Práctica Tema 8: Ajustar una función (Matlab DLT)

Motivación

Para la realización de esta práctica se ha escogido el ***Ajuste de una función*** utilizando el *dataset* de ***bodyFat*** de Matlab. Existen varios motivos por los que se ha escogido esta combinación: teniendo en cuenta el tipo de problema a resolver y teniendo en cuenta el *dataset*.

- Durante la realización del Máster en Big Data por la Universidad de Compostela, en un par de asignaturas estuve realizando varias prácticas tanto de regresión como de aplicaciones reales de redes neuronales. El principal problema, es que la mayoría de temáticas que me habían tocado con redes neuronales se enfocaron a problemas de clasificación y/o clusterización. Por este motivo he escogido la aplicación de redes neuronales a un problema de regresión.
- En cuanto al *dataset*, de las diferentes opciones disponibles (la de *chemical* ya estaba llena, y era la que más me interesaba de entre todas), escogí la de *bodyFat*. El motivo principal que me llevó a esta decisión es que actualmente, aún con todos los avances en material sanitario y la concienciación de llevar a cabo una vida saludable, los porcentajes de obesidad infantil, problemas de colesterol y número de infartos causados por una mala alimentación[1] han aumentado bastante en los últimos años. Este caso puede un perfecto escenario para aplicar una red neuronal que nos permita estimar el porcentaje de grasa en diferentes individuos e intentar cambiar los hábitos alimenticios o tomar otras medidas.

En definitiva, la motivación fue una mezcla del tipo de problema que se quería encarar con redes neuronales así como el tipo de problema que permitía resolver el *dataset* escogido.

Descripción conjunto de datos y problema a resolver

El *dataset* escogido está formado por un conjunto elevado de medidas en relación a un conjunto de características anatómicas. Se dispone de dos conjuntos de datos:

- un conjunto de 252 medidas de diferentes individuos formado por valores para los diferentes **atributos anatómicos** de éstos (constituyen la entrada de la red neuronal), y
- un conjunto de resultados que muestran el **porcentaje de grasa en el cuerpo** (constituye la salida de la red neuronal) para cada uno de esos individuos (en total 253 valores de porcentaje).

Con los diferentes valores de las características de un individuo, diseñando una red neuronal para este problema, se es capaz de estimar/predecir el porcentaje de grasa en el cuerpo para diferentes entradas (definidas por los valores de los atributos anatómicos de un individuo). Por lo tanto, la red

1 No estoy afirmando que el porcentaje de grasa en cuerpo sea uno de los indicadores principales, pero resulta de bastante ayuda para intentar prevenir estos casos.

neuronal diseñada constará de 13 valores de entrada en la capa de entrada y un valor de salida (resultado deseado) en la capa de salida con un número variable de neuronas en la capa intermedia.

El conjunto de medidas está formado por 13 atributos físicos/anatómicos que todo ser humano posee:

- Edad (medida en años)
- Peso (medido en libras)
- Altura (medida en centímetros)
- Cuello (circunferencia medida en centímetros)
- Pecho (circunferencia medida en centímetros)
- Abdomen (circunferencia medida en centímetros)
- Cadera (circunferencia medida en centímetros)
- Muslo (circunferencia medida en centímetros)
- Rodilla (circunferencia medida en centímetros)
- Tobillo (circunferencia medida en centímetros)
- Bíceps extendido (circunferencia medida en centímetros)
- Antebrazo (circunferencia medida en centímetros)
- Muñeca (circunferencia medida en centímetros)

El problema de la estimación del porcentaje de grasa dados unos valores de entrada, es claramente un problema de regresión debido a que la salida esperada es un valor numérico con propio significado para el problema. En este caso no hay nada que pueda ser clasificado ni clusterizado.

Estudio sobre el ajuste de las neuronas

Se han probado 4 valores diferentes para el número de neuronas, intentando seleccionar valores dispares entre ellos para apreciar como afectaban los valores pequeños respecto a valores grandes[2].

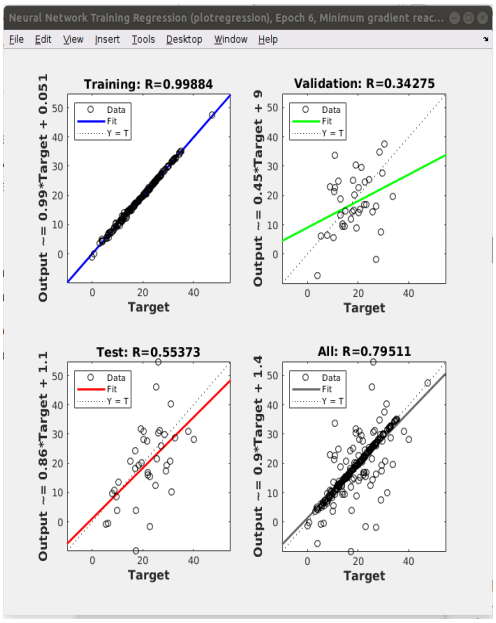

Imagen 1: Resultado regresión 500 neuronas

2 Se han realizado múltiples combinaciones posibles del número de neuronas y se ha observado que a partir de 75 neuronas, los resultados no aportaban ninguna información adicional (las mismas conclusiones se pueden obtener con 75 y 50 neuronas, *overfitting*).

Aparte de los valores mostrados a continuación, se han probado valores como 100 o 500 para apreciar los tiempos de entrenamiento (por encima de los 10 segundos en ambos casos) y como afectaba aumentar tanto el número de neuronas al *overfitting*. En la Imagen Error: Reference source not found se puede apreciar como para el conjunto de entrenamiento la red neuronal diseñada se ajusta perfectamente (error del 0.01), mientras que para el conjunto de validación y test, el modelo no es para nada el mejor (errores de 99 y 126 respectivamente).

En los casos mostrados a continuación, el tiempo de entrenar a la red neuronal no es un factor decisivo (en todos los casos tarda menos de 2 segundos). Factores determinantes serán los valores de errores obtenidos para los tres conjuntos y el número de neuronas (diseñar e implementar con un número pequeño de neuronas es mejor que tener que utilizar un valor superior).

75 neuronas

Utilizando 75 neuronas ya empezamos a notar un *overfitting* respecto al conjunto de entrenamiento (no muy lejos de los valores obtenidos con 500 neuronas) como se puede observar en las rectas de la Imagen 2.

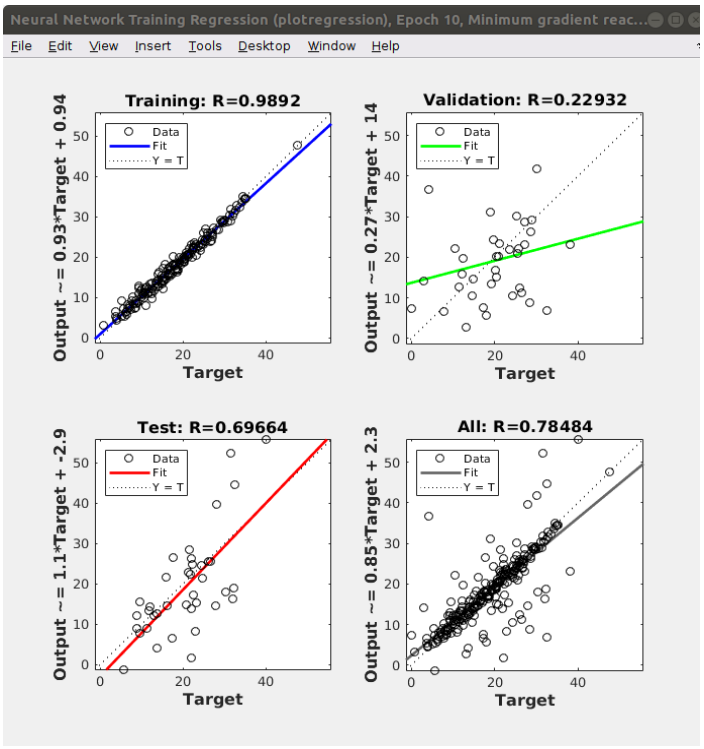

Imagen 2: Resultado regresión 75 neuronas

El error obtenido en el conjunto de entrenamiento es prácticamente 0, mientras que el error obtenido en el conjunto de test es de 80 (aprox.) y en el conjunto de validación de alrededor de 131. Estamos observando variaciones de errores de entre 80 y 130 (lo que es una burrada). Estos resultados confirman que la red neuronal está muy sobreajustada, y los resultados que se obtendrían en un entorno de producción no serían para nada buenos.

50 neuronas

Se ha decidido probar con 50 para ver si la evolución de los resultados mejoraba y se iba disminuyendo el sobreajuste sufrido con la configuración anterior. Como se puede apreciar en la gráfica de *performance* obtenida de Matlab, el grado de sobreajuste ha disminuido bastante.

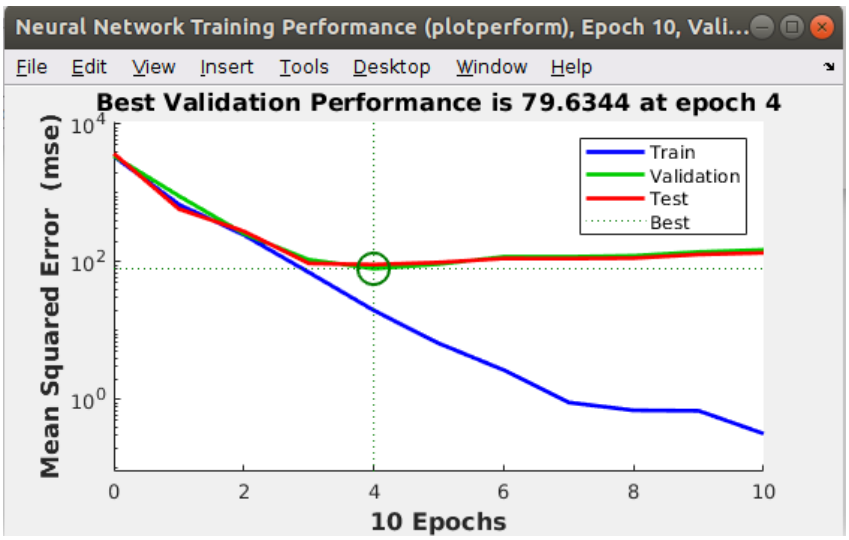

Imagen 3: Resultado performance 50 neuronas

El error obtenido en el conjunto de entrenamiento es de 5, y la diferencia respecto a los errores obtenidos en el conjunto de validación y test (prácticamente 50) es bastante menor que la observada en el apartado anterior, reduciéndose a prácticamente 45. Pero todavía apreciamos una red que se ajusta demasiado a los valores de entrenamiento.

15 neuronas

Los siguientes valores se han escogido teniendo en cuenta el valor que 10 neuronas escogido en el guión de prácticas. Se optó por probar 2 valores próximos a éste, a parte de que los resultados obtenidos con 5 neuronas eran bastante más prometedores que los que se obtuvieron con 10. La Imagen 4

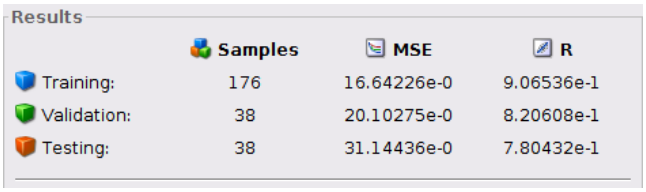

Imagen 4: Resultados MSE 15 neuronas

Como se puede apreciar, reduciendo el número de neuronas a 15 se aumenta el error en el conjunto de entrenamiento, pero se consigue que la diferente de errores entre conjuntos se reduzca debido a que la red neuronal obtenido no "*solo memoriza los valores de entrenamiento*".

En la Imagen 5 se puede apreciar la diferencia de la función obtenida para los diferentes conjuntos.

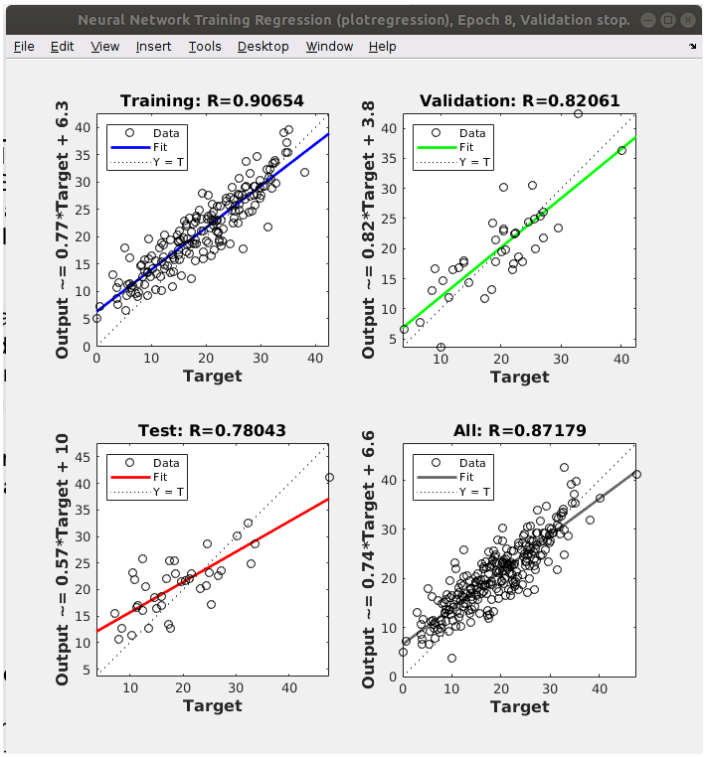

Imagen 5: Resultado regresión 75 neuronas

5 neuronas

Por último se decide probar con 5 neuronas. Realmente, al tratarse de un problema tan simple y bien definido como este, un número pequeño de neuronas debería de ser la mejor opción. Por regla general, problemas más simples necesitan pocas neuronas, mientras que problemas muy complejos necesitan un número elevado de neuronas.

Con 5 neuronas casi igualamos los valores de errores entre los diferentes conjuntos (una diferencia respecto al de entrenamiento de aproximadamente 5 puntos), obteniendo un error de 20 para el conjunto de entrenamiento. Los análisis y reflexiones propias acerca de este valor se comentarán en la sección Conclusiones.

A continuación se muestra el rendimiento obtenido por la red neuronal con 5 neuronas:

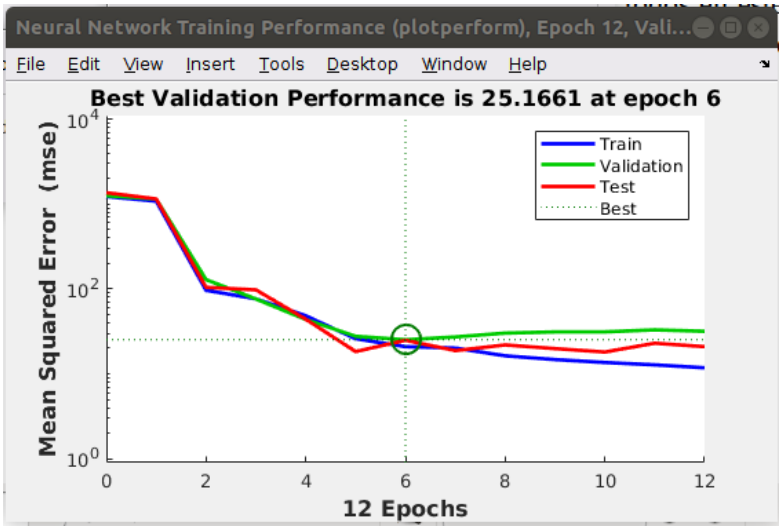

Imagen 6: Resultado performance 5 neuronas

En la carpeta BodyFat se pueden encontrar para cada número de neuronas diferentes capturas de pantalla de los resultados (no se mostraron todos en este documento) así como los scripts de Matlab necesarios par poder explotar la red y reproducir las pruebas realizadas.

Selección del mejor resultado

Observando los resultados presentados en la sección Estudio sobre el ajuste de las neuronas, parece bastante seguro que la mejor solución es la red neuronal con 5 neuronas ocultas debido que que no produce un *overfitting* de los datos de entrenamiento y los valores de errores obtenidos para los 3 conjuntos son bastante parecidos; lo que induce a pensar que se puede comportar de correcta en un entorno de producción. La segunda mejor opción sería una red neuronal con 15 neuronas, pero creo que sigue apreciándose un cierto sobreajuste de los datos.

Tanto las redes neuronales de 50 y 75 neuronas, producen un *overfitting* muy marcado, lo que producirá unos resultados bastante malos en producción. Por lo tanto la opción escogida basándonos en el rendimiento obtenido y los valores de MSE es la **red neuronal de 5 neuronas**.

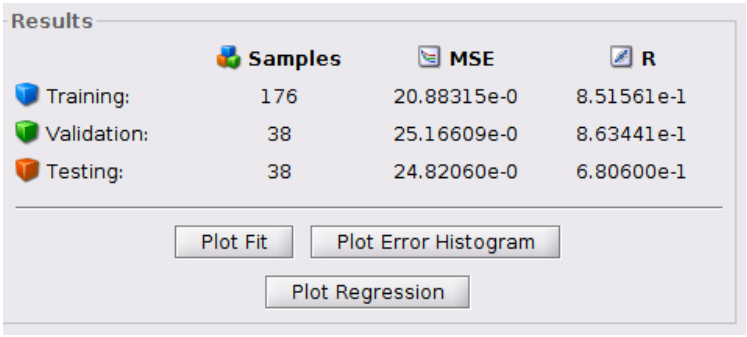

Imagen 7: Resultados MSE 5 neuronas

Explotación de la red

Para realizar las pruebas de funcionamiento de la red primero cargamos los datos de *dataset* para realizar 2 pruebas contra los datos utilizados (y compararlos con las salidas esperadas).

A continuación se muestran las salidas para las 2 pruebas contra valores del *dataset*:

```
>> load bodyfat_dataset.mat
>>
>> bodyFat5(bodyfatInputs(:, 1))

ans =

   14.9930

>> bodyfatTargets(:, 1)

ans =

   12.3000

>>
```

```
>> bodyFat5(bodyfatInputs(:, 123))

ans =

   13.8014

>> bodyfatTargets(:, 123)

ans =

   14.7000

>>
```

Como se puede observar, el resultado se queda un poco más lejos, desde mi punto de vista, del valor de porcentaje de grasa que debería de ser. A continuación se muestra una tabla con los valores seleccionados:

	Individuo 1	**Individuo 2**	**Individuo 3 (yo)**
Edad	23	23	26
Peso	154 (69.8 kg)	198 (90 kg)	136 (62 kg)
Altura	67 (170 cm)	65 (165 cm)	67 (170 cm)
Cuello	34	45	34
Pecho	89	100	92
Abdomen	81	100	75
Cadera	91	100	80
Muslo	55	65	50
Rodilla	33	40	34
Tobillo	17	25	20
Bíceps	28	40	27
Antebrazo	23	35	25
Muñeca	13	20	15
% grasa	***9.03***	***26.78***	10.30

Se muestran a continuación las salidas:

```
>> individuo1 = [23; 154; 67; 34; 89; 81; 91; 55; 33; 17; 28; 23; 13];
>> bodyFat5(individuo1)

ans =

    9.0395

>>
```

```
>> individuo2 = [23; 198; 70; 45; 100; 100; 100; 65; 40; 25; 40; 35; 20];
>> bodyFat5(individuo2)

ans =

   24.6405

>>
```

Como se puede apreciar, para el individuo 1 (se baso en el primer ejemplo ejecutado con el conjunto de datos *bodyfatInputs* reduciendo los valores), el índice de grasa corporal es menor que el del individuo 2. Lo que es normal considerando la diferencia en los valores establecidos:

- se puede considerar al individuo 1 como una persona de estatura normal (mas o menos en la media española de la altura media de un varón) y muy delgada (lo que resulta de en poco porcentaje grasa),
- el individuo 2 es una persona un tirando a bajita (considerando que ambos son hombres) pero con unos valores que indican que está por encima del peso normal (quizá tirando a sobrepeso).

En el caso del individuo 1, se puede considerar el porcentaje bastante acertado; mientras que en el caso del individuo 2 se esperaba un porcentaje un poco superior (quizás alrededor del 29-30 % de grasa), aunque según la *American Council on Excercise*, este individuo sería obeso[3]; pero el valor entra dentro de los márgenes de error obtenidos. Por último, en relación a los datos obtenidos relacionados a mi; me marca de forma aproximada un 10.3 % de grasa corporal, cuando en el último reconocimiento de la empresa me marcaba en torno al 12-14%.

Conclusiones

Durante la ejecución de las pruebas para seleccionar los diferentes números de neuronas, me fijé que los valores de error obtenidos fuera cual fuera el número de neuronas eran bastante superiores a los obtenidos durante la ejecución de la sesión práctica. Así mismo, los resultados obtenidos en la sección Explotación de la red parecen bastante mejorables.

Este hecho me ha llevado a pensar lo siguiente:

- quizá haya alguna variable que no aporta información y se contraproducente mantenerla o
- el tamaño del *dataset* sea muy pequeño (no llega ni a 500 medidas).

3 Cuando estaba preparando los datos del Individuo 2, pensaba en un hombre ancho en el límite del sobrepeso.