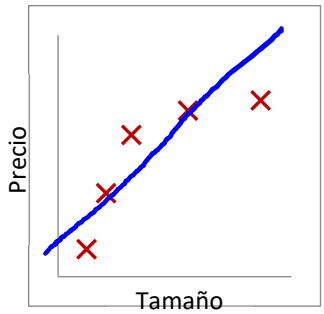
4. Regularización

La regularización viene a proponer una alternativa para evitar el problema del sobreajuste (overfitting), que consiste en la determinación de parámetros theta que permiten valores que pueden predecir con bastante precisión los valores de y, en el conjunto de entrenamiento, sin embargo, cuando se utilizan nuevos datos, estos no responden de manera eficiente. Para mejorar la explicación del overfitting se presentan las siguientes graficas:

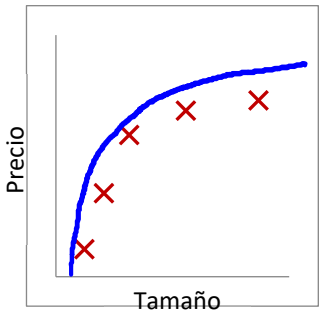
Ejemplo en el caso de una regresión lineal:

Consideremos el problema de predecir y a partir de x ∈ R. La figura de la izquierda muestra el resultado de ajustar $y = \theta_0 + \theta_1 x$ a un conjunto de datos. Vemos que los datos no se encuentran realmente en línea recta, por lo que el ajuste no es muy bueno como muestra el primer gráfico.



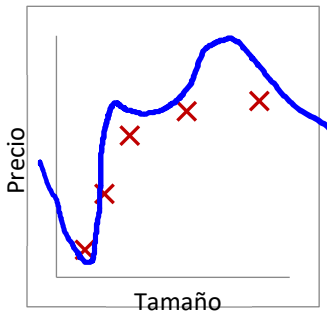
$\theta_0 + \theta_1 x$

Underfit – High bias
Subajuste



$\theta_0 + \theta_1 x + \theta_2 x^2$

Correcto



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Overfit – High variance
Sobreajuste

En cambio, si hubiéramos añadido una característica adicional x^2, y hubiéramos ajustado $y = \theta_0 + \theta_1 x + \theta_2 x^2$, obtendríamos un ajuste ligeramente mejor a los datos (véase la figura central). Ingenuamente, podría parecer que cuantas más características añadamos, mejor. Sin embargo, también existe el peligro de añadir demasiadas características: La figura de la derecha es el resultado de ajustar un polinomio de orden 5th $\sum_{j=0}^{5} \theta_j x^j$. Vemos que, aunque la curva ajustada pasa perfectamente por los datos, no esperaríamos que fuera un predictor muy bueno de, por ejemplo, los precios de la vivienda (y) para distintas zonas habitables (x). Sin definir formalmente lo que significan estos términos, diremos que la figura de la izquierda muestra un caso de infraajuste, en el que los datos muestran claramente una estructura no captada por el modelo y la figura de la derecha es un ejemplo de sobreajuste.

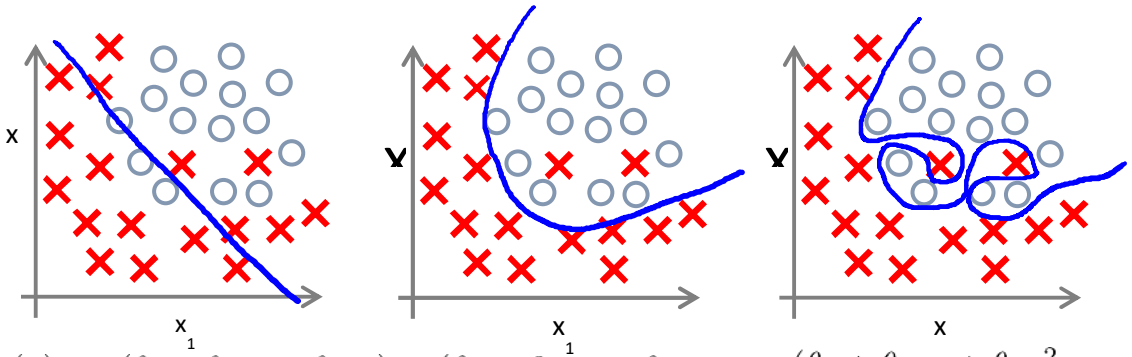
El infraajuste, o alto sesgo, se produce cuando la forma de nuestra función de hipótesis h se corresponde mal con la tendencia de los datos. Suele estar causada por una función demasiado simple o que utiliza muy pocas características. En el otro extremo, el sobreajuste, o alta varianza, está causado por una función de hipótesis que se ajusta a los datos disponibles, pero no generaliza bien para predecir nuevos datos. Suele estar causado por una función complicada que crea muchas curvas y ángulos innecesarios sin relación con los datos.

Esta terminología se aplica tanto a la regresión lineal como a la logística. Existen dos opciones principales para abordar el problema del sobreajuste:

1. Reducir el número de características:
    a. Seleccionar manualmente qué características conservar.
    b. Utilizar un algoritmo de selección de modelos.
2. Regularización
    a. Mantenga todas las características, pero reduzca la magnitud de los parámetros θ_j.
    b. La regularización funciona bien cuando tenemos muchas características ligeramente útiles.

Si se tienen muchas características, la hipótesis aprendida puede ajustar el conjunto de datos de entrenamiento muy bien ($J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0$), pero puede fallar al generalizar nuevos ejemplos (predecir precios sobre nuevos terrenos, como ejemplo).

Ejemplo en el caso de una regresión logística:



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g: sigmoid function)

Underfit – High bias

Subajuste

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

Correcto

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$

Overfit – High variance

Sobreajuste

4.1 Función de costes

Para comprender mejor la técnica de regularización, es importante analizar lo que sucede con la función de coste e hipótesis.

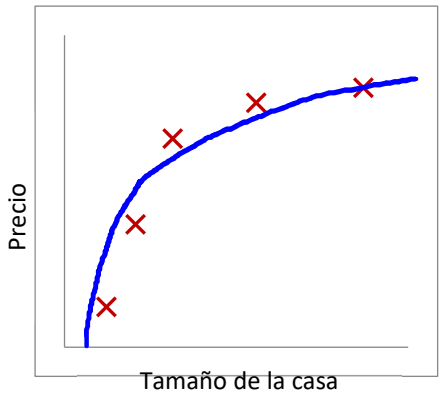
Si tenemos un sobreajuste de nuestra función de hipótesis, podemos reducir el peso que tienen algunos de los términos de nuestra función aumentando su coste.

Digamos que queremos hacer la siguiente función más cuadrática:$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

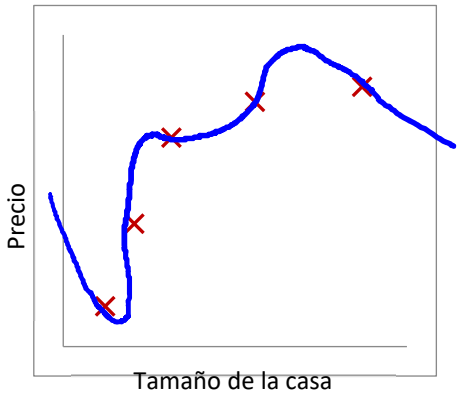
Querremos eliminar la influencia de $\theta_3 x^3$ y $\theta_4 x^4$ Sin deshacernos realmente de estas características ni cambiar la forma de nuestra hipótesis, podemos en cambio modificar nuestra función de coste:

$$\min_\theta \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + 100 * \theta_3^2 + 100 * \theta_4^2$$

Hemos añadido dos términos adicionales al final para inflar el coste de $\boldsymbol{\theta_3}$ $\boldsymbol{y}$ $\boldsymbol{\theta_4}$. Ahora, para que la función de coste se acerque a cero, tendremos que reducir los valores de $\boldsymbol{\theta_3}$ $\boldsymbol{y}$ $\boldsymbol{\theta_4}$ hasta casi cero. Esto, a su vez, reducirá en gran medida los valores de $\boldsymbol{\theta_3 x^3 y\ \theta_4 x^4}$ en nuestra función de hipótesis. Como resultado, vemos que la nueva hipótesis (representada por la curva rosa) parece una función cuadrática, pero se ajusta mejor a los datos debido a los pequeños términos $\boldsymbol{\theta_3 x^3 y\ \theta_4 x^4}$



$$\theta_0 + \theta_1 x + \theta_2 x^2$$



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Supongamos que penalizamos y hacemos muy pequeño $\boldsymbol{\theta_3}$ $\boldsymbol{y}$ $\boldsymbol{\theta_4}$

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + 1000 * \theta_3^2 + 1000 * \theta_4^2$$

Se logra que:

$$\theta_3 \approx 0\ y\ \theta_4 \approx 0$$

También podríamos regularizar todos nuestros parámetros theta en una única suma como:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Lambda, es el **parámetro de regularización**. Determina cuánto se inflan los costes de nuestros parámetros theta. Utilizando la función de costes anterior con el sumatorio adicional, podemos suavizar la salida de nuestra función de hipótesis para reducir el sobreajuste. Si lambda se elige demasiado grande, puede suavizar demasiado la función y provocar un ajuste insuficiente. Por lo contrario, si $\lambda = 0$ es demasiado pequeño la función no lograra el suavizado que se busca, es decir que no se lograra salir del overfitting.

Podemos aplicar la regularización tanto a la regresión lineal como a la regresión logística. Abordaremos primero la regresión lineal.

4.2 Regresión lineal regularizada

Antes de proceder con la explicación de la regularización aplicada a la regresión lineal es importante considerar que X es no invertible si m < n, y puede ser no invertible si m = n.

Para aplicar la regularización se debe modificar la función de descenso gradiente para separar θ_0 del resto de parámetros porque no queremos penalizar θ_0.

Donde el algoritmo se modificaría de la siguiente forma:

Repetir:

{

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_0^{(i)}$$

$$\theta_j = \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \quad j \in \{1, 2, \ldots, n\}$$

}

El término $\frac{\lambda}{m}\theta_j$ realiza nuestra regularización. Con alguna manipulación nuestra regla de actualización también puede representarse como:

$$\theta_j = \theta_j (1 - \alpha \frac{\lambda}{m}) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

El primer término de la ecuación anterior, $1 - \alpha \frac{\lambda}{m}$ siempre será menor que 1. Intuitivamente puede verse como la reducción del valor de θ_j en alguna cantidad en cada actualización. Observe que el segundo término es ahora exactamente igual que antes.

4.3 Ecuación de la normal

La regularización utilizando el método alternativo de la ecuación normal no iterativa. Para añadir la regularización, la ecuación es la misma que la original, salvo que añadimos otro término dentro de los paréntesis:

$$\theta = (X^T X + \lambda \cdot L)^{-1} X^T y$$

Donde:

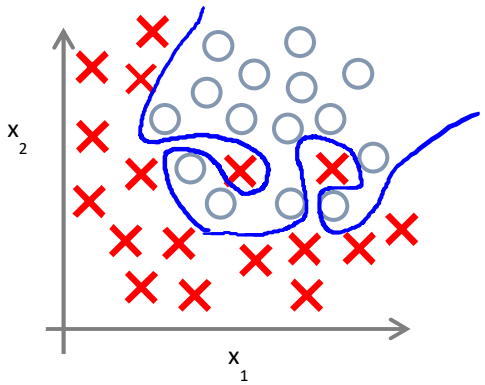
$$\boldsymbol{L} = \begin{bmatrix} \mathbf{0} & & & & \\ & \mathbf{1} & & & \\ & & \mathbf{1} & & \\ & & & \cdots & \\ & & & & \mathbf{1} \end{bmatrix}$$

$\boldsymbol{L}$ es una matriz con 0 en la parte superior izquierda y 1 en la diagonal, con 0 en todas las demás partes. Debe tener dimensión (n+1) × (n+1). Intuitivamente, se trata de la matriz identidad (aunque no incluimos $\boldsymbol{x_0}$), multiplicada por un único número real λ.

Recordemos que si m < n, entonces $X^T X$ es no invertible. Sin embargo, cuando añadimos el término λ·L, entonces $X^T X + \lambda \cdot L$ se vuelve invertible.

4.4 Regresión logística regularizada

Podemos regularizar la regresión logística de forma similar a como regularizamos la regresión lineal. Como resultado, podemos evitar el sobreajuste. La siguiente imagen muestra cómo la función regularizada, representada por la línea rosa, tiene menos probabilidades de sobreajustarse que la función no regularizada representada por la línea azul:



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \dots)$$

Donde la función de costo es la siguiente:

$$\boldsymbol{J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + \left(1 - y^{(i)}\right) \log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{m}\sum_{j=1}^{n} \theta_j^2}$$

Recordemos que la función de coste para la regresión logística era:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))\right]$$

Se puede regularizar esta ecuación añadiendo un término al final:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{m}\sum_{j=1}^{n} \theta_j^2$$

La segunda suma, $\sum_{j=1}^{n} \theta_j^2$ significa excluir explícitamente el término de sesgo, $\boldsymbol{\theta_0}$. Es decir, el vector θ está indexado de 0 a n (manteniendo n+1 valores, de $\boldsymbol{\theta_0}$ a $\boldsymbol{\theta_n}$), y esta suma se salta explícitamente $\boldsymbol{\theta_0}$, yendo de 1 a n, saltándose 0. Así, al calcular la ecuación, debemos actualizar continuamente las dos ecuaciones siguientes:

Repetir:

{

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_0^{(i)}$$

$$\theta_j = \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \quad j \in \{1, 2, \ldots, n\}$$

}

Similar a la regresión lineal.