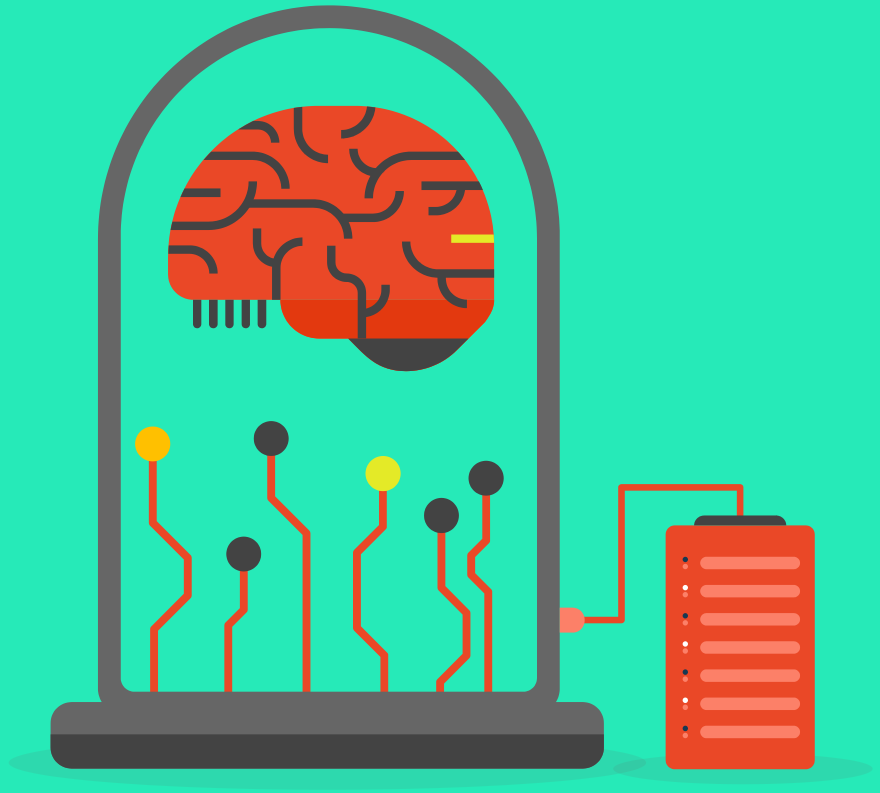
3. Regularización en Modelos Lineales



Regularización

Como vimos en el tema anterior, a medida que un modelo aumenta en su complejidad, más probabilidades tendrá de sobreentrenar los datos y, por lo tanto, peor será su capacidad de generalización. En el caso del modelo de regresión polinómica, vimos que una manera de evitar el sobreentrenamiento es reducir el grado del polinomio. Es decir, establecer restricciones en los coeficientes del modelo.

A las técnicas empleadas en el Machine Learning para restringir los parámetros (o pesos) y evitar el sobreentrenamiento de los modelos se le conoce como **Regularización**.

Para el caso de los modelos lineales, estudiaremos entonces las regresiones de **Ridge**, **Lasso** y **Elastic Net**. Todos ellos, modelos de regresión lineal regularizados.

Regresión Ridge

En el caso de la Regresión Ridge, a fin de regularizar el modelo lineal original, se agrega un término a la función de costo que depende del cuadrado de los coeficientes. Recordemos que la función de costo del problema de regresión lineal es:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

Que podemos escribir como:

$$J(\theta) = MSE(\theta)$$

Regresión Ridge

Entonces, la regularización vendrá dada por:

$$J(\theta) = MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$$

En donde α es un hiperparámetro que controla cuánto se desea regularizar el modelo. Si dicha constante es igual a cero, se obtiene la regresión lineal original. Agregar este término hará que, al optimizarlo, se garantice que no solo el error sea cercano a cero, sino también los valores de cada coeficiente.

Al término de regularización de la Regresión Ridge también se le conoce como norma l_2.

Regresión Lasso

La Regresión Lasso (Least Absolute Shrinkage and Selection Operator) es otra versión de regularización de la regresión lineal, en donde se agrega a la función de costo un término que depende del valor absoluto de los coeficientes:

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

A este término también se le conoce como norma l_1. A diferencia de la regresión Ridge, la regresión Lasso tiende a eliminar por completo (hacer cero) los coeficientes de las variables menos importantes en el modelo, lo que genera modelos *sparse* (modelos con pocos coeficientes no iguales a cero).

Regresión Elastic Net

La Regresión Elastic Net representa un punto medio entre Ridge y Lasso, en donde el término de regularización agregado a la función de costo es una combinación de las anteriores:

$$J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2} \alpha \sum_{i=1}^{n} \theta_i^2$$

En donde la constante r regula la mezcla entre una regularización y la otra.

Regularización

El cuándo utilizar cada caso dependerá del problema. De entrada, el usar un modelo con regularización siempre ayudará a evitar el sobreentrenamiento, de modo que la Regresión Ridge es un buen punto de partida o línea base. Si se sospecha que puedan existir variables que no sean buenas predictoras del problema, se puede aplicar la Regresión Lasso pues esta eliminará las variables no importantes. Por otro lado, si existen altas correlaciones entre variables, se recomienda usar Elastic Net y ensayar con la constante r a fin de encontrar el mejor valor.

Sin embargo, siempre que se trabaja con regularización, es importante aplicar a los datos lo que se conoce como **Escalado de Variables** (*Feature Scaling*), pues las técnicas de regularización pueden ser sensibles a las escalas de las variables.

Veamos qué significa esto.

Escalado de Variables (Feature Scaling)

En el Machine Learning, resulta común trabajar con conjuntos de datos que pueden provenir de fuentes muy variadas o que pueden representar características o variables (features) de naturalezas muy distintas o, en el caso de las variables numéricas, de **escalas** muy distintas. Esto quiere decir que, por ejemplo, el rango de valores de una variable podría estar entre [0, 1] mientras que el de otra entre [1.000, 10.000] o incluso mayores.

Cuando esto ocurre, es posible que un modelo de regresión (u otros) favorezca a estas últimas variables y les de un mayor peso, solo porque sus rangos de valores son mayores (aún cuando quizá la otra variable está más correlacionada con la salida esperada).

Por ello, cuando se construyen modelos de Machine Learning es recomendable realizar previamente un escalado de las variables a fin de disminuir los efectos descritos.

Escalado de Variables (Feature Scaling)

Los dos tipos de escalado de variables más comunes son la **Normalización** (o *MinMax Scaling*) y la **Estandarización** (o *Standard Scaling*).

La **Normalización** consiste en tomar cada valor de una variable, restarle el mínimo y dividirla entre el máximo menos el mínimo, de tal manera que el rango de valores de la nueva variable siempre estará entre [0, 1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Escalado de Variables (Feature Scaling)

La **Estandarización** consiste en tomar cada valor de una variable, restarle la media y dividirla entre la desviación estándar:

$$x' = \frac{x - \mu}{\sigma}$$

A diferencia de la normalización, la estandarización no restringe las variables a un rango fijo, sino que lleva a la variable a tener una distribución con media cero y varianza igual a la unidad.

Es importante resaltar que cuando se aplica el escalado de variables, se hace solo sobre el conjunto de entrenamiento. Luego, las variables asociadas a la normalización o estandarización se tienen que considerar al momento de realizar predicciones nuevas (veremos esto en la práctica).

4to Notebook Práctico

El escalado de variables garantizará, en la mayoría de los casos, que no se produzcan inestabilidades en la búsqueda de los mejores coeficientes del modelo, o bien que algunas variables tengan predominancia sobre otras debido a sus escalas.

De igual modo, el escalado de variables es una técnica de uso casi obligatorio cuando se realiza no solo Machine Learning, sino Deep Learning también.

Veamos, entonces, cómo se implementa el escalado de variables, así como las técnicas de regularización estudiadas.

¡Adelante con el Notebook!