Introducción al reconocimiento de patrones: Regresión paramétrica

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación, bachillerato en Ingeniería en Computación,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

17 de mayo de 2022

El presente trabajo hace un repaso del problema de ajuste de curvas, con el objetivo de introducir al estudiante gradualmente al problema de la clasificación de muestras (para el presente curso correspondiente a valores extraídos de imágenes digitales) y el diseño de clasificadores. El problema de clasificación se puede enfocar como una particularización discreta del ajuste de curvas, como se apreciará posteriormente.

1. El problema del ajuste de curvas

El problema de ajuste de curvas se refiere a la construcción de un modelo que permita predecir el comportamiento de un fenómeno a estudiar. Este fenómeno puede caracterizarse mediante una serie de valores a su entrada, con lo cual es posible medir su comportamiento a la «salida» del fenómeno.

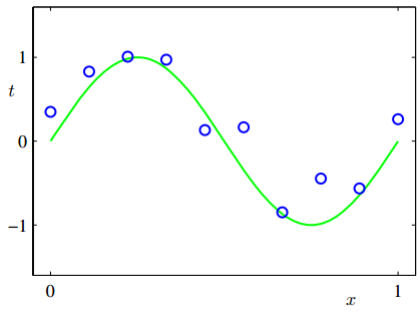

Figura 1: El problema del ajuste de curvas, los puntos azules corresponden a los datos de entrenamiento. Tomado de [1].

Supóngase entonces que x constituye la serie continua de valores a la entrada del fenómeno, con lo cual y' definida por,

$$y' = \sin(2\pi x), \tag{1}$$

es continua.

La función y' corresponde a la función que caracteriza a los datos del fenómeno por estudiar (sin ruido).

Si se tiene un arreglo finito de N observaciones $\vec{x} = [x_1, ..., .x_N]^T$ de la entrada al fenómeno en estudio, el conjunto de observaciones de los valores muestreados de salida del funcional bajo estudio viene dado por: $\vec{t} = [t_1, ..., .t_N]^T$. Nótese que utilizaremos la notación $\vec{x}, \vec{t} \in \mathbb{R}^N$ para denotar un conjunto de elementos ordenados, y no usaremos la notación de vector, puesto que los conceptos de dirección y magnitud no son necesarios para tales datos. Las salidas muestreadas en este caso presentan **ruido Gaussiano artificial** $\epsilon(x)$, con $\mu = 0$ y una desviación estándar σ (y precisión $\beta = 1/\sigma$) sobre y':

$$t = \sin(2\pi x) + \epsilon(x). \tag{2}$$

El objetivo del ajuste de curvas es encontrar un modelo que se ajuste al conjunto de observaciones $\mathcal{D} = \left\{\vec{x}, \vec{t}\right\}$, y que posibilite la predicción de de la salida t para una nueva muestra x.

En este trabajo práctico estudiaremos un **modelo con pesos lineales respecto al vector de pesos o parámetros del modelo** $\vec{w}$ con funciones base polinomiales, de la forma:

$$y(x, \vec{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j. \tag{3}$$

por lo que $\vec{w} \in \mathbb{R}^{M+1}$, definiendo $M + 1$ la dimensionalidad del modelo, y constituyendo la cantidad de parámetros desconocidos a estimar. Observe que la regresión lineal es un caso particular de la regresión polinomial con $M = 1$:

$$y(x, \vec{w}) = w_0 + w_1 x = \sum_{j=0}^{1} w_j x^j. \tag{4}$$

En general, si otras funciones base distintas a $\phi_j(x) = x^j$ son usadas, la expresión del modelo viene dada por:

$$y(x, \vec{w}) = \sum_{j=0}^{M} w_j \phi_j(x), \tag{5}$$

$$\Rightarrow y(x, \vec{w}) = \vec{w} \cdot \vec{\phi}(x), \tag{6}$$

donde el conjunto de funcionales $\{\phi_1(x), \phi_2(x), \ldots, \phi_M(x)\}$ se le denomina conjunto base, o **conjunto de funciones base**.

En los modelos lineales, el orden viene dado por la cantidad de términos en la combinación lineal M, el cual define la complejidad del modelo. Una vez definido el modelo a implementar (lineal o no lineal, conjunto de funciones base), el problema general de ajuste de modelo al conjunto de observaciones $\mathcal{D} = \left\{\vec{x}, \vec{t}\right\}$ se reduce a determinar el valor de los pesos $\vec{w}$. Un enfoque sencillo para calcular los valores de $\vec{w}$ óptimos es el de **mínimos cuadrados**, el cual propone la expresión de la función de error como sigue:

$$E(\vec{w}) = \frac{1}{2}\sum_{n=0}^{N}\left\{y\left(x_n, \vec{w}\right) - t_n\right\}^2 = \frac{1}{2}\sum_{n=1}^{N}\left\{\sum_{j=0}^{M} w_j x_n^j - t_n\right\}^2, \tag{7}$$

En tal esquema de mínimos cuadrados, el error se define entonces en como la diferencia al cuadrado de la evaluación del modelo en cada muestra de $\vec{x} = [x_0, ..., .x_N]^T$ con cada uno de las muestras de su salida $\vec{t} = [t_0, ..., .t_N]^T$ usando entonces el conjunto de datos de entrenamiento $\mathcal{D} = \left\{\vec{x}, \vec{t}\right\}$. Nótese que es posible **utilizar otras funciones de error** como por ejemplo la diferencia absoluta. El enfoque de mínimos cuadrados propone minimizar la expresión de error propuesta en la ecuación 7, por cada uno de los pesos en el vector $\vec{w}$. Para ello la ecuación 7 es derivada parcialmente respecto a w_i e igualada a cero:

$$\frac{\partial E\left(\vec{w}\right)}{w_i} = \sum_{n=1}^{N}\left\{\sum_{j=0}^{M} w_j x_n^j - t_n\right\} x_n^i = 0, \tag{8}$$

Despejando entonces la expresión 8, se obtiene un conjunto de $M+1$ ecuaciones ($i = 0, 1, 2, 3, \ldots, M$) donde los valores w_j son desconocidos, por lo que para una ecuación i tenemos:

$$\sum_{n=1}^{N}\sum_{j=0}^{M} w_j x_n^{j+i} = \sum_{n=0}^{N} t_n x_n^i, \tag{9}$$

y cambiando el orden a las sumatorias se obtiene:

$$\sum_{j=0}^{M}\sum_{n=1}^{N} w_j x_n^{j+i} = \sum_{n=0}^{N} t_n x_n^i \tag{10}$$

Desarrollando el sistema de matrices en 10, **para los M pesos**, obtenemos la siguiente expresión:

$$\begin{array}{ll}
(i=0) & w_0\sum_{n=1}^{N} 1 + w_1\sum_{n=1}^{N} x_n + w_2\sum_{n=1}^{N} x_n^2 + \ldots + w_M\sum_{n=1}^{N} x_n^M = \sum_{n=1}^{N} t_n \\
(i=1) & w_0\sum_{n=1}^{N} x_n + w_1\sum_{n=1}^{N} x_n^2 + w_2\sum_{n=1}^{N} x_n^3 + \ldots + w_M\sum_{n=1}^{N} x_n^{M+1} = \sum_{n=1}^{N} x_n t_n \\
(i=2) & w_0\sum_{n=1}^{N} x_n^2 + w_1\sum_{n=1}^{N} x_n^3 + w_2\sum_{n=1}^{N} x_n^4 + \ldots + w_M\sum_{n=1}^{N} x_n^{M+2} = \sum_{n=1}^{N} x_n^2 t_n \\
 & \vdots \\
(i=M) & w_0\sum_{n=1}^{N} x_n^M + w_1\sum_{n=1}^{N} x_n^{M+1} + w_2\sum_{n=1}^{N} x_n^{M+2} + \ldots + w_M\sum_{n=1}^{N} x_n^{2M} = \sum_{n=1}^{N} x_n^M t_n
\end{array} \tag{11}$$

Los valores w_j obtenidos al resolver este sistema de ecuaciones anterior se representan en el vector de pesos óptimo $\vec{w}_{\text{opt}}$.

En términos matriciales, lo anterior se desarrolla como:

$$\begin{bmatrix} \sum_{n=1}^{N} 1 & \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 & \cdots & \sum_{n=1}^{N} x_n^M \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 & & \sum_{n=1}^{N} x_n^{M+1} \\ \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 & \sum_{n=1}^{N} x_n^4 & & \sum_{n=1}^{N} x_n^{M+2} \\ & & \vdots & & \\ \sum_{n=1}^{N} x_n^M & \sum_{n=1}^{N} x_n^{M+1} & \sum_{n=1}^{N} x_n^{M+2} & & \sum_{n=1}^{N} x_n^{2M} \end{bmatrix} \vec{w} = \begin{bmatrix} \sum_{n=1}^{N} t_n \\ \sum_{n=1}^{N} x_n t_n \\ \sum_{n=1}^{N} x_n^2 t_n \\ \vdots \\ \sum_{n=1}^{N} x_n^M t_n \end{bmatrix}$$

Para lo cual, si A es la primer matriz de izquierda a derecha, y B la matriz al lado derecho de la igualdad, se tiene que:

$$\vec{w}_{\text{opt}} = A^{-1}\, B$$

Para evaluar el error resultante de ajustar un modelo de dimensión M y con pesos $\vec{w}_{\text{opt}}$, se utiliza el error normalizado o **error RMS**, expresado como sigue:

$$E\left(\vec{w}_{\text{opt}}\right)_{\text{RMS}} = \sqrt{2E\left(\vec{w}_{\text{opt}}\right)/N}. \tag{12}$$

Un modelo más complejo, es decir, con una dimensionalidad M más alta, minimiza el error frente al conjunto de datos de *entrenamiento* $\vec{t}$, como la Figura 2 muestra el ajuste de la curva con distintos M.

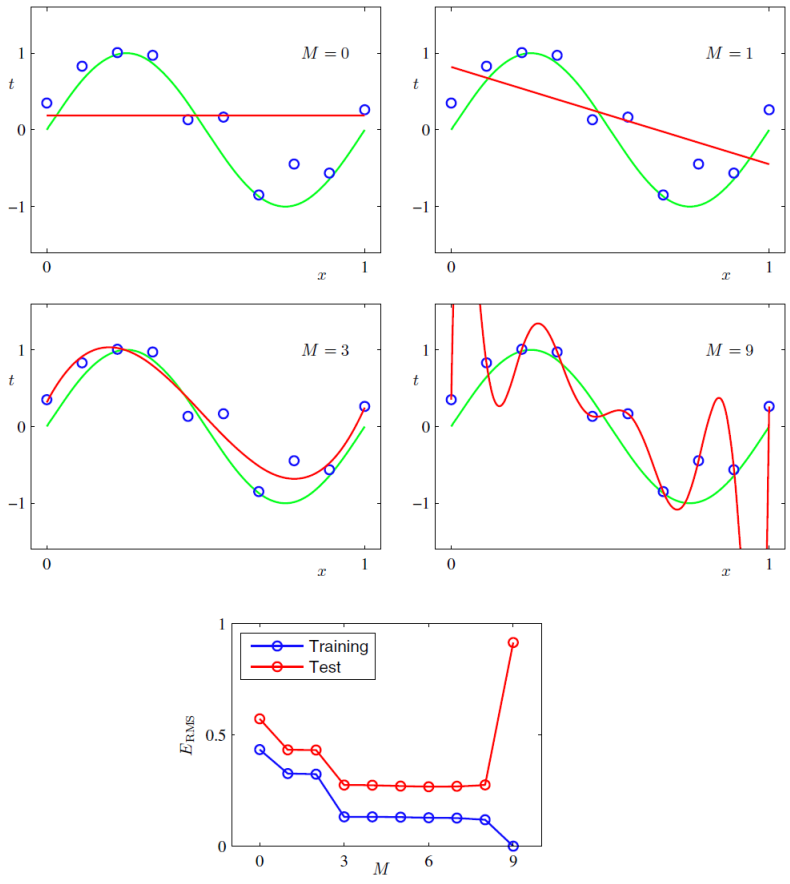

Figura 2: Gráfica del error conforme M aumenta.

El ajuste a los datos depende la cantidad de parámetros M y la cantidad de datos N, como además lo muestra la Figura 3, puesto que una cantidad de parámetros alta frente una cantidad de muestras baja, producirá un mayor sobreajuste.

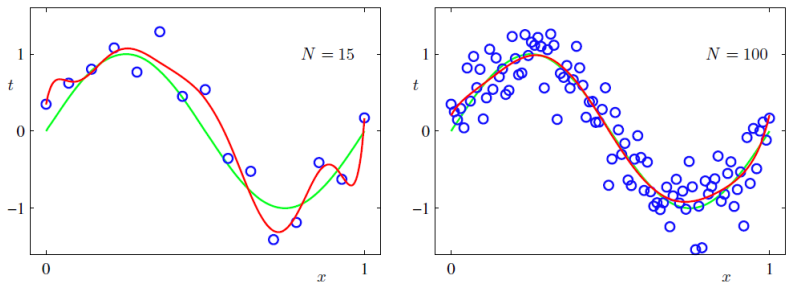

Figura 3: Ajuste en los datos con $M = 9$.

La Figura 3, muestra además que de sobre-ajustarse a una gran cantidad de datos con sesgos o ruido, el modelo no realizará una estimación adecuada ante nuevos datos. Es por ello que es recomendable **no sobre-ajustar el modelo a los datos,** escogiendo una cantidad de parámetros M apropiada. Para eliminar la relación entre la cantidad de datos y la cantidad de parámetros, se utilizan **modelos regularizados.**

2. Mínimos cuadrados regularizado

La expresión de error regularizada agrega un término $\frac{\lambda}{2}\left\|\vec{w}\right\|_2^2$ con magnitud $\ell = 2$, para controlar la magnitud del vector de pesos $\vec{w}$ y su dimensionalidad, de forma más simple:

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y\left(x_n, \vec{w}\right) - t_n\right\}^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|^2. \tag{13}$$

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\sum_{j=0}^{M} w_j x_n^j - t_n\right\}^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|^2, \tag{14}$$

En términos matriciales, se define

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^M \\ 1 & x_3 & x_3^2 & x_3^3 & \cdots & x_3^M \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \dots & x_N^M \end{bmatrix},$$

para el caso del modelo polinomial, por lo que X es de dimensiones $\mathbb{R}^{N\times M}$, $\vec{t}$ de $\mathbb{R}^{N\times 1}$ y $\vec{w} \in \mathbb{R}^{M\times 1}$. Recordando además que la norma al cuadrado de un vector como $\left\|\vec{w}\right\|^2 = \vec{w}^T\,\vec{w}$. Se puede reescribir matricialmente la salida del

modelo definido en la ecuación 3 como:

$$y(X,\vec{w}) = X\,\vec{w} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^M \\ 1 & x_3 & x_3^2 & x_3^3 & \cdots & x_3^M \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \dots & x_N^M \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix} = \begin{bmatrix} \sum_{m=0}^{M} w_m\, x_1^m \\ \sum_{m=0}^{M} w_m\, x_2^m \\ \sum_{m=0}^{M} w_m\, x_3^m \\ \vdots \\ \sum_{m=0}^{M} w_m\, x_N^m \end{bmatrix},$$

y la ecuación 13 es reescrita en términos matriciales:

$$E\left(\vec{w}\right) = \frac{1}{2}\left\|X\,\vec{w} - \vec{t}\right\|^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|^2 = \frac{1}{2}\left(X\,\vec{w} - \vec{t}\right)^T\left(X\,\vec{w} - \vec{t}\right) + \frac{\lambda}{2}\vec{w}^T\,\vec{w}, \quad (15)$$

$$\Rightarrow E\left(\vec{w}\right) = \frac{1}{2}\left(\vec{w}^T X^T - \vec{t}^T\right)\left(X\,\vec{w} - \vec{t}\right) + \frac{\lambda}{2}\vec{w}^T\,\vec{w}, \quad (16)$$

$$\Rightarrow E\left(\vec{w}\right) = \frac{1}{2}\vec{w}^T X^T X\,\vec{w} - \frac{1}{2}\vec{w}^T X^T\vec{t} - \frac{1}{2}\vec{t}^T X\,\vec{w} + \frac{1}{2}\vec{t}^T\,\vec{t} + \frac{\lambda}{2}\vec{w}^T\,\vec{w}, \quad (17)$$

Calculando el gradiente de $\vec{w}$ e igualando a cero, para encontrar el mínimo error, para lo cual recordamos las reglas básicas del cálculo matricial:

- $\nabla_{\vec{x}}\left(\vec{x}^T\vec{x}\right) = 2\,\vec{x}$
- $\nabla_{\vec{x}}\left(\left(A\,\vec{x}\right)^T\left(A\,\vec{x}\right)\right) = 2\,A\,\vec{x}$
- $\nabla_{\vec{x}}\left(\vec{b}^T\vec{x}\right) = \vec{b}$
- $\nabla_{\vec{x}}\left(\vec{x}^T\vec{b}^T\right) = \nabla_{\vec{x}}\left(\vec{b}\,\vec{x}\right)^T = \vec{b}^T$
- $\nabla_{\vec{x}}\left(\vec{x}^T\,A\,\vec{x}\right) = 2\,A\,\vec{x}$

la ecuación 15:

$$\nabla_{\vec{w}}\left(E\left(\vec{w}\right)\right) = 0 \Rightarrow \nabla_{\vec{w}}\left(\frac{1}{2}\vec{w}^T X^T X\,\vec{w} - \frac{1}{2}\vec{w}^T X^T\vec{t} - \frac{1}{2}\vec{t}^T X\,\vec{w} + \frac{1}{2}\cancelto{0}{\vec{t}^T\,\vec{t}} + \frac{\lambda}{2}\vec{w}^T\,\vec{w},\right) = 0,$$

evaluando el gradiente:

$$X^T X\,\vec{w} - \frac{1}{2}X^T\vec{t} - \frac{1}{2}X^T\vec{t} + \lambda\vec{w} = 0,$$

$$\nabla_{\vec{w}}\left(E\left(\vec{w}\right)\right) = X^T X\,\vec{w} - X^T\vec{t} + \lambda\vec{w} = 0,$$

que equivale a:

$$\left(X^T X + \lambda I\right)\vec{w} = X^T\vec{t},$$

para finalmente obtener la expresión de $\vec{w}$:

$$\vec{w} = \left(X^T X + \lambda I\right)^{-1} X^T \vec{t}.$$

La Figura 4 muestra como el valor de λ más alto disminuye el sobre-ajuste a los datos.

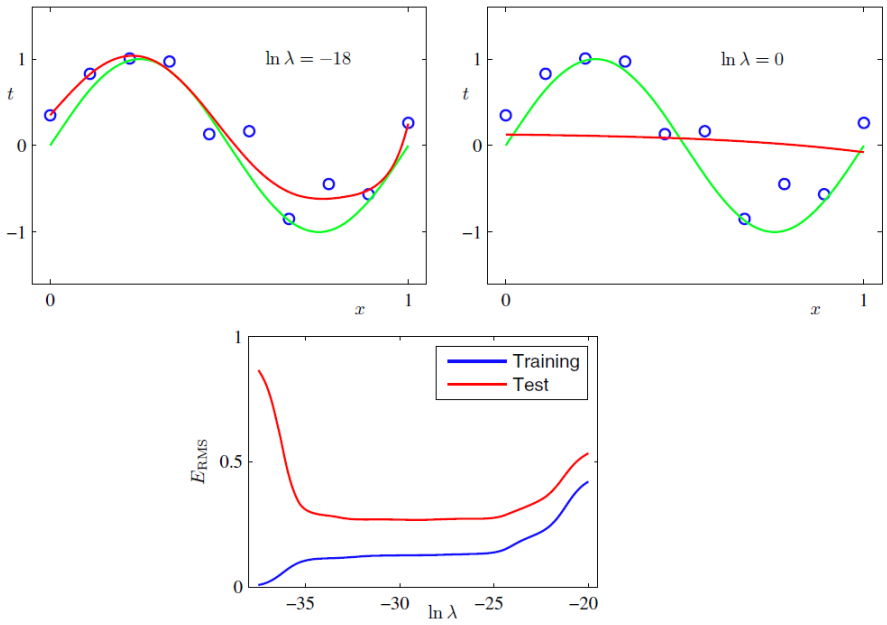

Figura 4: Ajuste a los datos de un modelo regularizado.

Preprocesamiento de las muestras: Para evitar problemas de *castigos* en el término de regularización dependientes de la escala de las muestras en las N observaciones $\vec{x} = [x_1, ..., .x_N]^T$, se recomienda normalizar las muestras respecto a la desviación estándar, de modo que:

$$\tilde{x}_i = \frac{x_i}{\sqrt{\frac{1}{N-1}\sum_{j=0}^{N-1}\left(x_j - \overline{x}\right)^2}}$$

En estadística, usualmente el modelo se define de la siguiente forma:

$$y(x, \vec{w}) = \beta + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=1}^{M} w_j x^j + \beta$$

donde β se le llama el **sesgo o *bias***.

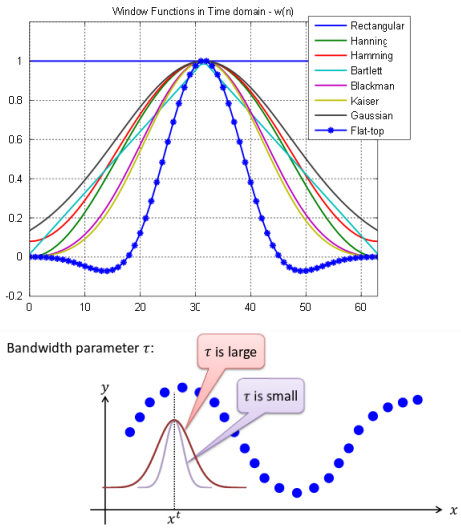

Figura 5: Funciones ventana.

3. Regularizacion por ponderamiento local

Otra alternativa para evitar usar modelos con muchos parametros y minimizar el riesgo de sobre-ajuste, es realizar un pesado o **ponderamiento local del vecindario**. De esta forma, es posible ajustar un modelo mas simple, al vecindario del punto a estimar. Esto se expresa en la funcion de error como sigue:

$$E\left(\vec{w}\right) = \frac{1}{2}\overrightarrow{\theta}^{T}\left\|X\,\vec{w} - \vec{t}\right\|^{2}$$

donde $\overrightarrow{\theta} \in \mathbb{R}^{N}$(un peso por observacion). Por ejemplo, para estos efectos, se puede utilizar un ponderado Gaussiano, el cual calcula los pesos usando una funcion Gaussiana:

$$\theta_i = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\left(\frac{x_i - x}{\tau}\right)^2}$$

De esta forma, los x que estan lejos del valor x_i, tendran un peso θ_i bajo, es decir, de forma inversa, si $|x_i - x| \to 0$, entonces $\theta_i \to 1$. El valor τ controla la influencia de los vecinos en la regresion. Pueden utilizarse distintas funciones ventana, como las que se ilustran en la Figura 5.

Esto quiere decir que se da un peso mucho mas alto a los errores cercanos al x a estimar. En *test-time* necesitaremos conservar el conjunto de datos en X, para hacer el ajuste o entrenamiento local, calculando $\overrightarrow{\theta}^{T}$segun la entrada x que reciba el modelo. La Figura 6 muestra una ilustracion sobre el resultado de una regresion local con ponderado local. Similar a los algoritmos **no-parametricos**, para este modelo necesitamos guardar el conjunto de datos completo.

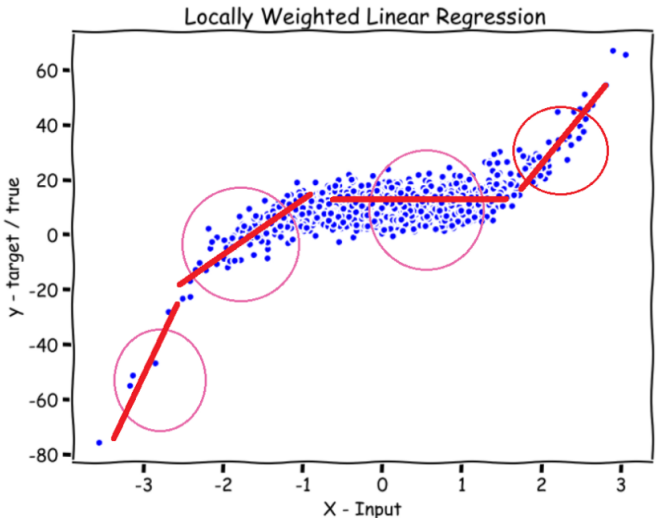

Figura 6: Regresion lineal con ponderado local.

4. Regresión multivariable y selección de características por el método empotrado

En la regresión multivariable, se define cada muestra como un vector de D dimensiones, componentes o características de forma que una muestra i está dada por $\vec{x}_i \in \mathbb{R}^D$. De esta forma, la salida del modelo y viene dada por:

$$y\left(\vec{x}, W\right) = \sum_{i=0}^{D-1} \sum_{j=0}^{M} x_i^j W_{i,j} = \sum_{i=0}^{D-1} \sum_{j=1}^{M} x_i^j W_{i,j} + \beta$$

donde entonces los pesos a estimar pasan a estar representados en la matriz $W \in \mathbb{R}^{D \times M}$, y se le asigna entonces un peso distinto a cada dimensión y a cada componente del polinomio de grado M. Observe como entonces el aumentar la dimensionalidad D de cada muestra, aumenta la cantidad de parámetros a estimar, y aumenta por ende el riesgo de sobreajuste, lo cual se corresponde con la **maldición de la dimensionalidad**.

Observe que por ejemplo, si se desea implementar un modelo lineal ($M = 1$) para una muestra de $D = 3$, en este caso, la expresión desarrollada del modelo viene dada por:

$$y\left(\vec{x}, W\right) = \sum_{i=0}^{2} \sum_{j=0}^{1} x_i^j W_{i,j} = W_{0,0} + x_0 W_{0,1} + W_{1,0} + x_1 W_{1,1} + W_{2,0} + x_2 W_{2,1} = \beta + x_0 W_{0,1} + x_1 W_{1,1} + x_2 W_{2,1}$$

con el sesgo definido por $\beta = W_{0,0} + W_{1,0} + W_{2,0}$, con lo que la cantidad total de parámetros a estimar viene dada por $D \times M + 1$.

Para la **selección de características** o dimensiones más importantes, es usual utilizar modelos sencillos y evaluar su error en un conjunto de datos. Por ejemplo se puede seleccionar para muestras $\vec{x}_i \in \mathbb{R}^D$ la búsqueda de los parámetros de un modelo de orden $M = 1$, con lo que entonces el modelo se define como:

$$y\left(\vec{x}, W\right) = \sum_{i=0}^{D-1} x_i W_i + \beta$$

la regularización LASSO como se verá más adelante, es posible utilizar interpretar el resultado para posibilitar la eliminación de características que aportan poco valor al modelo.

De forma similar a la expresion matricial para el modelo en una variable, se puede reescribir la funcion de error para la regresion multi-variable como:

$$E\left(\vec{w}\right) = \frac{1}{2}\left\|X\,\vec{w} - \vec{t}\right\|^2$$

Donde en este caso, definimos X como:

$$X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,1}^M & x_{1,2} & x_{1,2}^M & \dots & x_{1,D} & x_{1,D}^M \\ 1 & x_{2,1} & \dots & x_{2,1}^M & x_{2,2} & x_{2,2}^M & \dots & x_{2,D} & x_{2,D}^M \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \dots & x_{N,D} & \vdots \\ 1 & x_{N,1} & \dots & x_{N,1}^M & x_{N,2} & x_{N,2}^M & \dots & x_{N,D} & x_{N,D}^M \end{bmatrix},$$

Y de forma similar, para ajustar los pesos, podemos usar el metodo de minimos cuadrados expuesto para regresion en una variable:

$$\vec{w} = \left(X^T X\right)^{-1} X^T \vec{t}.$$

4.1. Regularizacion: Minimos cuadrados ponderado (Weighted Least Squares)

El método de los mínimos cuadrados ordinarios supone que existe una varianza constante en los errores (lo que se denomina **homocedasticidad**). El método de mínimos cuadrados ponderados se puede utilizar cuando se viola el supuesto de mínimos cuadrados ordinarios de varianza constante en los errores (lo que se denomina **heterocedasticidad**). Es decir, si las varianzas de cada dimension o caracteristica son muy diferentes, podemos utilizarlas para dar un peso diferente a cada dimension, usando la matriz R:

$$R = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_1^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_D^2 \end{bmatrix}$$

Esta matriz se puede extender (repitiendo las filas y columnas pertinentes) en R' para pesar cada dimension de acuerdo a su varianza y modificar el calculo

de los pesos optimos como:

$$\vec{w} = \left(X^T R' X\right)^{-1} R' X^T \vec{t}.$$

4.2. Regularizacion: Métodos de achicamiento

La regularización generalizada con la norma $\ell = p$:

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y\left(\vec{x}_n, \vec{w}\right) - t_n\right\}^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|_p^2$$

agrega el término de regularización $\frac{\lambda}{2}\left\|\vec{w}\right\|_p^2$ el cuál es también referido en el aprendizaje estadístico como la **penalización de achicamiento** o *shrinkage penalty*. Cuando $\lambda = 0$ observe la regularización es nula, y termina en una regresión básica por mínimos cuadrados, aumentando la probabilidad de sobre mientras que a medida que $\lambda \to \infty$, los pesos estimados en $\vec{w}$ se achican. La selección del valor λ adecuado es crítico, pues un λ_0 específico generará un arreglo de pesos $\vec{w}_{\lambda_0}$ asociado. Para su selección se utilizarán técnicas de *validación cruzada* y *bootstrap* los cuales se estudiarán posteriormente.

4.2.1. Regresión Ridge

La regresión Ridge implementa una penalización de achicamiento $\ell = 2$, con $\vec{w} \in \mathbb{R}^{M-1}$

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y\left(x_n, \vec{w}\right) - t_n\right\}^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|_2^2$$

donde en este caso el sesgo β no es regularizado, de estar forma el sesgo queda libre e independiente del coeficiente de regularización λ, por lo que usualmente cuando $\lambda \to \infty$, el sesgo tiende también a crecer $\beta \to \infty$. Cuando $\lambda \to \infty$ la flexibilidad o varianza del modelo $y\left(x_n, \vec{w}\right)$ tiende a decrecer, lo que se conoce como el **intercambio entre varianza y sesgo**. Una cantidad alta de parámetros $M \to \infty$ hará que **todos los coeficientes se achiquen** con un λ adecuado, por lo que la regresión Ridge da un peso similar a todas las muestras x.

4.2.2. Regresión por valor absoluto menor para selección y achicamiento o LASSO

En la regresión LASSO se implementa una regularización con norma $\ell = 1$ ($\left\|\vec{w}\right\|_1 = |\vec{w}|$). La regresión LASSO (*least absolute shrinkage and selection operator*) permite poner en cero los coeficientes de las dimensiones o características en cada muestra $\vec{x}$, lo cual permite seleccionar un subconjunto de dimensiones o características específicas. Para ello usaremos el modelo de orden $M = 1$ comentado anteriormente,

$$y\left(\vec{x}, W\right) = \sum_{i=0}^{D-1} x_i W_i + \beta$$

con regularización por norma $\ell = 1$:

$$E(\vec{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\vec{x}_n, \vec{w}) - t_n\}^2 + \lambda |\vec{w}|$$

de forma similar con Ridge, LASSO, cuando $\lambda \to \infty$, los coeficientes se achican. Sin embargo, en el caso de LASSO, la norma $\ell = 1$ tiene el **efecto de forzar a algunos de los pesos a ser exactamente cero** o a ser muy pequeños respecto a los demás coeficientes, cuando λ es lo suficientemente grande. Es por esto que LASSO es utilizado para la **selección de características**. Esto facilita la interpretación de los modelos generados por LASSO, respecto a los modelos generados por Ridge, pues los pesos que tienden a cero indican características que contribuyen poco al modelo.

4.3. El balance entre preprocesado y clasificación

El preprocesado de las muestras como bien se estudió en apartados anteriores puede disminuir el error en la etapa posterior de clasificación. Para ilustrar esto de forma simple, tómese el siguiente conjunto de datos $\overrightarrow{x} = [x_1, x_2, \ldots, x_N]$, con cada muestra $x_i \in \mathbb{R}$, y un sistema para el cual tenemos por objetivo construir un modelo:

$$y = \log(x), \tag{18}$$

Las salidas muestreadas en este caso presentan **ruido Gaussiano artificial** $\epsilon(x)$, con $\mu = 0$ y una desviación estándar σ (y precisión $\beta = 1/\sigma$) sobre y':

$$t = y(x) + \epsilon(x). \tag{19}$$

como lo muestra la Figura 7.

Construir un modelo polinomial $y(x, \vec{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$ requerirá sin dudas una cantidad de parámetros $M > 1$ para lograr un error bajo, por lo que un modelo lineal respecto a las entradas $y(x, \vec{w}) = w_0 + w_1 x$ será insuficiente. El costo computacional y sobre todo el riesgo de sobre ajuste aumenta conforme aumentamos los parámetros M del modelo, por lo que es de interés encontrar alternativas que permitan utilizar un modelo de regresión simple.

Una alternativa es preprocesar los datos X utilizados para entrenar al modelo $y(x, \vec{w})$. Dado que al graficar los datos se aprecia una relación cercana a logarítmica, ello sugiere que al preprocesar los datos de entrada

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \ldots & x_1^M \\ 1 & x_2 & x_2^2 & x_2^3 & \ldots & x_2^M \\ 1 & x_3 & x_3^2 & x_3^3 & \cdots & x_3^M \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 & \ldots & x_N^M \end{bmatrix}$$

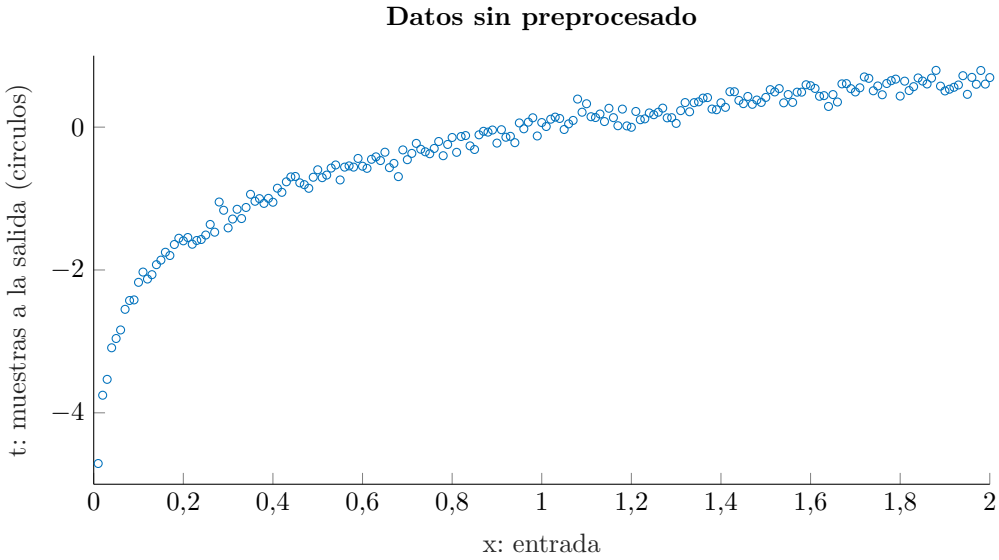

Figura 7: Sistema artificial con comporamiento logarítmico y ruido gaussiano aditivo con $N = 100$.

utilizando una transformación similar a la del modelo, se recupere una relación lineal entre las entradas x y las salidas y' del fenómeno. Por ejemplo, al preprocesar los datos aplicando el logaritmo sobre las entradas originales:

$$X' = \log(X)$$

y graficar los datos preprocesados X' y las salidas $\overrightarrow{t} = [t_1, t_2, \ldots, t_N]^T$ se observa una relación casi lineal, puesto que $\overrightarrow{x}' \approx \overrightarrow{t}$ puesto que

$$e^{\log(X)} \approx e^{\log(X)+\epsilon(x)} \Rightarrow X' \approx X' + \epsilon(x)$$

lo cual no es más que una correlación lineal dereciente como lo ilustra la Figura 8. Se obtiene el vector de pesos con los datos preprocesados, usando $M = 1$ parámetros:

$$\vec{w}' = \left(X'^T X' + \lambda I\right)^{-1} X'^T \vec{t},$$

obtieniendo un modelo lineal $y(x', \vec{w}') = w'_0 + w'_1 x'$, el cual se ajusta a los datos preprocesados como se visualiza en la misma Figura 8.

Para evaluar el error cuadrático es necesario comparar la salida del modelo lineal $\overrightarrow{y}' = y(\overrightarrow{x}, \vec{w}')$ con los datos de verdad muestrados $\overrightarrow{t} = [t_1, t_2, \ldots, t_N]^T$ es necesario transformar tales datos con la función exponencial de forma que el error se evalúa como sigue:

$$E(\overrightarrow{y}') = \left\| \overrightarrow{y}' - e^{\overrightarrow{t}} \right\| = \left\| \log(\overrightarrow{y}') - \overrightarrow{t} \right\|$$

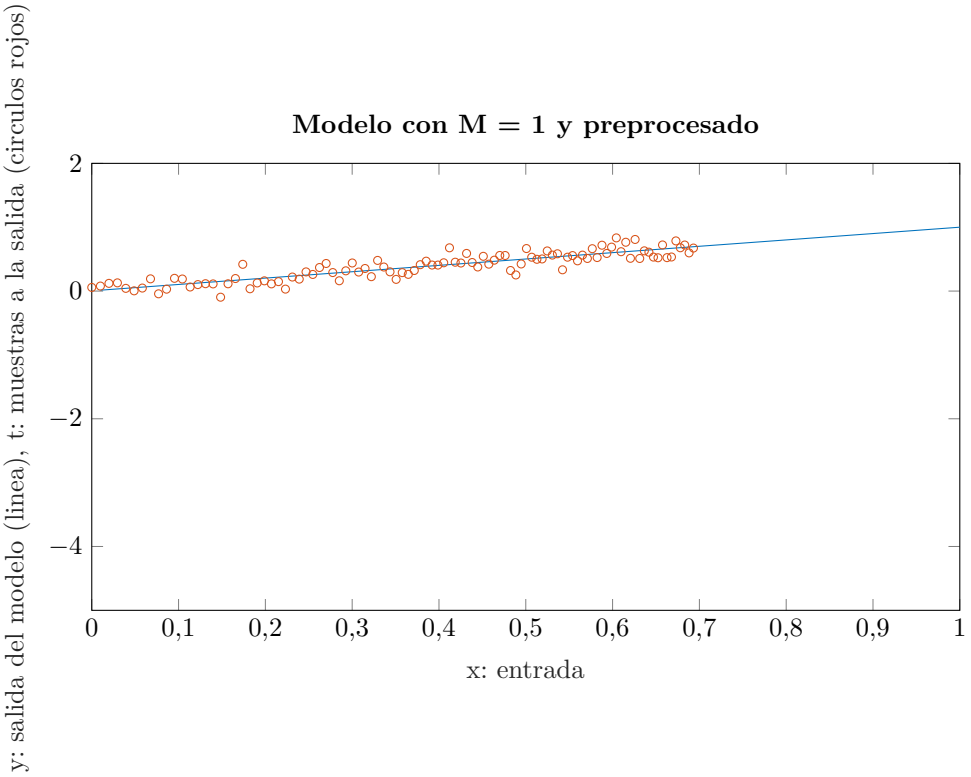

Figura 8: Modelo estimado $y\left(X', \vec{w}'\right)$ con las muestras preprocesadas $\overrightarrow{x}'$.

con lo cual realizando las simulaciones se puede verificar que

$$E\left(\overrightarrow{y}'\right) \ll E\left(\overrightarrow{y}\right).$$

Referencias

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.