Universidad de Ingeniería y Tecnología - UTEC
Departamento de Ingeniería Mecatrónica
Curso: Tópicos de Ingeniería Mecatrónica (2020 -2)

Tema 3: Regularización

Prof. Oscar E. Ramos Ponce

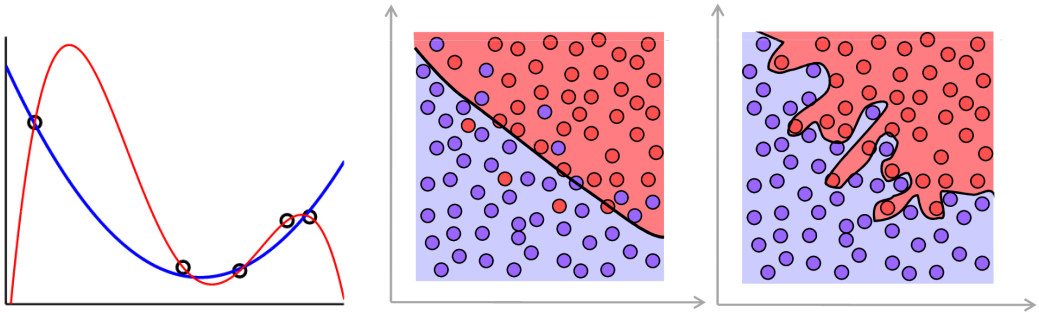
La regularización es una forma de complementar algoritmos de aprendizaje, como la regresión lineal o la regresión logística, para evitar algunos problemas que surgen en el modelamiento debido, usualmente, a la cantidad de parámetros. En particular busca resolver, o eliminar, el problema conocido como *overfitting* a través de la eliminación automática de algunos parámetros que podrían estar haciendo que el modelo sea demasiado complejo. En este sentido, busca "simplificar" el modelo.

1. Overfitting

El *overfitting*, llamado a veces sobreajuste, es un problema muy importante en todos los algoritmos de aprendizaje automático. Consiste en tener una función de hipótesis $h_w(\boldsymbol{x})$ que predice perfectamente, o casi perfectamente, el conjunto de entrenamiento, pero no generaliza a nuevos datos. En otros términos, cuando se aplica al conjunto de entrenamiento, el error (medido, por ejemplo, por la función de costo) es bajo o casi nulo, pero al aplicarse a un conjunto de prueba, el error es alto.

Este problema se origina usualmente cuando la función de hipótesis tiene bastantes parámetros (o grados de libertad), ya que estos parámetros son utilizados para memorizar de alguna manera los datos de entrenamiento. En cierta forma, esto hace que el sistema se comporte como una tabla de consulta, más que como un algoritmo de aprendizaje, fallando en puntos donde no ha habido entrenamiento.

De manera matemática se puede considerar, por simplicidad, que todas las posibles instancias pertenecen a una distribución probabilística fija, pero desconocida. Cada hipótesis h en esta distribución tiene un error real $J^*(h)$, que es el error esperado cuando los datos se obtienen de la distribución. Sin embargo, cuando no se tiene todas las instancias, sino solamente el conjunto de entrenamiento, se mide el error, denotado por $J_t(h)$, en este conjunto. Supóngase que se tiene dos funciones de hipótesis h_1 y h_2, y que en el conjunto de entrenamiento se cumple que $J_t(h_1) < J_t(h_2)$; es decir, la función h_1 se ajusta mejor a los datos de entrenamiento generando un menor error. Si, por el contrario, al aplicar ambas funciones a los datos de la distribución se obtiene $J^*(h_2) < J^*(h_1)$, indicando que en realidad h_2 es mejor que h_1, se tiene un problema con lo obtenido en el conjunto de entrenamiento. En este caso, este problema se denomina *overfitting*, y se dice que h_1 está haciendo un sobreajuste a los datos de entrenamiento, ya que en realidad h_2 es mejor.



(a) Regresión (b) Sin overfitting (c) Con overfitting

Figura 1: Ejemplos de *overfitting*. En (a) se muestra una regresión (en azul) sin overfitting, y una regresión (en rojo) con overfitting, que se ajusta demasiado a los datos de entrada. Un ejemplo de clasificación sin overfitting se muestra en (b) y con overfitting en (c).

Ejemplo. En la Fig. 1 se muestra un ejemplo de regresión y uno de clasificación. En el ejemplo de regresión (a), la curva azul es una hipótesis aceptable para los datos mostrados. Sin embargo, a medida que se incrementa el grado del polinomio, se incrementa la capacidad de sobreajustarse al conjunto de entrenamiento. Así, la curva roja es un polinomio de grado 4 (con 5 parámetros) que pasa por cada uno de los puntos de entrenamiento pero, como se puede observar, no generaliza adecuadamente. Algo similar sucede en la clasificación: en (b) se observa un clasificador relativamente adecuado, pero en (c) se observa uno que está demasiado ajustado al conjunto de entrenamiento. En este caso, para poder tener la forma presentada, la frontera de decisión está descrita por bastantes parámetros.

Para reducir el *overfitting* al momento de generar y entrenar una función de hipótesis se puede seguir alguna(s) de las siguientes recomendaciones.

- *Reducir del número de atributos.* La forma más básica consiste en manualmente seleccionar qué atributos se debería mantener. Alternativamente se puede utilizar algún algoritmo que permita la selección automática.
- *Regularizar.* Consiste en mantener todos los atributos, pero reducir la magnitud o valor de los parámetros w_i, de tal modo que algunos sean tan pequeños que su efecto sea casi nulo. Funciona bien cuando hay bastantes atributos, donde todos los atributos contribuyen en algo a la predicción.

2. Regularización

Para reducir el *overfitting* de los datos de entrenamiento se desea obtener una función de hipótesis más simple, lo cual es equivalente a tener pequeños valores para los parámetros $w_1, w_2, \cdots, w_n$. La regularización es un método que permite de forma relativamente automática alcanzar este efecto: reduce los valores de cada parámetro w_i generando funciones de hipótesis más simples.

Ejemplo. Si se tiene una función de hipótesis dada por $h_w(\boldsymbol{x}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$, pero se obtiene coeficientes w_3 y w_4 muy pequeños tales que $w_3 \approx w_4 \approx 0$, la función de hipótesis resultante será $h_w(\boldsymbol{x}) \approx w_0 + w_1 x + w_2 x^2$ que es una función más simple que la anterior.

2.1. Forma Genérica de Regularización

Si se denota el espacio de hipótesis como $\mathscr{H}$, se define la medida de complejidad de un modelo como $\Omega : \mathscr{H} \to [0, \infty)$ y en realidad depende de la función de hipótesis como $\Omega(h_w)$. Sin embargo, por simplicidad, dado que los modelos con los que se está trabajando dependen del parámetro $\boldsymbol{w}$, se representará la medida de complejidad como $\Omega(\boldsymbol{w})$. El objetivo genérico de la regularización consiste en reducir esta medida de complejidad.

Regularización de Ivanov. Es una forma de representar la regularización como un problema con restricciones, donde la función objetivo es la función de costo normalmente utilizada, y la restricción impone una cota superior r a la medida de complejidad. Normalmente a r se le llama la máxima complejidad del modelo. Matemáticamente, esta regularización se representa como

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) \\ \text{s.a.} \quad & \Omega(\boldsymbol{w}) \leq r \end{aligned}$$

donde $r \geq 0$ y constante, y $\ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)})$ es la función de costo asociada a cada instancia del modelo que se esté utilizando (como la regresión lineal o logística). Notar que en la expresión anterior, las siglas *s.a.* significan "sujeto a". Se observa, además, que el objetivo de la minimización se encuentra, escalado por $\frac{1}{n}$, pero en algunos casos específicos puede escalarse usando otras constantes, como $\frac{1}{2n}$ en la regresión lineal. Estos factores suelen utilizarse solamente por conveniencia matemática.

Regularización de Tikhonov. Es otra forma de representar la regularización, pero en este caso no se utiliza restricciones, sino que se añade la medida de complejidad al objetivo de minimización. Esto se expresa como

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) + \lambda \Omega(\boldsymbol{w}),$$

donde $\lambda \geq 0$ y se conoce como el parámetro de regularización y determina la importancia relativa que tiene el término de regularización. En esta forma, a la medida de complejidad $\Omega(\boldsymbol{w})$ también se le llama el término "regularizador". En algunos casos resulta más conveniente utilizar la forma de Tikhonov debido a que el problema de minimización no posee restricciones. En adelante, debido a esta simplificación de las restricciones, se utilizará solamente la forma de Tikhonov.

Para la mayoría de medidas de complejidad $\Omega(\boldsymbol{w})$, tanto la regularización de Ivanov como la de Tikhonov son equivalentes. En particular, esto se cumple para los tipos de regularización mayormente utilizados, como la regularización L_1 y L_2. Esto significa que cualquier solución $\boldsymbol{w}^*$ que se obtenga utilizando la forma de Ivanov, se puede también obtener de Tikhonov, y viceversa. En la práctica, ambas formas son igual de efectivas.

2.2. Tipos Principales de Regularización

Según la medida de complejidad que se utilice, existen tres tipos principales de regularización utilizados en aprendizaje automático y se conocen como regularización de arista (*ridge*), regularización Lasso, y regularización *elastic net*.

a. Regularización *Ridge*

Esta regularización, también conocida como regularización de arista o regularización L_2, consiste en escoger la norma L_2 como medida de complejidad; es decir,

$$\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2,$$

donde $\|\boldsymbol{w}\|_2^2 = w_1^2 + w_2^2 + \cdots + w_d^2$, asumiendo $\boldsymbol{w} \in \mathbb{R}^d$ (notar que no se suele incluir el término w_0 en la regularización). Usando este tipo de regularización, el problema de minimización se puede escribir como

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) + \lambda \|\boldsymbol{w}\|_2^2,$$

donde $\lambda \geq 0$ es el parámetro de regularización. En esta regularización, la norma L_2 controla la magnitud del vector de parámetros $\boldsymbol{w}$, tratando de hacer pequeños los valores de sus componentes. De manera equivalente, este problema de regularización se puede expresar como

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) + \lambda \sum_{j=1}^{d} w_j^2 \tag{1}$$

donde se aplica de forma explícita la norma L_2 para los parámetros w_i. No se suele regularizar el término w_0 debido a que es un *bias*. Por este motivo solamente se aplica regularización a los parámetros de w_1 hasta w_d, suponiendo que $\boldsymbol{x}$ tiene d atributos (para la regresión lineal o logística).

De manera gráfica, la Figura 2 muestra el efecto de aplicar esta regularización. Los elipses concéntricos son las curvas de nivel de la función de costo cuando no se incluye regularización. En este caso, el mínimo sería el punto azul señalado con la flecha. Si solamente se tuviese el término de regularización, lo que se desearía seria llevar a cero la norma L_2 del vector, y el mínimo sería el punto rojo señalado con la flecha. Cuando se añade la regularización a la función de costo, se encuentra un balance entre ambos casos; en este caso, ese balance se muestra con el punto verde.

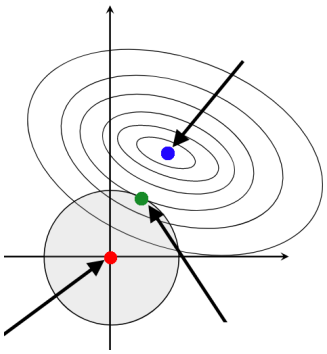


Figura 2: Efecto de la regularización L_2 (*ridge*).

Puede observarse que, el punto que minimiza la función de costo regularizada no es la solución al problema sin regularización; es más, es un punto un tanto "peor" para la función de costo inicial. Sin embargo, lo que se gana al regularizar es que la norma del vector $\boldsymbol{w}$ sea más pequeña. En la figura anterior, el alejamiento del punto del punto verde con respecto al punto azul se controla mediante el parámetro λ. Así, se tiene los siguientes casos.

- Si $\lambda = 0$, se elimina por completo el término de regularización y su efecto (se obtiene el punto azul).
- Si $\lambda \to \infty$, se elimina el efecto de la función de costo inicial y se hace que la norma L_2 de $\boldsymbol{w}$ tienda a cero (se obtiene el punto rojo). Esto genera un *underfitting* ya que se tendrá $h_w(\boldsymbol{x}) \approx 0$.
- Si $\lambda > 0$, pero no tan alto, se obtiene un compromiso entre la función de costo inicial y el término de regularización, de tal modo que $\boldsymbol{w}$ será menor que sin regularización (punto verde).

En general, escoger un valor de λ adecuado puede ser un proceso complejo. A veces se utiliza validación cruzada (*cross-validation*) para encontrar un valor aceptable. Se debe además notar que si existiesen atributos de $\boldsymbol{x}$ que no son importantes, esta regularización puede hacer a los parámetros asociados pequeños, pero no los llega a eliminar.

Como se verá en las aplicaciones específicas (regresión lineal o regresión logística), cuando se utiliza esta regularización sí se puede obtener la derivada. Por tanto, sí son aplicables los métodos que hacen uso de la derivada, como el descenso del gradiente.

b. Regularización LASSO

La regularización LASSO (del inglés *Least Absolute Shrinkage and Selection Operator*), también llamada regularización L_1, consiste en escoger la norma L_1 como medida de complejidad; es decir,

$$\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_1,$$

donde $\|\boldsymbol{w}\|_1 = |w_1| + |w_2| + \cdots + |w_d|$, asumiendo $\boldsymbol{w} \in \mathbb{R}^d$. Al igual que en la regularización L_2, no se suele incluir el término w_0. El problema de regularización usando esta medida de complejidad se define como

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) + \lambda \|\boldsymbol{w}\|_1,$$

donde $\lambda \geq 0$ es el factor de regularización y tiene el mismo efecto que el descrito anteriormente. Equivalentemente, este problema de regularización se puede expresar como

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(\boldsymbol{x}^{(i)}), y^{(i)}) + \lambda \sum_{j=1}^{d} |w_j|,$$

donde se aplica de forma explícita la norma L_1 a cada uno de los parámetros w_i que definen la función de hipótesis.

De manera gráfica, la Figura 3 muestra el efecto de este tipo de regularización. Las elipses representan las curvas de nivel que se obtendría si solamente se considerase la función de costo del modelo sin incluir la regularización, y el punto azul indica el mínimo que se obtendría de esta manera. Por otro lado, el rombo muestra el efecto que tiene el restringir la norma L_1 a una región (es decir, es la grafica de $|w_1| + |w_2| = r$, con $r > 0$, considerando que el eje horizontal es w_1 y que el eje vertical es w_2). Con solo el término de regularización, el punto rojo sería el mínimo. Al combinar el efecto de ambas partes se obtiene el punto verde. Notar que este punto verde tiene una coordenada vertical diferente de cero, pero tiene una coordenada horizontal igual a cero. En general, para la regularización L_1, cuando λ es grande, el punto que brinda el mínimo se encuentra en los

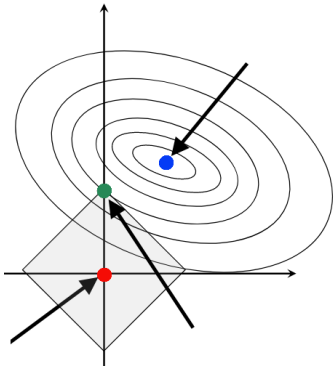


Figura 3: Efecto de la regularización L_1 (*lasso*).

vértices del rombo que define la norma. Por este motivo, algunos parámetros (w_j) serán exactamente iguales a cero, a lo que se conoce como dispersión (*sparsity*).

Debido a la propiedad de dispersión que presenta este tipo de regularización, en algunos casos no se llega a calcular todos los parámetros w_i ya que son cero. Esto puede significar una mejora en la memoria, especialmente cuando se trabaja con algoritmos que poseen muchos parámetros. Visto de otra manera, el hecho de eliminar algunos parámetros significa que los atributos asociados no son muy importantes para el modelo. Sin embargo, el problema que presenta esta regularización es que puede ser computacionalmente más costosa, ya que no se puede aplicar el algoritmo del descenso del gradiente de manera directa.

c. Regularización Elastic Net

Este tipo de regularización es la mezcla de las dos regularizaciones vistas anteriormente, quedando la medida de complejidad definida como

$$\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 + \|\boldsymbol{w}\|_2^2$$

Con esta función $\Omega(\boldsymbol{w})$, el problema de regularización se representa como la siguiente minimización:

$$\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(\boldsymbol{x}^{(i)}), y^{(i)}\right) + \lambda_1 \sum_{j=1}^{d} |w_j| + \lambda_2 \sum_{j=1}^{d} w_j^2,$$

donde λ_1 y λ_2 son factores que indican el peso que tendrá cada uno de los términos de regularización.

3. Regularización en Regresión Lineal

En la regresión lineal se puede aplicar los tipos de regularización descritos en la sección 2.2, utilizando la función de costo $J(w)$ definida para este modelo. Debido a que en este caso el uso de la norma L_2 es más frecuente, se describirá solamente este método a continuación.

3.1. Regresión Lineal Ridge

Al uso de la regularización de tipo *ridge* en la regresión lineal se le suele denominar *regresión ridge*. Utilizando la función de costo $J(\boldsymbol{w})$ para el modelo de regresión lineal en

el esquema general de la regularización L_2 dado en (1), se tiene que la función de costo regularizada en este caso es

$$J(\boldsymbol{w}) = \frac{1}{2n}\left(\sum_{i=1}^{n}(h_w(\boldsymbol{x}^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{d} w_j^2\right). \tag{2}$$

Sin embargo, a diferencia de (1), se ha utilizado el factor $\frac{1}{2n}$ en lugar del factor $\frac{1}{n}$ solo por conveniencia matemática. Igualmente, el factor de regularización es ahora, en realidad, $\frac{\lambda}{2n}$ a diferencia de λ en (1), pero esto no tiene mayor consecuencia debido a que sigue siendo un número que se escoge según el efecto que se desea que tenga la regularización. Debido a esto, se seguirá llamando factor de regularización a λ.

Para encontrar los parámetros que minimizan esta función de costo, es necesario determinar primero su derivada con respecto a cada uno de los parámetros w_i. La derivada del primer término de la función de costo (2), que describe el error del modelo, fue obtenida con anterioridad. La derivada del segundo término, que representa la regularización, se obtiene de forma directa como

$$\frac{\partial}{\partial w_i}\left(\frac{\lambda}{2n}\sum_{j=1}^{d} w_j^2\right) = \frac{\lambda}{n}w_i.$$

Una vez determinada la derivada de la función de costo, se puede utilizar el método del descenso del gradiente para encontrar los parámetros óptimos de la función de hipótesis. Debido a que w_0 en general no se regulariza, se tratará por separado. Así, la actualización de w_0 será

$$w_0 := w_0 - \alpha\frac{1}{n}\sum_{i=1}^{n}(h_w(\boldsymbol{x}^{(i)}) - y^{(i)}),$$

donde no aparece ningún efecto de regularización. Los demás parámetros sí tienen el efecto de regularización, por lo que la derivada anterior deberá ser tomada en cuenta. De este modo, la actualización de w_j, con $j = 1, \cdots, d$, está dada por

$$w_j := w_j - \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\left\{(h_w(\boldsymbol{x}^{(i)}) - y^{(i)})x_j^{(i)}\right\} + \frac{\lambda}{n}w_j\right).$$

Factorizando w_j en esta expresión se puede obtener la expresión equivalente

$$w_j := w_j\left(1 - \alpha\frac{\lambda}{n}\right) - \alpha\frac{1}{n}\sum_{i=1}^{n}(h_w(\boldsymbol{x}^{(i)}) - y^{(i)})x_j^{(i)},$$

donde se observa que, si se tiene bastantes instancias (n es bastante grande), el coeficiente de w_j será aproximadamente 1. Debido a esto, resulta común realizar la aproximación $w_j(1 - \alpha\frac{\lambda}{n}) \approx 0.99w_j$ cuando n es grande.

3.2. Solución Analítica de la Regresión Lineal Ridge

Al igual que con la regresión lineal sin regularización, es posible encontrar una solución analítica al caso de la regresión lineal ridge. Anteriormente se mostró que cuando se define la matriz $X \in \mathbb{R}^{n\times d}$ que contiene todas las n instancias $\boldsymbol{x}^{(i)}$ de entrada, y el vector $\boldsymbol{y} \in \mathbb{R}^n$ que contiene todas las $y^{(i)}$ salidas deseadas, la función de costo de la regresión lineal se

puede expresar como $\frac{1}{2n}\|X\boldsymbol{w}-\boldsymbol{y}\|^2$. Al añadir el efecto del término de regularización, esta función de costo se convierte en

$$J(\boldsymbol{w}) = \frac{1}{2n}\left(\|X\boldsymbol{w}-\boldsymbol{y}\|^2 + \lambda \boldsymbol{w}^T I_0 \boldsymbol{w}\right), \tag{3}$$

donde

$$I_0 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Con esta definición de I_0, y reemplazando término a término los productos indicados, se puede demostrar que $\sum_{j=1}^{d} w_j^2 = \boldsymbol{w}^T I_0 \boldsymbol{w}$, por lo que (3) es equivalente a (2).

La solución analítica de (3) se obtiene tomando la derivada con respecto a $\boldsymbol{w}$ e igualando a cero. La derivada del primer término se obtuvo anteriormente. La derivada del término de regularización es

$$\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{2n}\lambda \boldsymbol{w}^T I_0 \boldsymbol{w}\right) = \frac{\lambda}{n}\boldsymbol{w}^T I_0.$$

Utilizando estos términos, los parámetros óptimos se obtienen cuando la derivada de (3) se hace cero; es decir cuando

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{n}(\boldsymbol{w}^T X^T X - \boldsymbol{y}^T X) + \frac{\lambda}{n}\boldsymbol{w}^T I_0 = 0,$$

donde n es un factor que puede ser eliminado multiplicando a ambos lados por n, quedando $\boldsymbol{w}^T X^T X - \boldsymbol{y}^T X + \lambda \boldsymbol{w}^T I_0 = 0$. Debido a que se quiere despejar el valor de $\boldsymbol{w}$, se puede tomar la transpuesta de la expresión anterior, obteniendo $X^T X\boldsymbol{w} - X^T\boldsymbol{y} + \lambda I_0 \boldsymbol{w} = 0$. Si se factoriza $\boldsymbol{w}$ se obtiene $(X^T X + \lambda I_0)\boldsymbol{w} = X^T\boldsymbol{y}$, de donde resulta sencillo despejar $\boldsymbol{w}$ como

$$\boldsymbol{w} = (X^T X + \lambda I_0)^{-1} X^T \boldsymbol{y}.$$

A esta expresión se le denomina la *ecuación normal*, y es la solución analítica al problema cuando la regularización utiliza la norma L_2. Otros nombres que se suelen utilizar para esta solución son solución de *mínimos cuadrados amortiguados* o método de Levenberg-Marquardt. Notar que si $\lambda > 0$, el término $(X^T X + \lambda I_0)$ siempre será invertible.

4. Regularización en Regresión Logística

Al igual que para la regresión lineal, para la regresión logística también se puede aplicar cada uno de los tipos de regularización. Sin embargo, el método más usado es la regularización de norma L_2 o *ridge* y será el que se utilizará en adelante.

Al añadir el efecto de regularización $L-2$ a la función de costo de la regresión logística se obtiene la siguiente ecuación de costo

$$J(\boldsymbol{w}) = -\left(\frac{1}{n}\sum_{i=1}^{n}\left\{y^{(i)}\log(h_w(\boldsymbol{x}^{(i)})) + (1-y^{(i)})\log(1-h_w(\boldsymbol{x}^{(i)}))\right\}\right) + \frac{\lambda}{2n}\sum_{j=1}^{d} w_j^2.$$

Las derivadas de esta ecuación de costo son semejantes a las obtenidas para la regresión lineal, por lo que el descenso del gradiente de la sección 3.1 será igualmente aplicable.