

Práctica 2. Regularización y Selección de Modelos

Objetivo

El objetivo es aplicar y comparar técnicas de regularización y validación cruzada para selección de modelos en una aplicación de regresión lineal en un caso real. Se trata encontrar el mejor modelo polinómico para predecir el precio de un coche de segunda mano en función de tres atributos: antigüedad, kilómetros y potencia.

Estudio previo (es necesario prepararlo **por escrito, antes** de acudir a la práctica)

1. Repasa las transparencias de clase y estudia las funciones auxiliares proporcionadas para esta práctica. Escribe el algoritmo de k-fold cross-validation para elegir el grado de una regresión polinómica o el valor del parámetro de regularización. Ten en cuenta que para resolver la regresión es necesario realizar la expansión polinómica de los atributos y su normalización.

Desarrollo de la práctica

Copia a tu directorio de trabajo los ficheros proporcionados, y comprueba que funcionan correctamente en Matlab. A continuación escribe los programas necesarios para resolver la regresión polinómica con los datos de precios de coches de segunda mano, siguiendo los siguientes pasos:

2. **Selección del grado del polinomio para la antigüedad del coche.** Programa el algoritmo de k-fold cross-validation para elegir el grado del polinomio de antigüedad (entre 1 y 10), dejando fijos los kilómetros y la potencia con grado 1. Dibuja las curvas de evolución de los errores RMSE de entrenamiento y de validación.
3. **Selección del grado del polinomio para los kilómetros.** Dejando fijo el mejor grado encontrado para la antigüedad y grado 1 para la potencia, elige mediante k-fold cross-validation el grado del polinomio de los kilómetros (entre 1 y 10). Dibuja las curvas de evolución de los errores RMSE de entrenamiento y de validación. Con el mejor modelo encontrado, entrena con todos los datos y calcula el error RMSE con los datos de test.
4. **Regularización.** Realiza el ajuste de un polinomio de grado 10 para la antigüedad y 5 para los kilómetros y la potencia, utilizando regularización. Elige el parámetro λ mediante k-fold cross-validation. Dibuja las curvas de evolución de los errores RMSE de entrenamiento y de validación. Con el mejor modelo encontrado, entrena con todos los datos y calcula el error RMSE con los datos de test. Compara con los resultados obtenidos en el apartado 3. ¿Cuál de los dos modelos es mejor?

A entregar (en Moodle, dentro de un fichero .zip)

- Programa P2.m, junto con las funciones auxiliares que hayas programado, que vaya mostrando por pantalla los resultados de todos los apartados.
- Si no presentas la práctica durante la sesión, además deberás entregar la memoria de la práctica en un fichero P2.pdf ó P2.doc con los resultados de todos los apartados, su interpretación y las conclusiones que hayas obtenido.