Laboratorio 1
Máquinas de vector de soporte y redes de neuronas

Autores: Eloy Alfageme, Inés Heras, Jorge García, Manuel Pasieka

Introducción

El juego de datos proporcionado corresponde a la biblioteca StatLib mantenida por la Universidad Carnegie Mellon, y son de fecha 1993. Son datos acerca del consumo en millas por galón de diferentes características de automóviles. Es una versión ligeramente modificada de los datos reales iniciales. El objetivo del estudio es predecir el consumo en función de características como la cilindrada, los caballos o el año del modelo, entre otros. Son 8 categorías y tenemos una muestra de 398 datos de cada una. Es un dataset pequeño.

Metodología utilizada

Los datos presentan algunas pérdidas (missing) en la potencia en caballos. Hemos asignado NaN a los valores perdidos. Al ser muy pocos (6), no consideramos relevante un tratamiento numérico. El conjunto de datos se ha dividido posteriormente en un grupo de entrenamiento (75%) y un grupo de test (25%). El grupo de test a su vez se ha dividido en test y validación al 50%. Los datos incluyen una variable categórica sobre el nombre del vehículo, que ha sido codificada a etiqueta numérica mediante LabelEnconder.

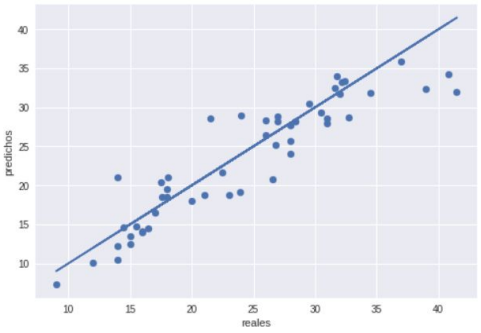
Máquina de Vector de Soporte (SVM)

Para la SVM hemos puesto a competir los siguientes hiper parámetros:

- kernels: rfb y **lineal**
- Softmargin: 0.001, 0.01, 0.1, **1**, 10
- gamma: **0.001**,0.01,0.1,1

La configuración ganadora ha sido con el kernel lineal, el softmargin de 1 y el gamma de 0.001, con un coeficiente de determinación de 83%.

MSE: 10.63; MAE: 2.51; RMSE: 3.26



Mediante un diagrama de dispersión de valores reales contra valores estimados, apreciamos una pérdida de precisión en la predicción a partir de los 38 mpg, siendo los valores subestimados por el modelo.

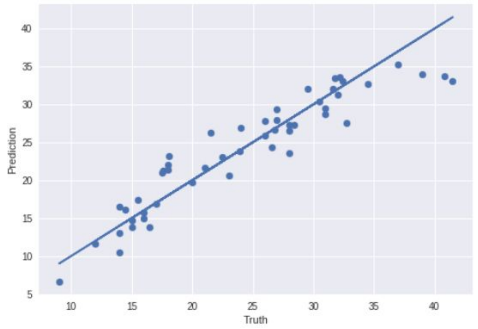
Red Neuronal

Hemos tratado los datos con el StandardScaler de sklearn con el objetivo de normalizarlos y que la red no sufra un sesgo inapropiado debido a las diferentes escalas. Al igual que con la

svm, hemos generado una competición de hiperparámetros para buscar la configuración óptima de la red. En este caso los valores probados han sido:

- Capas ocultas: (10,), **(50,)**, (100,), (10, 10,), (50, 50,)
- Función de activación: relu, **lineal**
- Alpha: **0.0001,** 0.01, 0.10
- Ratio de aprendizaje: [**0.1**, 0.01, 0.001]

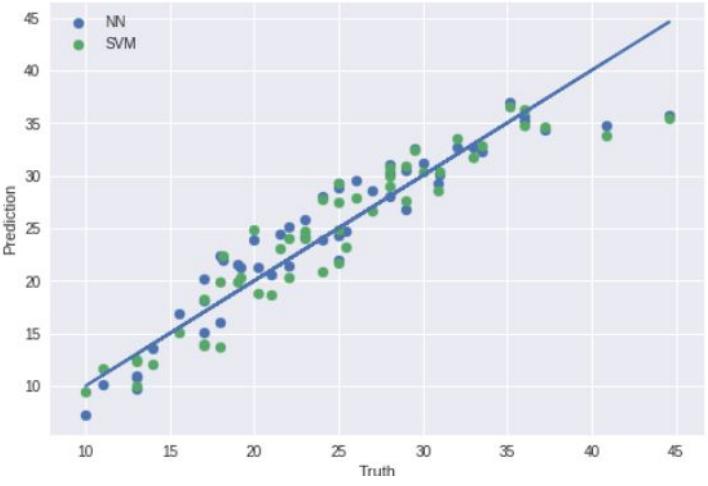
Hemos generado 90 configuraciones diferentes. La mejor configuración ha sido la resaltada en negrita, con un coeficiente de determinación del 87% y MSE: 8.07



Al igual que se puede apreciar con las svm, el modelo pierde precisión para valores altos de mpg, subestimando las predicciones. Sin embargo es más preciso en el cuerpo central de los datos, para valores entre 20 y 30 mpg.

Comparativa

Hemos usado el juego de datos de test para comprar los dos modelos. Ninguno de los dos modelos ha probado este juego de datos. Los de validación fueron usados para la búsqueda de los hiper parámetros.



Ambos modelos empatan en el coeficiente de determinación 87%. La máquina de vector de soporte obtiene un MSE ligeramente más bajo que la red neuronal, 7.41 contra 7.86, pero es despreciable.

Conclusión

Para el tipo de problema dado, una regresión lineal con un dataset reducido de datos, ambos modelos se desempeñan bien, cometiendo los mismos errores de subestimación en valores altos de mpg, seguramente producidos por las características de los datos, ya que hay muy poca información para valores tan altos de consumo.