una determinada visualización, sino que debe explicarse qué se concluye de ella*).

Tarea 3

Construcción de un modelo de regresión lineal | 2 puntos

Con los datos disponibles, se construirá un modelo de **regresión lineal** y se evaluará sobre el mismo conjunto de entrenamiento.

¿Cómo se comporta este modelo? Se deben reportar al menos dos métricas:

- Coeficiente R^2
- RMSE (raíz del error cuadrático medio)

Tarea 4

Mejora del modelo de regresión lineal | 2 puntos

Se investigará cómo mejorar el resultado del proceso de regresión lineal. Para ello, los estudiantes deben realizar **al menos una** de las siguientes tareas:

- Investigar y probar alternativas a la regresión lineal que empleen regularización de los parámetros, ya sea L1 (*Lasso*), L2 (*Ridge*), o ambas.
- Reducir el número de variables empleadas en la regresión lineal. En ese caso, se debe indicar qué variables se emplearán y por qué.
- Creación de nuevas variables *sintéticas* que combinen las ya existentes.

En cualquiera de los casos se tratará de concluir si el resultado es mejor o peor que el obtenido por la regresión lineal y se tratará de averiguar (o aventurar) el porqué.

Nota: se podrá alcanzar la máxima calificación realizando solo una de las tareas propuestas, no es necesario realizarlas todas. Tampoco es necesario mejorar las métricas, pero sí realizar un análisis crítico de los resultados obtenidos.

Tarea 5

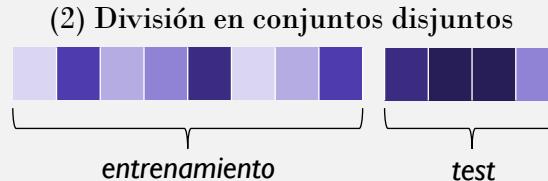
Generalización del modelo de regresión lineal | 2 puntos

El enfoque de evaluación del modelo empleado anteriormente es muy limitado, pues no permite evaluar la capacidad del modelo de generalizar a datos nuevos. Un proceso más adecuado involucra los siguientes pasos:



Al trabajar con procesos que introducen ! aleatoriedad, se pierde la *reproducibilidad* de los experimentos.

Para evitarlo, es importante controlar bien las semillas aleatorias que se emplean.



En esta tarea se pide realizar este proceso, empleando un 70% de los datos para entrenamiento y un 30% para test.

Se debe volver a entrenar el modelo de regresión (con los datos de entrenamiento) y evaluarlo sobre el conjunto de test. ¿Cómo cambian los resultados? ¿Funcionan mejor (o peor) las mejoras probadas en la tarea 4?

Tarea Extra

Investigación de otros modelos | +1 punto

En esta tarea **opcional** se permitirá a los estudiantes probar con otros modelos de regresión diferentes (incluso si no han sido vistos en clase).

También podrán realizar cualquier otro procesamiento de los datos que consideren oportuno con el fin de intentar mejorar los resultados de la regresión.

En cualquier caso, es importante tratar de justificar los resultados obtenidos, o explicar por qué se considera que pueden ser mejores o peores (no basta con escribir y ejecutar el código).

Evaluación

Cada tarea será evaluada con una puntuación, cuyo valor máximo viene indicado en las diapositivas anteriores.

Sean T_1, \dots, T_5 las puntuaciones obtenidas por un equipo de prácticas en las cinco primeras tareas, y T_E la puntuación obtenida en la tarea extra opcional. La nota de la práctica se calculará como:

$$P = \min(T_1 + T_2 + T_3 + T_4 + T_5 + T_E, 10)$$

Por lo tanto, la máxima nota alcanzable en la práctica es un 10.

Para evaluar cada tarea, se tendrá en cuenta que se haya implementado el código requerido, y que se hayan explicado las decisiones tomadas y los resultados obtenidos.

No es fundamental saber explicar todos los resultados de forma precisa, pero sí se debe intentar entenderlos.



Entrega

- Se debe trabajar **individualmente o en parejas**.
- El trabajo vale 1 punto sobre la evaluación continua.
- **¡Importante!** Se debe revisar que todo el código funcione.
 - Es útil, antes de entregar la práctica, reiniciar el entorno y ejecutar todas las celdas.
- **Entrega:**
 - Viernes 30 de octubre a las 23:59 a través de Aula Global.
 - Se entregará el cuaderno en formato IPYNB (y, si es posible, también en PDF).