Trabajo práctico 4: Validación Cruzada

Laboratorio de Datos
Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires

Verano 2022

Este trabajo práctico debe entregarse en un `notebook` de **R**. Intercale texto código y gráficos. Asegurese de incorporar a la presentación de código lo que usted aprendió y las conclusiones que obtuvo del análisis. No todas las exploraciones necesitan estar presentes en el notebook final, sólo retenga el contenido que considere necesario. Pienselo como un informe o una historia que narra a alguien interesado en aprender del dataset.

Vamos a trabajar con un subconjunto del dataset de Properati que usamos para la práctica de modelo lineal (`datos_alquiler_crossvalidation.csv`).

El objetivo será analizar los datos teniendo en cuenta un subconjunto de *variables de interés*: el tipo de propiedad, su superficie (cubierta y fondo), cantidad de ambientes, precio, fecha de publicación (`start_date`) y la ubicación (`lat-lon`).

1. Implemente una función que compute MAE:

$$MAE(x, y) = \frac{\sum_i |x_i - y_i|}{n}$$

   donde x es el vector de n observaciones e y el vector de predicciones.

2. Implemente una función que compute PMAE:

$$PMAE(x, y) = \frac{\sum_i |x_i - y_i|}{\sum_i x_i}$$

   donde x es el vector de observaciones e y el vector de predicciones.

3. Considere el modelo que ajusta el precio en función de la superficie cubierta. Calcule el MAE y PMAE. Agregue la variable fondo y compare.

4. Construya una función `crossval(datos, modelo, n_obs, fun_error, n_muestras=10)` para calcular el error promedio de predicción haciendo validación cruzada. La misma debe recibir como parámetros:

   `datos` que es el dataset a utilizar,

   `modelo` que representa una fórmula (que debe construirse con la función `formula` y un string, por ejemplo: `formula('x ~ y')`) que será input del modelo lineal (invocado mediante `lm`),

   `n_obs` que indica el número de muestras que se usará para evaluación,

   `fun_error` que será la función que se usará para evaluar el modelo[1],

   `n_muestras` que indica la cantidad de veces que se debe repetir el procedimiento de muestreo.

   La función debe seleccionar al azar `nrow(datos)-n_obs` observaciones (para elegir estas muestras explore el procedimiento `sample`); con este subconjunto debe ajustar empleando la formula provista en `modelo`; luego debe computar, para las `n_obs` que fueron excluidas de la construcción del modelo, el error de predicción utilizando `fun_error`. Este procedimiento se debe repetir `n_muestras` veces. La función debe retornar una lista con los errores obtenidos, el error promedio, su varianza, la fórmula del modelo empleado y el modelo ajustado usando todos los datos.

[1] Pasar una función como input permite cambiar la función que usaremos para calcular el error

5. Utilizando las funciones anteriores evalue el comportamiento de un modelo que ajuste el precio en función de la superficie cubierta utilizando validación cruzada. Use como función de error al PMAE.

6. Considere el modelo que ajusta el precio en función de la superficie cubierta. Explore como varía el error al usar valores de `n_obs` iguales a distintos porcentajes del tamaño del dataset, por ejemplo `seq(1,100,5)` y `n_muestras=100`. Grafique el error de validación cruzada en función de la cantidad de observaciones separadas. ¿Qué le indica esto sobre la cantidad de observaciones que debe usar para validar el modelo?

7. Construya modelos usando las potencias de la variable `fondo`. Es decir, si p es el precio y f la el fondo, considere modelos:

$$p = \sum_{i=0}^{N} a_i f^i$$

con $N = 1, \ldots, 8$. Considere como valor de `n_obs` un 20 % de dataset y `n_muestras=20` Compare sus desempeños.¿Qué modelo tiene menor de ajuste? ¿Y cuál tiene mayor error de predicción? Grafique el error de predicción promedio y el error de ajuste en función del grado del polinomio. *Tip: Para escribir* `y ~ x+x^2` *puede usar* `y ~ poly(x,2)`. Use un ciclo para automatizar la construcción de los distintos modelos.

8. Compare la calidad de predicción de 3 modelos: precio en función de superficie cubierta, precio en función de fondo y precio en función de ambas variables. Emplee en su comparación el error de ajuste y el error de predicción.

9. Construya un programa que calcule el error de los 16 modelos posibles que incluyen las variables superficie cubierta, fondo, tipo de propiedad y ubicación (considere las variables latitud y longitud conjuntamente). Compare los modelos en términos de su error de ajuste y su error en validación cruzada, usando un 20 % de los datos para validar el modelo. ¿Cuál es el ranking entre los modelos? ¿Considerar más variables siempre mejora la predicción?