PRÁCTICA 2 – REGULARIZACIÓN Y SELECCIÓN DE MODELOS

Fernando Aliaga Ramón - 610610

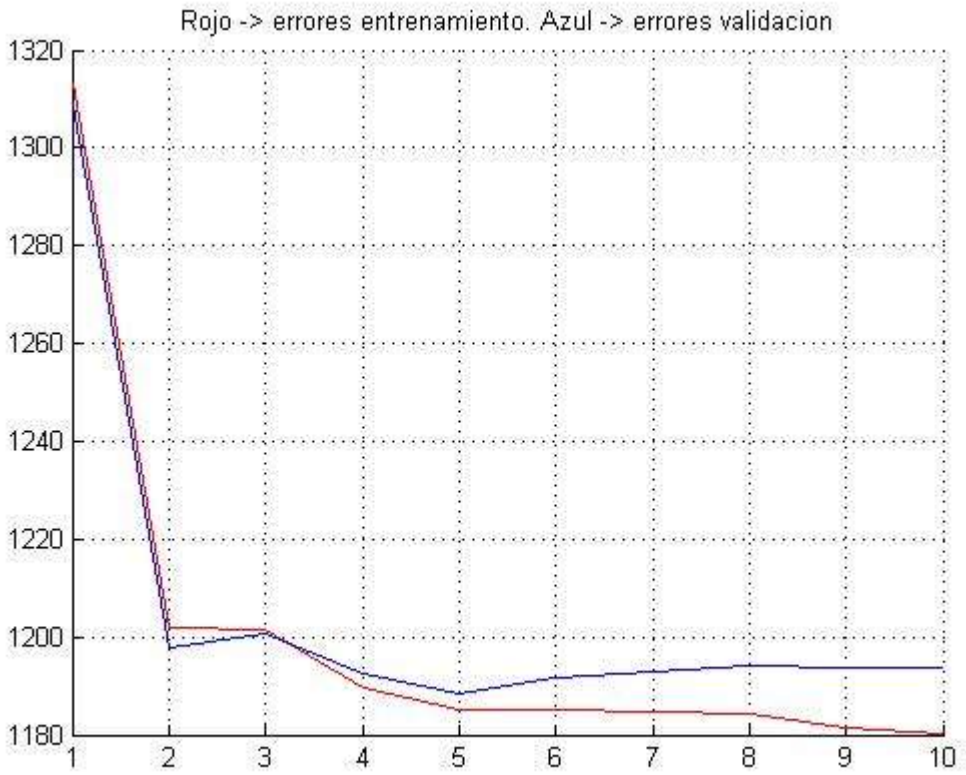
2. Seleccion del grado del polinomio para la antigüedad del coche

Para seleccionar el grado del polinomio que mejor funcione a la hora de reconocer nuevos datos nos vamos a ayudar del algoritmo k-fold. En él, iteraremos con los diferentes grados, entrenando el sistema en cada paso con parte de los datos y validando con el resto de los datos que no han sido utilizados para entrenar. Almacenaremos el mejor grado encontrado hasta el momento, siendo este reemplazado en caso de encontrar un nuevo mejor valor. La medida del error se realiza con el erro RMSE. A continuación se describe el algoritmo utilizado:

```
for grado = 1:n
   error_T = 0;
   error_V = 0;
   [Xexp] = expandir (Xdatos, [grado 1 1]);
   [ Xn, mu, sig ] = normalizar( Xexp );
   for i = 1:k
      [ Xcv, ycv, Xtr, ytr ] = particion( i, k, Xn, ydatos );
      h = Xtr\ytr;
      error_T = error_T + RMSE (h, Xtr, ytr);
      error_V = error_V + RMSE (h, Xcv, ycv);
   end
   error_T = error_T / k;
   error_V = error_V / k;
   errores_T = [errores_T error_T];
   errores_V = [errores_V error_V];
   if (error_V < mejor_error )
      fprintf('Actualizo mejor error y tamaño\n');
      fprintf('Nuevo mejor tamaño: %d\n',grado);
      fprintf('Nuevo mejor error: %d\n',error_V);
      mejor_tam = grado;
      mejor_error = error_V;
   end
end
```

Ejecutaremos en esta primera parte de la práctica el algoritmo para encontrar el mejor grado del polinomio para el atributo de antigüedad, dejando el grado para los otros dos atributos fijo en 1.

Las curvas de los errores durante la realización del algoritmo k-fold son las siguientes:



Podemos ver que el menor valor del error de validación se encuentra en el polinomio con grado 5 en el atributo antigüedad.

Volvemos a entrenar con todos los datos y calculamos el error con los datos de test no utilizados en el algoritmo k-fold.

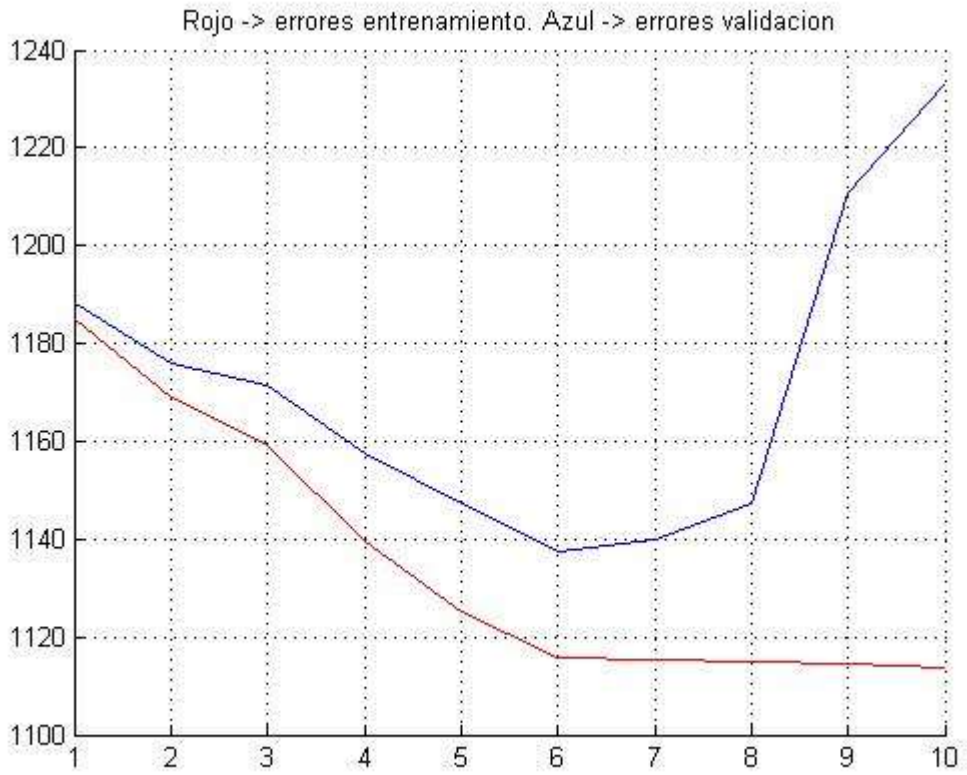
Mejor grado del polinomio = [5,1,1] Error RMSE con datos de test = 1.167842e+003

El erro RMSE es la raíz cuadrada del error cuadrático medio.

3. Seleccion del grado del polinomio para los kilómetros

Para encontrar el mejor grado del polinomio para los kilómetros volvemos a ejecutar el algoritmo k-fold, dejando fijos ahora el grado del atributo antigüedad (5) y el de la potencia (1).

Las curvas de los errores durante la realización del algoritmo k-fold son las siguientes:



Podemos ver que el menor valor del error de validación se encuentra en el polinomio con grado 6 en el atributo de los kilómetros.

Volvemos a entrenar con todos los datos y calculamos el error con los datos de test no utilizados en el algoritmo k-fold.

Mejor grado del polinomio = [5,6,1] Error RMSE con datos de test = 1.073405e+003

Podemos ver que reduce el error RMSE encontrado para el polinomio [5,1,1] (1.167842e+003).

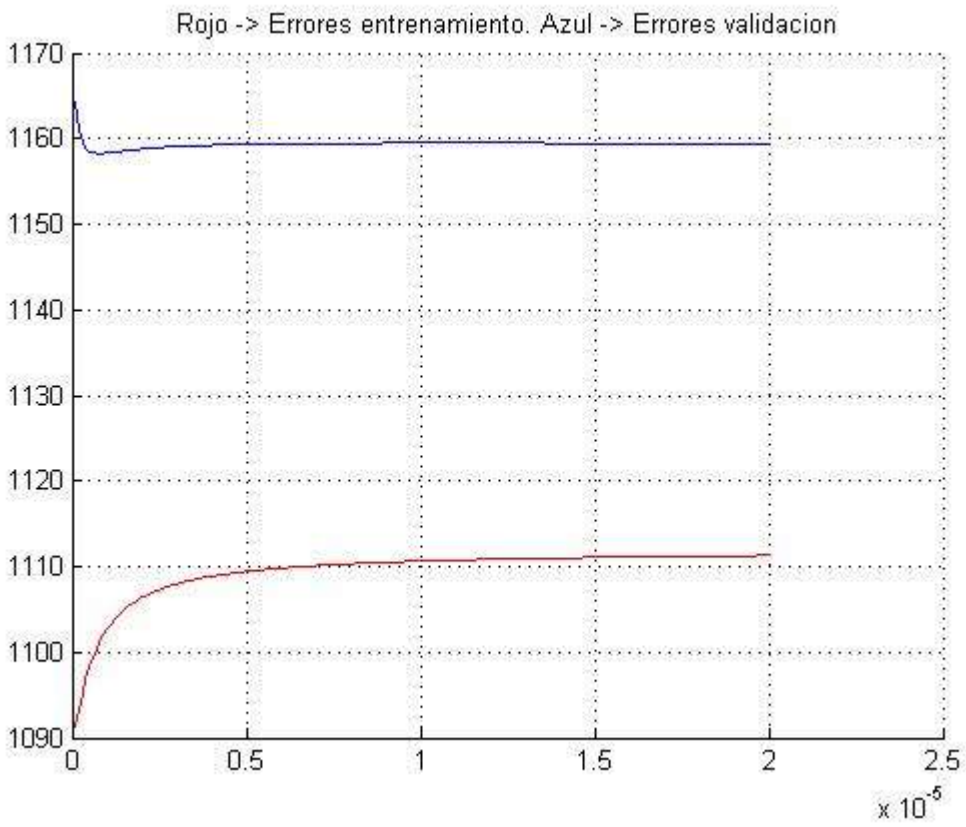
4. Regularizacion

Para usar una regresión lineal regularizada debemos escoger un valor de lambda. Dicho valor lo hemos escogido mediante validación cruzada.

La implementación de dicha validación cruzada tiene como claves los siguientes aspectos:

- Los valores de lambda se han buscado en el intervalo de 0 a 0.00002 con pasos de 0.0000001.
- El valor de 'k' ha sido definido como 10.
- Para encontrar el mejor valor de lambda hemos tomado como referencia el porcentaje de errores respecto de los datos de validación. Dicho porcentaje se ha obtenido de la misma forma que en el apartado 2 y 3, haciendo uso de la predicción y comparando con los datos de salida.
- Los valores de los errores respecto de los datos de entrenamiento y de validación se almacenarán en dos vectores para ser mostrados en una gráfica comparativa.

Las curvas de los errores durante la realización del algoritmo k-fold son las siguientes:



Podemos ver que el menor valor del error de validación se encuentra con lambda=8.000000e-007.

Volvemos a entrenar con todos los datos y calculamos el error con los datos de test no utilizados en el algoritmo k-fold.

```
 Mejor lambda encontrado = 8.000000e-007
Error RMSE con datos de test = 1.061630e+003
```

Si comparamos todos los errores obtenidos con los datos de test podremos comprobar que sistema se comporta mejor ante nuevos datos:

```
Comparativa errores:
Error RMSE (2) = 1.167842e+003
Error RMSE (3) = 1.073405e+003
Error RMSE (4) = 1.061630e+003
```

Viendo los errores RMSE podemos afirmar que el modelo regularizado es el que mejor se comporta, ya que reduce en mayor medida el error obtenido con los datos de test.