2. Primero describa como esta construida la base de datos utilizada en el problema de regresión; para esto puede graficar los datos, o analizar la función que se usa para generarlos. Explique cuál es la función polinomial con la cual están construidos los datos en matlab. Diga cuantas muestras de entrenamiento y de validación se usan para resolver el problema, también escriba cuantas caracterısticas tienen las muestras.

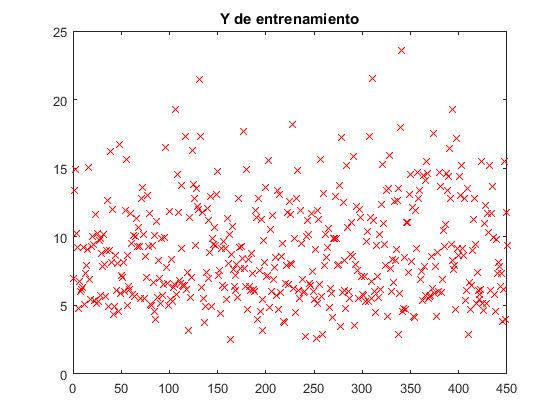
**Descripción de la construcción de la base de datos**

Para la construcción de los datos se crean dos vectores fila con la función linspace de Matlab. El primero se hace con un límite inferior de -20, un límite superior de 20 y 500 elementos. El segundo de manera similar al primero pero con límite inferior de -50 y límite superior de 50. Estos dos vectores fila se transponen y se agregan a una matriz de 500 x 2, esta matriz es llamada X. Luego se normaliza la distribución de datos de la matriz X con la función zscore de Matlab, resultando una distribución de datos de media 0 y desviación 1. Esta matriz normalizada es usada para la construcción del vector de las salidas de la función polinomial inicial Y=2\*X.^3 + 3\*X.^2 - 4\*X + 6 , luego se hace otro proceso de randomización usando los valores mínimo y máximo de la función Y. Al finalizar la generación de la función polinomial, se revierte la normalización en X, para extender la matriz con la función potenciaPolinomio que depende del grado del polinomio. Se particionan los vectores en dos, uno para el entrenamiento y otro para las pruebas, se utiliza la función randperm para permutar aleatoriamente los índices del 1 al 500, luego, estos índices randomizados son usados para distribuir los datos en las particiones creadas anteriormente. Por último se agrega a las particiones Xtrain y Xtest una columna de 1s al final, esto es necesario para hacer el respectivo ajuste con los pesos W y evitar procesamiento posterior. Así es como se genera la información en la base de datos. Veremos algunas gráficas que ilustran los resultados.

**Función polinomial**

La función polinomial con la cual están generados los datos es

Y=2\*X.^3 + 3\*X.^2 - 4\*X + 6;



**Muestras de entrenamiento y validación**

Se usaron 450 muestras para entrenamiento y 50 muestras para la validación.

**Caracteristicas de la muestra**

Inicialmente hay dos variables: X1 y X2. En base al grado ingresado se generan otras variables en función de las dos primeras.

|  |  |  |
| --- | --- | --- |
| Taza de Aprendizaje | Grado del Polinomio | Error Cuadratico Medio (ECM) |
| 0,1 | 1 | NaN |
| 2 | NaN |
| 3 | NaN |
| 4 | NaN |
| 5 | NaN |
| 0,001 | 1 | 80.3105 |
| 2 | NaN |
| 3 | NaN |
| 4 | NaN |
| 5 | NaN |
| 0,00001 | 1 | 108.2567 |
| 2 | Inf |
| 3 | NaN |
| 4 | NaN |
| 5 | NaN |

2-a. **Usando la siguiente fórmula**: W = W - eta\*(1/N)\*(((X\*W)-Y)'\*X)';

Mientras más pequeña se pone la taza de aprendizaje y el grado del polinomio, el ECM va disminuyendo considerablemente, pero si aumenta tanto el grado del polinomio como la tasa de aprendizaje el ECM aumenta tanto que en ocasiones puede llegar a tomar valores NaN donde se supone que es un número tan grande que matlab no lo puede poner

2-b. Se obtienen 9 valores para W, por que son 8 que acompañan a las variables con exponentes desde 1 hasta 4, y un término independiente

2-c. La respuesta es la A. X1X2X1^2X2^2X1^3X2^3X1^4X2^4

3.

|  |  |  |
| --- | --- | --- |
| Taza de Aprendizaje | Grado del Polinomio | Error Cuadratico Medio (ECM) |
| 0,1 | 1 | 19.2542 |
| 2 | 3.8366 |
| 3 | 6.8282 |
| 4 | 6.6197 |
| 5 | 7.7729 |
| 0,001 | 1 | 79.981 |
| 2 | 70.722 |
| 3 | 72.2338 |
| 4 | 65.9872 |
| 5 | 68.0568 |
| 0,00001 | 1 | 108.2935 |
| 2 | 108.1641 |
| 3 | 108.1892 |
| 4 | 108.0443 |
| 5 | 108.0933 |

3-a. ¿Qu´e proceso hace la normalizaci´on sobre los datos? Consulte porqu´e

es necesaria la normalizaci´on en el modelo de regresi´on log´ıstica y

cu´ales son los tipos de normalizaci´on m´as comunes. ¿Cu´al de ellos se

aplic´o en el laboratorio?

El proceso de normalización permite tener una distribución de datos con media 0 y varianza 1, esto facilita las tareas de análisis y comparación. En este caso utilizamos las medias y las desviaciones estándar usadas en Xtrain para centrar la distribución de datos, y utilzamos estos valores para normalizar las muestras de validación

La normalización debe realizarse ya que los valores fueron obtenidos aleatoriamente y puede que no exista ninguna relación entre las variables, esto afectaría los procesos de regresión.

La normalización utilizada es la estandar, a traves de la funcion de matlab zscore.

Zscore = (x-u)/

Max - min: convierte las variables en rangos entre 0 y 1.

Al descomentar las líneas de normalización disminuye aún más el ECM

3-b. Compare el resultado con el punto anterior donde no se us´o la normalizaci

´on, ¿Cu´al es el grado que mejor representa el conjunto de

muestras y por qu´e?

En el proceso de regresion con grado 4 y eta 0,001 se logra un ECM de 80, cuando las muestras no estan normalizadas. Esto es una buena representación del conjunto de muestras.

Cuando se uso la normalización, se encontro que con el grado 2 y eta 0.1 se encuentra un ECM de 3.8, siendo la mejor representación del comportamiento de las muestras.

c. función polinomial generada????

4 BD del problema de clasificación

Se usa la misma fórmula que en regresión múltiple?(En los W)

B dónde sale la función polinomial de la mejor frontera encontrada

C gráfica?

D cassify

5-a

Si bien los datos son transformados por la normalización, está no es tan evidente como en la regresión multiple. Esto es debido a que los conjuntos X1 y X2 de la regresión logistica fueron construidos a partir de funciones de aleatoriedad similares y no hay mucha diferencia en sus medias.

5-b