PRÁCTICA REGRESIÓN

Daniel García
garciad@ifca.unican.es

Datos

- CSV con serie temporal de datos físico - químicos medidos en el embalse de Cuerda del Pozo (Soria) durante los años 2014 y 2015

	A	B	C	D	E	F
1	date	AVG(Temp)	AVG(Press)	AVG(Cond)	AVG(Salinity)	AVG(DO)
2	01/01/2014 0:03	3.997746302	9.621590016	0.040665008	0.030426135	13.64697077
3	02/01/2014 0:05	4.057666958	3.944665552	0.039898844	0.029823479	13.60542035
4	03/01/2014 0:05	4.13361175	3.991832818	0.03766908	0.028097318	13.58907035
5	04/01/2014 0:14	4.188685341	3.945494965	0.0345822	0.025744588	13.57723625
6	05/01/2014 0:08	4.346171706	9.391423477	0.03129767	0.023243731	13.50922589
7	06/01/2014 0:05	4.392661279	9.359303401	0.03112603	0.023105117	13.42513025
8	07/01/2014 0:05	4.362913301	9.504405446	0.032435575	0.024089534	13.36802177
9	08/01/2014 0:04	4.331744297	9.513480077	0.035868421	0.026672862	13.38517309
10	09/01/2014 0:05	4.262987126	3.962276253	0.035115589	0.026128516	13.32167398
11	10/01/2014 0:05	4.309596222	3.938766667	0.033933667	0.025224667	13.30987211
12	11/01/2014 0:05	NAN	NAN	NAN	NAN	NAN
13	12/01/2014 0:05	NAN	NAN	NAN	NAN	NAN
14	13/01/2014 0:05	NAN	NAN	NAN	NAN	NAN
15	14/01/2014 0:05	NAN	NAN	NAN	NAN	NAN
16	15/01/2014 0:05	NAN	NAN	NAN	NAN	NAN
17	16/01/2014 14:00	4.517913	3.363136583	0.031323917	0.023233167	13.25917875
18	17/01/2014 1:00	4.465651	4.118094783	0.032199348	0.023893478	13.18188126

Cargar el dataset

Seleccionar la conductividad y la salinidad de 2014.

Gráfica Salinidad 2014 vs Conductividad 2014

- Consejos:

  - Formatear la fecha para poder seleccionar por fecha
  - Omitir valores nan

Regresión lineal

1. Modelo de Regresión lineal con un “feature” para los datos Salinidad 2014 vs Conductividad 2014

$$y = f(x) = \alpha_0 + \alpha_1 x$$

1.1 - Utilizando como función de coste el cuadrado de la distancia euclídea. Calcular de forma analitica, los coeficientes que minimizan la función de coste.

$$Loss = (y - X\,\alpha)^T (y - X\,\alpha) = \sum_{j=1}^{N} (y_j - \alpha_0 - \alpha_1 x_j)^2$$

$$\nabla_\alpha Loss = 0 \Rightarrow \begin{cases} \dfrac{\partial Loss}{\partial \alpha_0} = -2 \sum_{j=0}^{N} (y_j - \alpha_0 - \alpha_1 x_j) = 0 \\ \dfrac{\partial Loss}{\partial \alpha_1} = -2 \sum_{j=0}^{N} (y_j - \alpha_0 - \alpha_1 x_j)\, x_j = 0 \end{cases}$$

$$\alpha_0 = \bar{y} - \alpha_1 \bar{x}$$

$$\alpha_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$

Regresión lineal

1.2 Caso general. Para los datos Salinidad 2014 vs Conductividad 2014, calcular los coeficientes usando matrices.

Construye una función que reciba los vectores “x” e “y” y que calcule los valores de los coeficientes que minimizan la función de coste.

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(N)} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_M^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_M^{(2)} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_1^{(N)} & x_2^{(N)} & \dots & x_M^{(N)} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_M \end{bmatrix}$$

$$\nabla_\alpha Loss = -X^T(y - X\alpha) = -X^T y + X^T X\alpha = 0$$

$$\alpha = (X^T X)^{-1} X^T y$$

1.3 Compara ambos resultados con el obtenido mediante la función “lm” implementada en R.

Regresión no lineal

2. Modelo no lineal para los datos Salinidad 2014 vs Conductividad 2014

Construye una función que reciba los vectores “x” e “y” y calcule los valores de los coeficientes que minimizan la función de coste.

2.1 Polinomio de grado 2
2.2 Polinomio de grado 3.

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(N)} \end{bmatrix} = \begin{bmatrix} 1 & x^{(1)} & x^{(1)2} & \dots & x^{(1)M} \\ 1 & x^{(2)} & x^{(2)2} & \dots & x^{(2)M} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x^{(N)} & x^{(N)2} & \dots & x^{(N)M} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_M \end{bmatrix}$$

$$y = f(x) = \alpha_0 + \alpha_1 x + \dots + \alpha_M x^M \qquad \alpha = (X^T X)^{-1} X^T y$$

2.3 Compara los coeficientes calculados, con los obtenidos con la función “glm” implementada en R. “glm(y~poly(x,grado del polinomio))”

Regresión

1. Comprueba los modelos calculando la función de coste con los datos de Salinidad vs Conductividad de 2015
   a. Usamos como función de coste el cuadrado de la distancia euclídea.

$$Loss = (y - X\alpha)^T (y - X\alpha)$$