El Problema de la Generalización

Luis Zúñiga

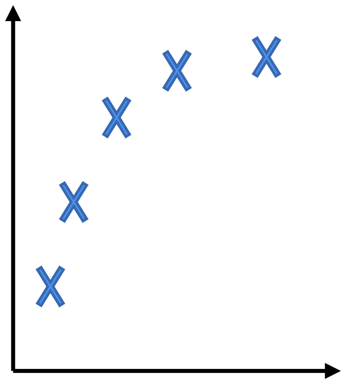
Agenda

1. Overfitting
2. Regularización
3. Idea básica de la Regularización
4. Regularización en Regresión Lineal

Introducción

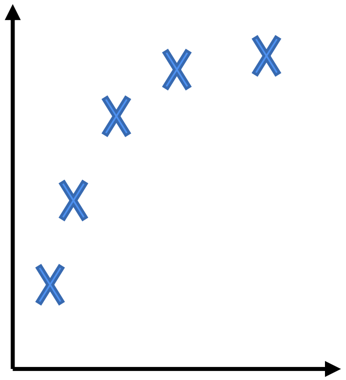
- Hasta el momento, trabajamos la idea de ajustar datos a un modelo matemático (regresión lineal, polinomial, logística).
- El fin principal es que, si hay un nuevo dato x_i, el modelo debe ser capaz de **predecir su valor o clase** de manera correcta.
- En pocas palabras, debe ser capaz de **generalizar**.

Overfitting



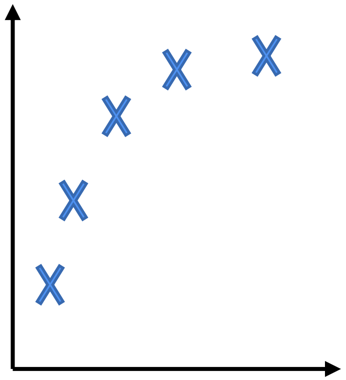
$\theta_0 + \theta_1 x$

Underfitting
(high bias)



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

Just Right



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

Overfitting
(high variance)

Overfitting

Sobreajuste (overfitting)

Para los puntos del conjunto de datos D, el algoritmo de aprendizaje se ajusta a los datos casi de manera perfecta

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 \approx 0$$

pero el modelo falla al generalizar nuevas instancias (aquellos con los que no se entrenó), i.e., no predice de manera correcta nuevos elementos. Esto se traduce en **bajo error en el modelo, gran error al predecir nuevos datos**.

Underfitting

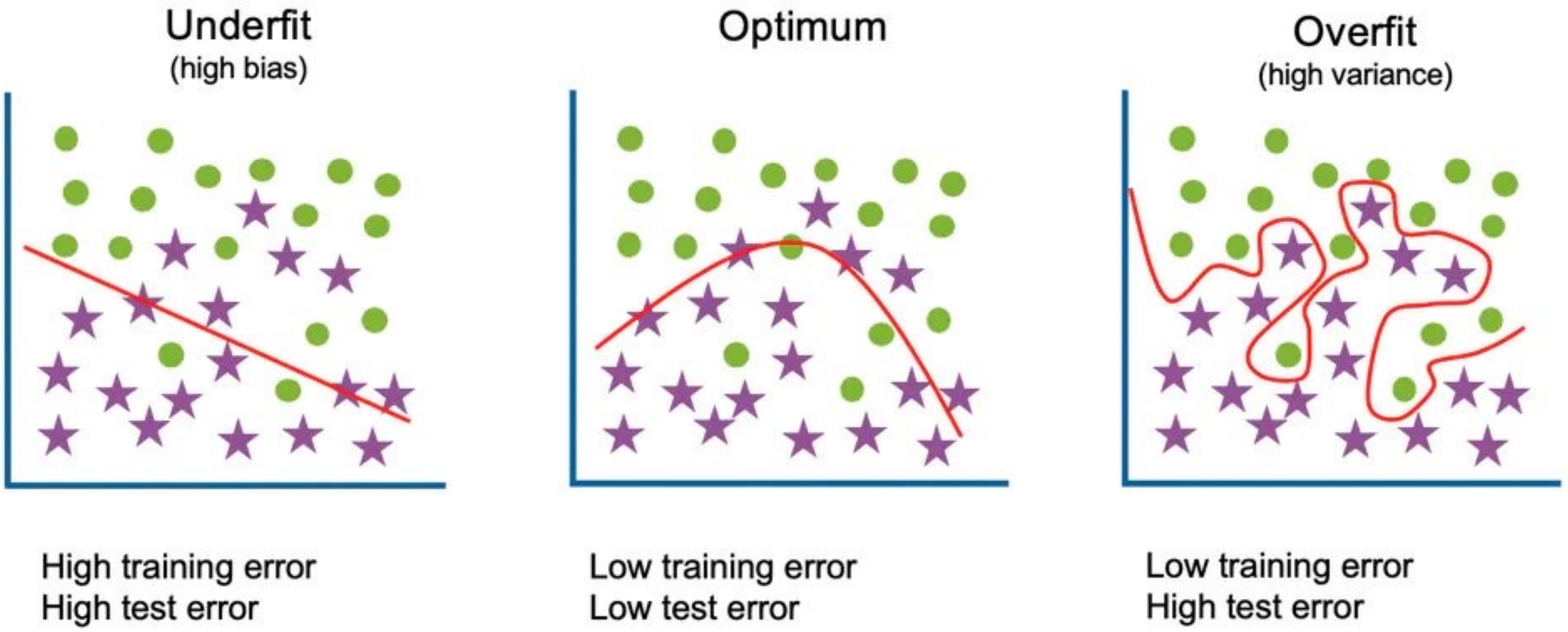
Desajuste (underfitting)

Para los puntos del conjunto de datos D, el algoritmo de aprendizaje no logra ajustarse a los datos

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 > 0$$

y falla al generalizar nuevas instancias (aquellos con los que no se entrenó), i.e., no predice de manera correcta nuevos elementos. Esto se traduce en **gran error en el modelo, gran error al predecir nuevos datos**.

Overfitting



Fuente: IBM

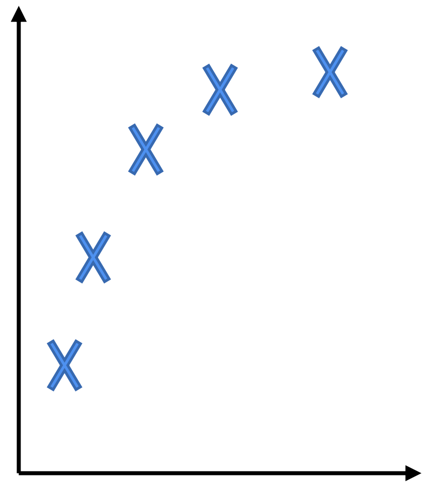
¿Cómo evitar el sobreajuste?

1. Reducir el número de características
    1. De manera manual
    2. Con algoritmos
2. Regularización
    1. Reducen el impacto o magnitud de los parámetros del modelo.
    2. Permite conservar características.
    3. No es un método único, cambia dependiendo del algoritmo de aprendizaje.

La idea de la Regularización

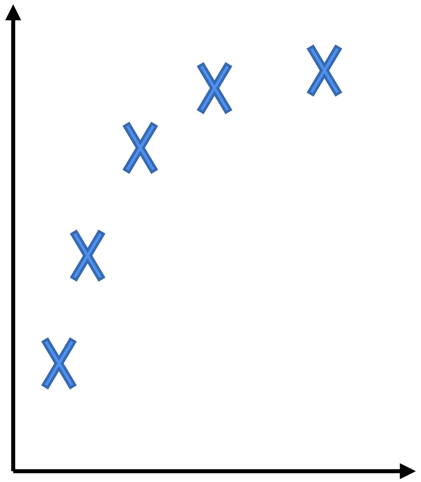
La idea principal de regularizar un modelo es ajustar los parámetros para evitar funciones muy complejas que no permitan generalizar el aprendizaje.

Consideremos el modelo de regresión que sobreajusta los datos.



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

La idea de la Regularización



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

Es posible que tanto el término cúbico como el cuártico afecten el modelo. Hasta el término cuadrático parece ser suficiente.

Vamos a disminuir la contribución de x^3 y x^4. ¿Cómo se les ocurre hacerlo?

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

La idea de la Regularización

Regularización

- Ajustar los parámetros θ_i para que la hipótesis sea más «simple».
- Menos propensa al sobreajuste.

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

- El parámetro λ controlará la contribución o magnitud de cada parámetro del modelo. Noten que no se afecta θ_0.

La idea de la Regularización

¿Qué pasaría con los parámetros θ_i y, en consecuencia, con la hipótesis $h_{\boldsymbol{\theta}}$ si $\lambda = 10^{10}$?

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

Regularización en Regresión Lineal

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

Gradiente Descendiente

$$\theta_0 = \theta_0 - \eta\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}_0^{(i)}$$

$$\theta_j = \theta_j - \eta\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}_j^{(i)} + \frac{\lambda}{n}\theta_j$$

Actividad: Demostrar estos resultados.

Regularización en Regresión Lineal

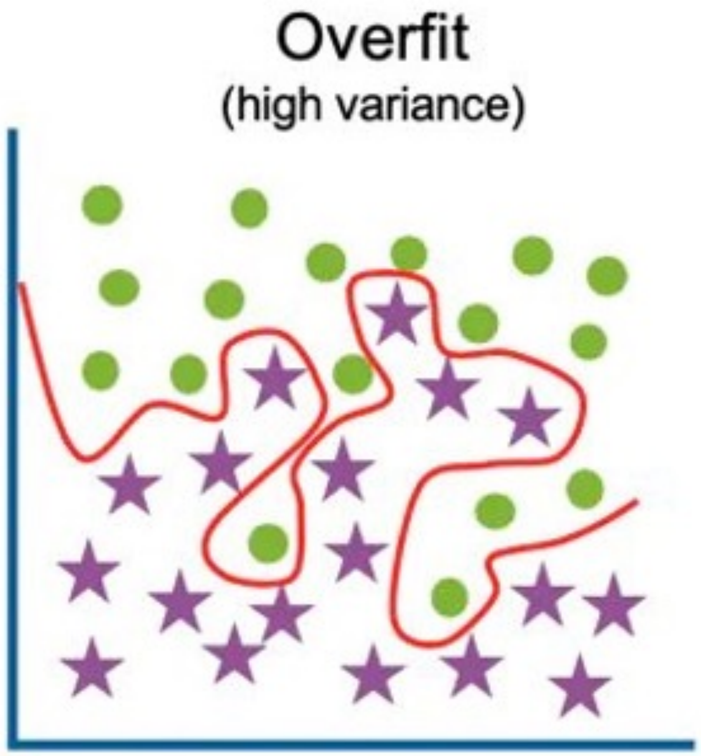
Tarea: Demostrar que ambas expresiones son equivalentes.

$$\theta_0 = \theta_0 - \eta \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right) \boldsymbol{x}_0^{(i)}$$

$$\theta_j = \theta_j - \eta \left[\frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right) \boldsymbol{x}_j^{(i)} + \frac{\lambda}{n} \theta_j \right]$$

$$\theta_j = \theta_j \left(1 - \eta \frac{\lambda}{n} \right) - \eta \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right) \boldsymbol{x}_j^{(i)}$$

Regularización en Regresión Logística



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

Función de Costo

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

Regularización en Regresión Logística

Función de Costo

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

$$\theta_0 = \theta_0 - \eta \frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}_0^{(i)}$$

$$\theta_j = \theta_j - \eta \left[\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}_j^{(i)} + \frac{\lambda}{n}\theta_j\right]$$

Tarea

- Investigar en qué consiste la regresión Lasso.
  - Mencionar cómo se modifica el modelo de regresión.
  - Determinar las reglas de actualización de gradiente descendiente bajo esta idea.

Gracias

Luis Zúñiga

p40887@correo.uia.mx

Sitio web