Regresión y Regularización

Hay veces en las que queremos que nuestro modelo no se ajuste tan bien a los datos de entrenamiento para evitar el overfitting. Por ejemplo, una forma de hacer esto en una regresión polinomial es bajar el grado del polinomio, ya que "reducimos los grados de libertad" del modelo.

Ahora bien, para un modelo de regresión lineal, esto se hace limitando el peso de los coeficientes. Vamos a revisar tres técnicas para regularizar una regresión lineal: esto es, castigar al modelo en la fase de entrenamiento para prevenir el overfitting.

Ridge Regression

Ridge Regression fuerza a que el modelo intente mantener los pesos lo más bajo posible. Esto se logra cambiando la función objetivo:

$$\underbrace{MSE(\beta)}_{\text{Parte habitual en la regresión lineal}} + \underbrace{\alpha \frac{1}{2} \sum_{i=1}^{m} \beta_i^2}_{\text{Término de la regularización}}$$

El hiperparámetro α controla cuanto pesa la regularización. Si α es grande, va a forzar que todos los pesos sean cercanos a 0, lo que generaría una línea "sin pendiente". Notemos además que el término β_0 (coeficiente de posición) no está regularizado.

Lasso Regression

Esta técnica es similar a Ridge, pero la función objetivo en este caso es:

$$MSE(\beta) + \alpha \sum_{i=1}^{n} |\beta_i|$$

Una característica importante de Lasso es que tiende a "eliminar" el peso de

las variables poco importantes.

Elastic Net

Elastic Net es una mezcla entre Ridge y Lasso. Introducimos un hiperparámetro r que va entre 0 y 1, que "mezcla" en cierta proporción ambas regularizaciones.

$$MSE(\beta) + r\left(\alpha \sum_{i=1}^{n} |\beta_i|\right) + 1-r\left(\frac{1}{2}\alpha \sum_{i=1}^{n} \beta_i^2\right)$$

Lasso Ridge

Algunos detalles

Tenemos que discutir dos detalles importantes. El primero es identificar cuando usar una regresión simple vs Ridge vs Lasso vs Elastic Net.

En general, en la literatura, nunca se recomienda usar regresión simple. En general Ridge suele ser una buena decisión

por defecto, pero si sospechamos que hay features que no aportan, usamos Lasso o Elastic Net.

El segundo punto es que al penalizar por el valor de los β_i, estas técnicas son sensibles a las escalas de las features. Por ejemplo, si hay una columna que se mueve entre 0 y 100 y otra entre 0 y 10.000.000. Esto porque la escala de los datos influye en la magnitud del coeficiente. En general, antes de regularizar se estandarize cada columna, restando la media y dividiendo por la desviación estándar.