Aprendizaje automático

Práctica 2. Regularización y Selección de Modelos

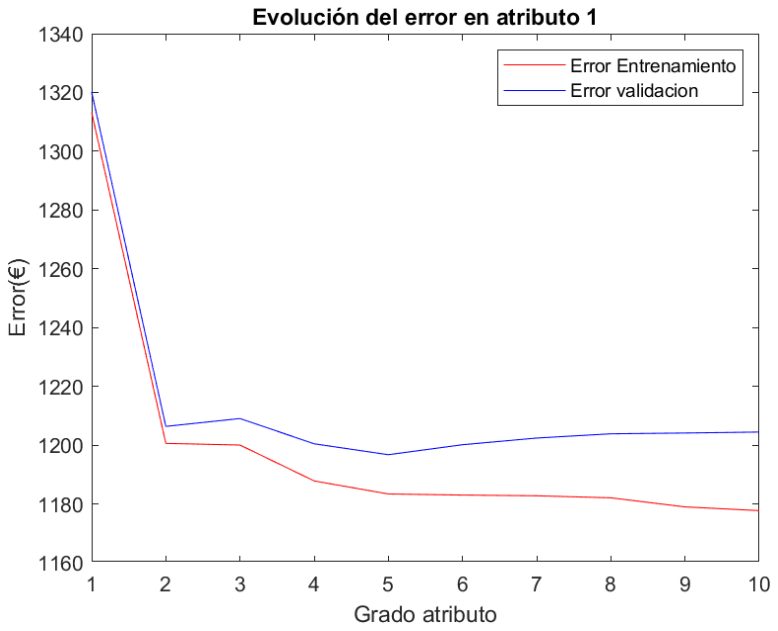
Universidad de Zaragoza, curso 2022/2023

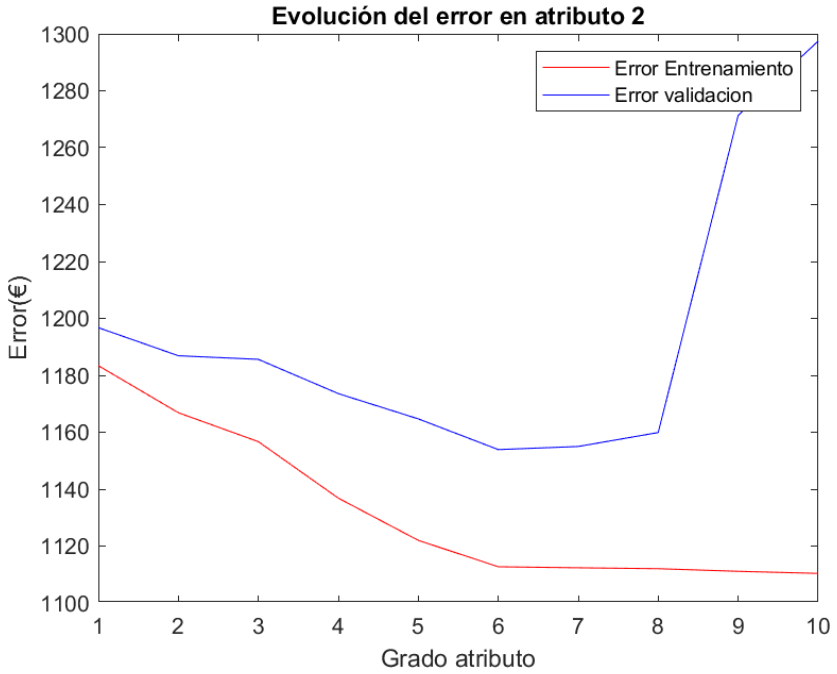
Juan Eizaguerri Serrano 816079

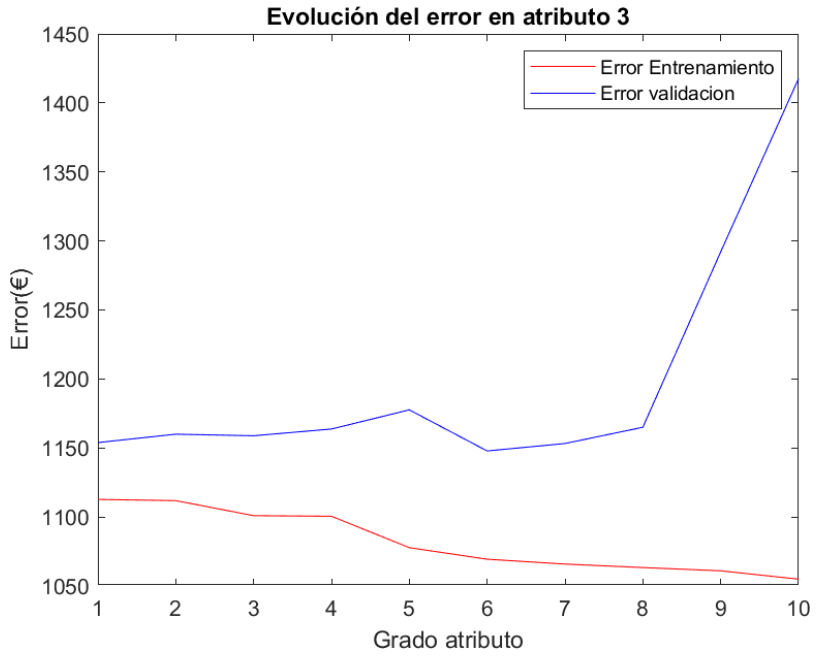
Selección de modelos mediante búsqueda heurística

Se ha utilizado el algoritmo de k-fold cross-validation para buscar los grados óptimos del polinomio fijándolos uno a uno. Para cada iteración del las k particiones de datos, se expanden los datos de entrenamiento con los grados del polinomio que se quieren probar, se normalizan para que estén en la misma escala y afecten en la misma medida en el entrenamiento, y se calcula el vector de pesos mediante la ecuación normal. Una vez calculados los pesos se desnormalizan para poder utilizarlos con los datos de entrada no normalizados. Se calcula la media del error RMSE con los datos de validación para cada valor probado de los atributos, y se guarda el mejor. Además, en cada iteración se realiza una partición de los datos en datos de entrenamiento y datos de validación, para comprobar sobre-ajuste y sub-ajuste.

Se ha utilizado el algoritmo con grado máximo 10 y pliegues k=5 y k=10, obteniendo exactamente los mismos resultados: Grados óptimos (5, 6, 6). A continuación se muestra la evolución del error RMSE con los datos de entrenamiento y validación para los distintos grados de los polinomios.







Cabe a destacar que el grado de los atributos se fija de izquierda a derecha, lo que quiere decir que en la primera figura se muestra el error de los datos con grados (1,1,1) a (10,1,1), para la segunda ya se ha fijado el primer valor a 5, por lo que se prueban los grados (5,1,1) a (5,10,1), y para la tercera se prueban de (5,6,1) a (5,6,10).

Una vez seleccionado el mejor modelo, se entrena con todos los datos de entrenamiento, sin particiones, también mediante la ecuación normal, expandiendo y normalizando los datos de entrada, y desnormalizado después los pesos calculados. Se evalúa tanto para los datos con los que se ha entrenado como para los datos de test. Como métricas de error se utiliza RMSE y MAE, menos afectada por espurios.

	RMSE (€)	**MAE (€)**
Datos entrenamiento	1076	804
Datos test	1009	763

Se observa que los errores para los datos de test son cercanos a los de entrenamiento, por lo que no se ha producido sobre-ajuste, y además son razonablemente buenos para el problema (700-1000 € de error cuando se trata de precios de coches no es demasiado grande) así que tampoco hay sub-ajuste. Cabe destacar que el error RMSE para los datos de test es incluso más bajo que para los datos de entrenamiento, lo que se puede deber a cómo se ha realizado el corte en los datos originales.

Selección de modelos mediante búsqueda exhaustiva (grid search)

El objetivo de este apartado es encontrar los grados de los polinomios óptimos para los atributos de entrada mediante una búsqueda exhaustiva que utilice k-fold cross-validation.
Se ha desarrollado una implementación recursiva que pueda funcionar con cualquier número de atributos para que no sea específica al problema de esta práctica.

la búsqueda exhaustiva calcula la media del error con los datos de validación de los k pliegues de datos de entrada para todas las posibles combinaciones de grados de los polinomios y guarda el mejor.

El número de posibles combinaciones es de ($grados\_max^{nº\ entradas}$). Este algoritmo tiene un coste exponencial y puede ser muy costoso para problemas con un número de parámetros elevado.

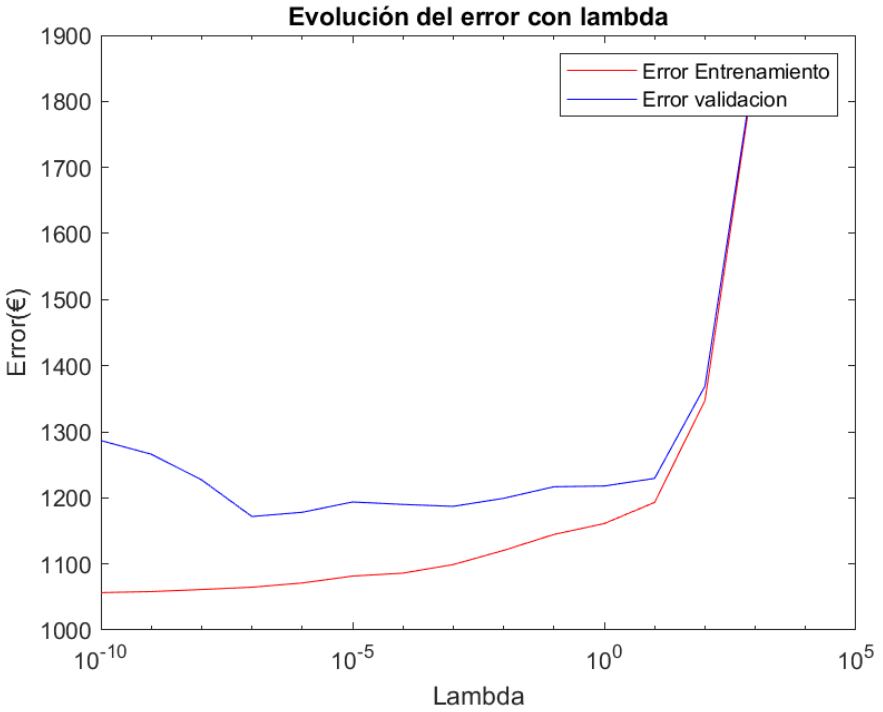
La búsqueda exhaustiva llega a la misma solución que la búsqueda heurística, grados (5, 6, 6), que produce una media de error RMSE de 1147.84€. Al tratarse del mismo modelo que el del apartado anterior, la evaluación tras el entrenamiento con todos los datos es la misma.

	RMSE (€)	**MAE (€)**
Datos entrenamiento	1076	804
Datos test	1009	763

Se pueden extraer las mismas conclusiones.

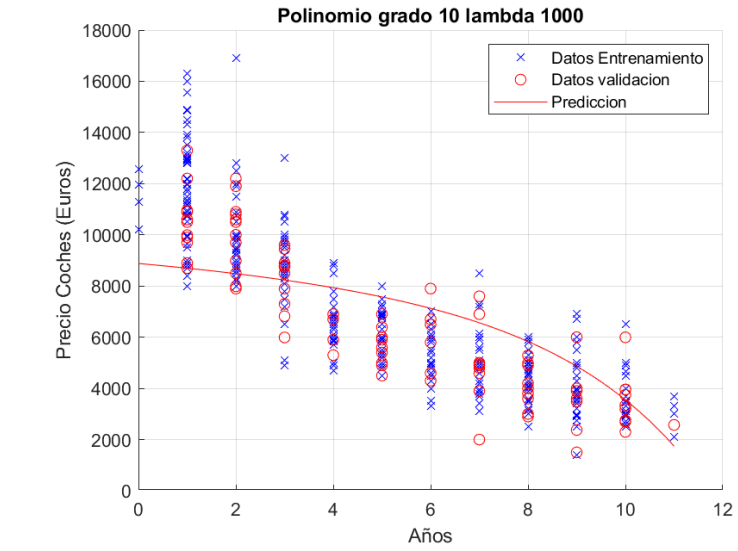
Regularización

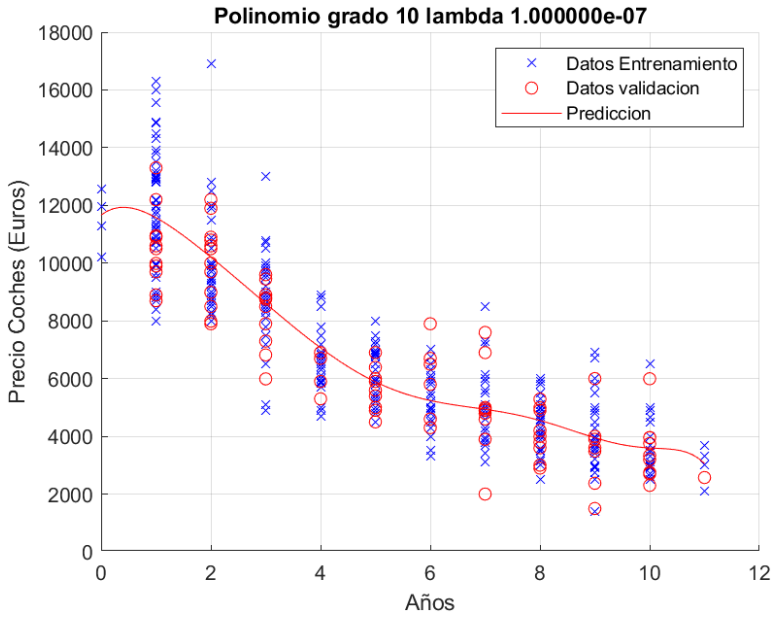
En este apartado se va a utilizar el k-fold cross-validation para encontrar el mejor valor posible de λ para un modelo de grado 10 para todos los atributos, en lugar de para encontrar los grados del polinomio. Se van a probar valores en progresión geométrica entre 10^{-10} y 10^{5} con k = 5. A continuación se muestra la gráfica de la evolución del error RMSE para los distintos valores de λ probados, para los datos de entrenamiento y validación.

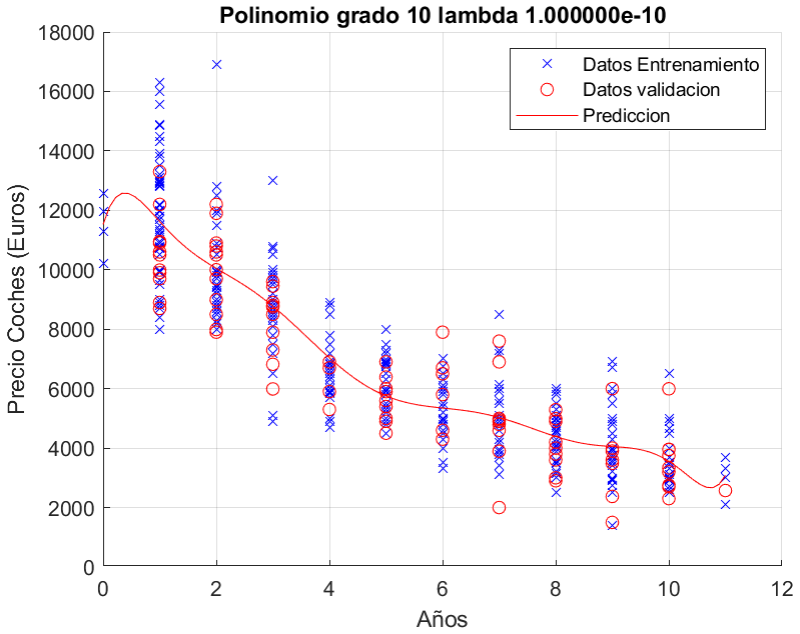


La mejor λ encontrada es 10^{-7}, que da lugar a una media de error RMSE con los datos de validación de 1171.8€.

Dado que la salida del problema está en los miles, los valores altos de λ tienen un efecto demasiado grande en el coste a la hora de entrenar teniendo un efecto regularizador demasiado alto. Se puede comprobar este efecto con distintas λ para un sólo atributo años del coche:


Polinomio grado 10 lambda 1000
Datos Entrenamiento
Datos validacion
Prediccion
Precio Coches (Euros)
Años


Polinomio grado 10 lambda 1.000000e-07
Datos Entrenamiento
Datos validacion
Prediccion
Precio Coches (Euros)
Años


Polinomio grado 10 lambda 1.000000e-10
Datos Entrenamiento
Datos validacion
Prediccion
Precio Coches (Euros)
Años

Para valores demasiado altos de λ se produce un sub-ajuste muy grande, el punto óptimo esta en $\lambda=10^{-7}$, donde la línea de predicción se ajusta a los datos sin que haya sobre-ajuste como ocurre para valores demasiado bajos de λ.

Una vez entrenado el modelo con todos los datos y el mejor λ encontrado se puede evaluar dando lugar a los siguientes resultados.

	RMSE (€)	**MAE (€)**
Datos entrenamiento	1072	801
Datos test	1071	755

Los errores son incluso más pequeños que para los del modelo con grados de polinomio óptimos sin regularización.

Conclusiones finales

A lo largo de la práctica se ha utilizado el algoritmo k-fold cross-validation para seleccionar los hiperparámetros de un modelo de regresión multivariable. Ha sido importante la expansión polinómica de los atributos y su normalización.

Se ha visto que seleccionar los grados óptimos para los atributos influye mucho en los errores a la hora de evaluar los modelos. La regularización también ha demostrado ser muy efectiva, con un λ adecuado ha permitido obtener resultados muy buenos para un modelo con grados de los atributos a priori demasiado altos.

Una búsqueda heurística de los grados del polinomio con una buena heurística puede obtener igualmente los grados óptimos con un coste mucho menor que el de la búsqueda exhaustiva.

Es interesante comprobar la mejora en los resultados si se utiliza los grados de los polinomios encontrados en los dos primeros apartados y se aplica regularización.

Con grados (5,6,6) hay tan poco margen de mejora que en la búsqueda del mejor λ se obtiene 10^{-13} como resultado, que al ser tan pequeño apenas penaliza los valores grandes de los valores altos de θ, dando lugar incluso a peores resultados al evaluarlo. A continuación se muestra una recopilación de la evaluación de los distintos modelos probados:

		RMSE (€)	**MAE (€)**
Datos entrenamiento	Búsqueda heurística	1076	804
	Búsqueda exhaustiva	1076	804
	Regularización	1072	801
	Búsqueda heurística y regularización	1197	899
Datos test	Búsqueda heurística	1009	763
	Búsqueda exhaustiva	1009	763
	Regularización	1071	755
	Búsqueda heurística y regularización	1167	864