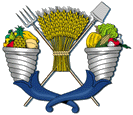
****

POSGRADO EN INGENIERÍA AGRÍCOLA

Y USO INTEGRAL DEL AGUA

UNIVERSIDAD AUTÓNOMA CHAPINGO

**MODELOS BASADOS EN APRENDIZAJE PROFUNDO**

**TAREA 7.**

* **Regularización L1**
* **Regularización L2**
* **Decaimiento de los pesos (Weight decay)**

**Presentan:**

***LÓPEZ VELASCO BALTAZAR***

**Catedrático:**

***DR. JUAN CARLOS OLGUIN ROJAS***

***Diagrama

Descripción generada automáticamente***

***Chapingo, México. Junio de 2025***

**INTRODUCCIÓN**

En el ámbito del aprendizaje automático y el modelado estadístico, uno de los desafíos recurrentes es el "sobreajuste" (overfitting), donde un modelo se adapta demasiado a los datos de entrenamiento, perdiendo su capacidad de generalizar correctamente a datos nuevos. Para contrarrestar este problema, surge la regularización como una técnica fundamental. La regularización introduce restricciones o penalizaciones adicionales durante el entrenamiento del modelo, con el propósito de controlar su complejidad y evitar que se aferre excesivamente a particularidades del conjunto de datos de entrenamiento. Las formas más comunes de regularización son la L1 (Lasso) y L2 (Ridge), cada una con un enfoque particular en la forma de penalizar los coeficientes del modelo.

**¿QUÉ ES LA REGULARIZACIÓN?**

La regularización es una técnica utilizada en el aprendizaje automático y el modelado estadístico para evitar el sobreajuste y mejorar la capacidad de generalización de los modelos. Cuando un modelo se sobre ajusta, ha aprendido demasiado bien los datos de entrenamiento y podría no funcionar correctamente con datos nuevos e inéditos (Figura 1).

La regularización introduce restricciones o penalizaciones adicionales al modelo durante el proceso de entrenamiento, con el objetivo de controlar su complejidad y evitar una dependencia excesiva de características o patrones específicos en los datos de entrenamiento. De esta forma, la regularización ayuda a lograr un equilibrio entre un buen ajuste a los datos de entrenamiento y su correcta generalización a nuevos datos.

Las técnicas de regularización más comunes son la regularización L1 (Lasso), la regularización L2 (Ridge) y la regularización de Red Elástica. La regularización L1 suma los valores absolutos de los coeficientes del modelo a la función de pérdida, lo que fomenta la dispersión y la selección de características. La regularización L2 suma los valores al cuadrado de los coeficientes del modelo, lo que permite coeficientes menores, pero distintos de cero. Finalmente, la regularización de Red Elástica combina las regularizaciones L1 y L2.

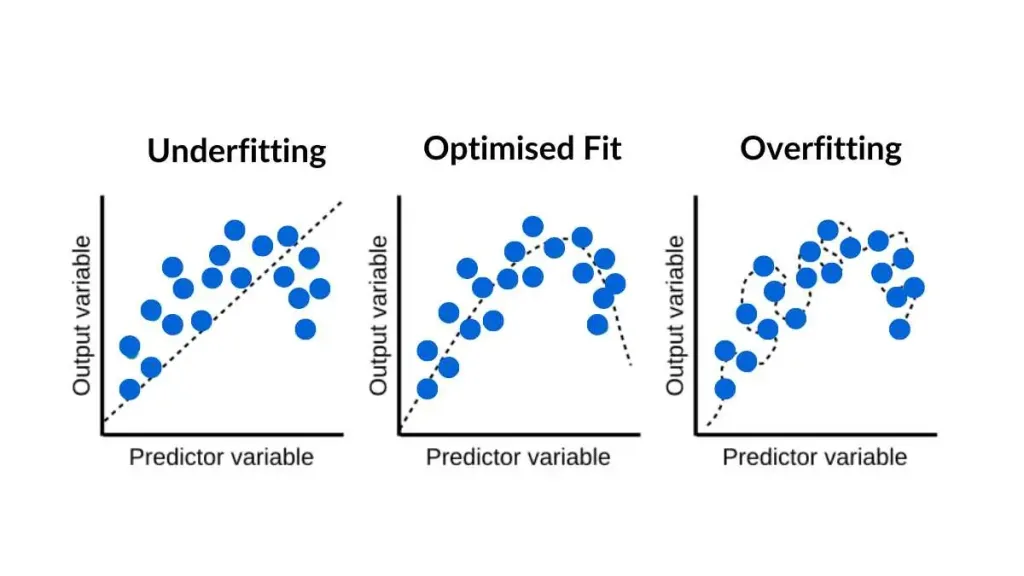


Figura. 1. Formas de ajuste de un modelo

**REGULARIZACIÓN L1 (LASSO REGRESSION)**

La regularización L1 agrega una penalización proporcional al valor absoluto de los coeficientes a la función de pérdida.

Función de pérdida (Loss Function):

(1)

donde es el error cuadrático medio, son los pesos del modelo, el parámetro de regularización y n es el número de características.

(2)

donde m es el número de muestras de entrenamiento, son los valores reales de los datos de entrenamiento y son los valores de salida previstos para el almacenamiento de datos.

**Para minimizar la función se hace;**

Penalización por Coeficientes Grandes:

* El término de penalización L1 aumenta a medida que los coeficientes crecen en valor absoluto.
* Esta penalización desalienta al modelo de permitir coeficientes grandes que podrían resultar de un sobreajuste al ruido o a características irrelevantes.

Fomenta la Esparsidad:

* La característica clave de Lasso es que la penalización L1 fomenta la esparsidad en el modelo. A diferencia de la regresión Ridge (L2), que reduce los coeficientes, pero no los establece en cero, Lasso puede llevar algunos coeficientes a cero exacto, eliminando efectivamente esas características del modelo.
* Esto significa que Lasso realiza una selección automática de características, manteniendo solo las características más significativas.

Efecto del Parámetro de Regularización λ:

* Cuando λ es grande: El término de penalización domina, y más coeficientes se llevan a cero. El modelo se vuelve más regularizado, y muchas características se descartan (lo que lleva a un modelo más simple). Sin embargo, si λ es demasiado grande, el modelo puede subajustarse.
* Cuando λ es pequeño: El término de penalización es débil, y el modelo se comporta de manera similar a la regresión lineal ordinaria. Los coeficientes pueden crecer más, y el modelo puede ajustarse a los datos de cerca (lo que podría llevar a un sobreajuste).

**¿Qué efecto tiene?**

* Fomenta la escasez en el modelo (algunos coeficientes se vuelven exactamente cero), realizando de manera efectiva la selección de características.
* Si una característica tiene un coeficiente pequeño, la regularización L1 tiende a reducirlo a cero, volviéndolo irrelevante para el modelo.

Para minimizar la función de pérdida de Lasso utilizando el descenso de gradiente, calculamos el gradiente de la función de pérdida con respecto a los coeficientes w:

Gradiente del MSE: El gradiente del término MSE con respecto a es:

(3)

* Este gradiente guía las actualizaciones de los coeficientes basándose en qué tan bien el modelo se ajusta a los datos.

Gradiente de la Penalización L1: El gradiente del término de penalización L1 con respecto es:

(4)

* Aquí, da el signo (ya sea +1 o -1), lo que hace que la actualización del gradiente se comporte de manera diferente para coeficientes positivos y negativos.

Gradiente Total: El gradiente total para cada coeficiente es la suma del gradiente del MSE y el gradiente de la penalización L1:

(5)

La regla de actualización de descenso de gradiente para los coeficientes es:

**¿Cuándo utilizar la regularización L1?**

* Selección de características: utilice la regularización L1 si cree que solo unas pocas características contribuyen significativamente al modelo y desea seleccionar características relevantes automáticamente.
* Modelos dispersos: cuando desea un modelo que dé como resultado menos coeficientes distintos de cero.

(6)

Impacto de la regresión Lasso en los coeficientes

λ grande: El término de penalización se vuelve más significativo y los coeficientes se reducen drásticamente, con muchos coeficientes llevados exactamente a cero. Esto resulta en un modelo más simple con menos características.

λ pequeño: el término de penalización se debilita y se permite que los coeficientes crezcan, lo que lleva al modelo a ajustarse mejor a los datos, pero con el riesgo de sobreajuste.

**REGULARIZACIÓN L2 (RIDGE REGRESSION)**

La regularización L2 agrega una penalización proporcional al valor cuadrado de los coeficientes a la función de pérdida.

Función de pérdida (Loss Function):

(7)

donde es el error cuadrático medio, son los pesos del modelo, el parámetro de regularización y n es el número de características.

La regresión de cresta reduce los coeficientes hacia cero sin establecerlos exactamente en cero. Esto se logra mediante la penalización L2, que es la suma de los cuadrados de los coeficientes. La intuición detrás del proceso de reducción es la siguiente:

Penalización por Coeficientes Grandes:

El termino añade una penalización proporcional al cuadrado de los coeficientes. Si un coeficiente se vuelve muy grande, la penalización aumenta. Esto anima al modelo a reducir los coeficientes grandes para minimizar la función de pérdida total.

Minimización de la Función de Pérdida:

El modelo intenta minimizar tanto el MSE como la penalización L2. Como resultado, aunque el modelo quiere ajustar bien los datos (minimizar el MSE), se ve forzado a mantener también los coeficientes pequeños, lo que ayuda a reducir el sobreajuste.

Impacto del Parámetro de Regularización λ:

Cuando λ es grande: El término de regularización se vuelve más significativo, y el modelo reduce más los coeficientes. Esto lleva a coeficientes más pequeños y modelos más regularizados.

Cuando λ es pequeño: El modelo se comporta más como una regresión lineal ordinaria, con menos regularización. Se permite que los coeficientes crezcan más para ajustar mejor los datos.

Para minimizar la función de pérdida utilizando el descenso de gradiente, calculamos el gradiente de la función de pérdida con respecto a los coeficientes w:

Gradiente del MSE: El gradiente del término MSE con respecto a los coeficientes es:

(8)

* Este gradiente guía las actualizaciones de los coeficientes basándose en qué tan bien el modelo se ajusta a los datos.

Gradiente de la Penalización L2: El gradiente del término de penalización L2 con respecto a es:

(9)

* Este término hace que los coeficientes se reduzcan hacia cero con cada actualización.

Gradiente Total: La actualización del gradiente total para cada coeficiente es una combinación del gradiente del MSE y el gradiente de la penalización L2:

(10)

La regla de actualización de descenso de gradiente para los coeficientes es:

(11)

donde η es la tasa de aprendizaje.

La conclusión clave es que la actualización del descenso del gradiente implica tanto reducir el término de error (MSE) como reducir los coeficientes debido a la penalización L2.

**¿Qué efecto tiene?**

* Reduce los coeficientes: La regularización L2 fomenta coeficientes más pequeños, pero no los reduce a cero. Reduce la complejidad del modelo sin eliminar ninguna característica por completo.
* Evita coeficientes grandes: ayuda a evitar el sobreajuste al penalizar pesos grandes, que pueden resultar de características ruidosas o irrelevantes.

**¿Cuándo utilizar la regularización L2?**

* Multicolinealidad: L2 es útil cuando hay características altamente correlacionadas (multicolinealidad), ya que evita que surjan coeficientes grandes de dichas correlaciones.
* Evitar el sobreajuste: utilice L2 cuando desee reducir los coeficientes y reducir el sobreajuste sin descartar por completo ninguna característica.

**Ventajas y desventajas de la regularización L1 y la regularización L2**

|  |  |  |
| --- | --- | --- |
| **Técnica de regularización** | **Ventajas** | **Desventajas** |
| Regularización L1 (Lasso) | * Realiza la selección de características, llevando algunos coeficientes a cero. * Ayuda a gestionar conjuntos de datos de alta dimensión. * Puede manejar funciones irrelevantes o menos importantes. * Útil para construir modelos dispersos | * Puede generar una alta escasez, lo que hace que el modelo sea menos interpretable. * No es eficaz cuando existen fuertes correlaciones entre las características. * Computacionalmente más costoso que la regularización L2. |
| Regularización L2 (Ridge) | * Ayuda a prevenir el sobreajuste y mejorar la generalización. * Eficaz cuando existen fuertes correlaciones entre las características. * Calcula soluciones estables. * Computacionalmente eficiente. | * No realiza selección de características como la regularización L1. * El modelo resultante aún puede contener muchos coeficientes pequeños distintos de cero. * Puede no ser adecuado para conjuntos de datos de alta dimensión |

**Interpretación geométrica de la regresión de Lasso y Ridge**

En la Regresión de Ridge (L2), los coeficientes están restringidos por una región esférica en el espacio de parámetros, donde la suma de sus cuadrados es limitada. Esto resulta en coeficientes decrecientes, pero ninguno es exactamente cero, manteniendo así todas las características del modelo. Mientras que, en la regresión Lasso (L1) , la restricción forma una región en forma de diamante, donde la suma de los valores absolutos de los coeficientes es limitada. Esta forma geométrica propicia que algunos coeficientes sean exactamente cero, lo que permite una selección de características eficaz al eliminar las irrelevantes.

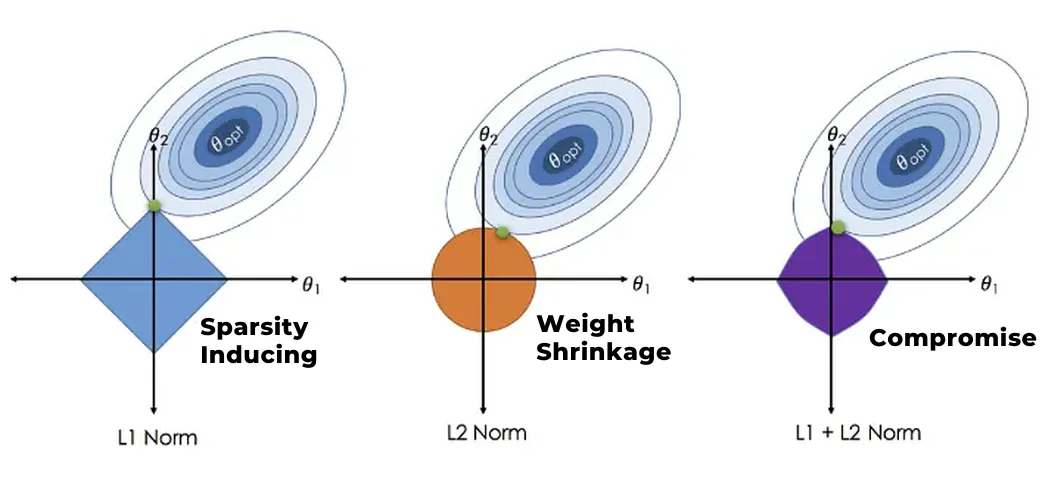


Figura 2. Interpretación gráfica de la regularización L1 y la regularización L2.

**Regularización L1 y L2 en el aprendizaje profundo**

En el aprendizaje profundo, la regularización L1 y L2 se incorporan típicamente al proceso de entrenamiento añadiendo sus términos de penalización correspondientes a la función de pérdida. Los términos de regularización se multiplican por un parámetro de regularización (λ) para controlar la fuerza de la regularización

En la práctica, un enfoque común consiste en utilizar una combinación de regularización L1 y L2, conocida como regularización de red elástica. Esto equilibra la selección de características (regularización L1) y la reducción de peso (regularización L2).

En el aprendizaje profundo, la elección entre la regularización L1 y L2 (o su combinación) depende del problema específico, las características de los datos y el comportamiento deseado del modelo. A menudo se requiere experimentación y ajuste de los parámetros de regularización para lograr los mejores resultados.

**WEIGHT DECAY**

El decaimiento de los pesos o la regularización L2 es una técnica de regularización aplicada a los pesos de una red neuronal. Minimizamos una función de pérdida que compromete tanto la función de pérdida principal como una penalización en la L2, Norma de los pesos:

dónde es un valor que determina la fuerza de la penalización (fomentando pesos menores).

La disminución de peso puede incorporarse directamente en la regla de actualización de peso, en lugar de definirse implícitamente mediante la función objetivo. A menudo, la disminución de peso se refiere a la implementación donde se especifica directamente en la regla de actualización de peso (mientras que la regularización L2 suele ser la implementación que se especifica en la función objetivo).

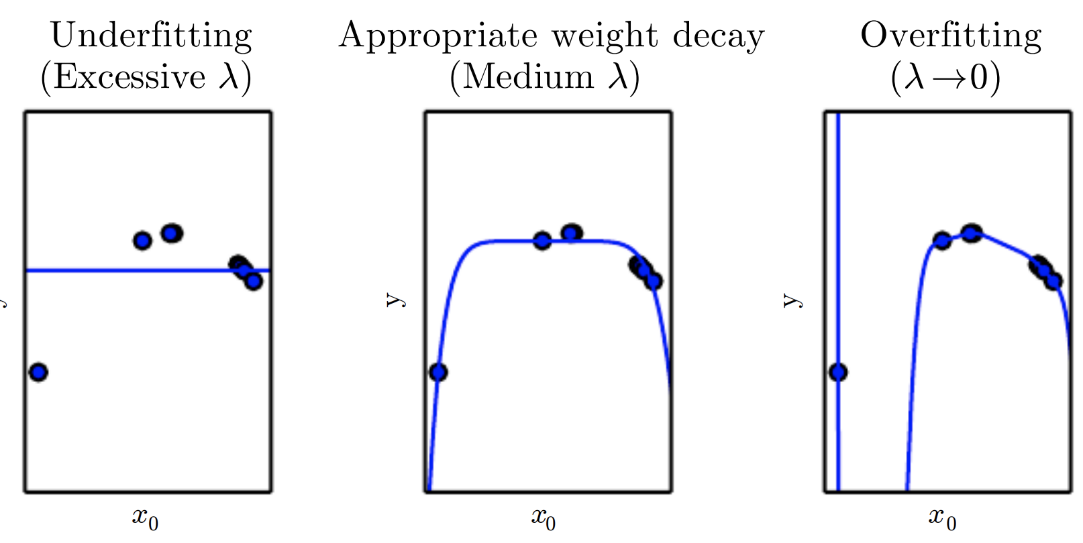


Figura 3. Efecto del parámetro de regularización (λ) en el ajuste del modelo.

CONCLUSIÓN

La regularización es una herramienta para reducir el sobreajuste y mejorar la capacidad de generalización de los modelos. Tanto la regularización L1 (Lasso) como la L2 (Ridge), son mecanismos efectivos para controlar la complejidad del modelo a través de la penalización de sus coeficientes. Mientras que Lasso se distingue por su capacidad de realizar selección de características al llevar algunos coeficientes a cero, Ridge se enfoca en reducir los coeficientes sin eliminarlos por completo, siendo especialmente útil en presencia de multicolinealidad. La elección de la técnica de regularización, y el ajuste de su parámetro λ, dependerá de las características específicas del problema y de los datos, buscando siempre construir modelos que no solo se ajusten bien a la información conocida, sino que también sean fiables y precisos ante condiciones desconocidas.