Overfitting, Regularización y Validación

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

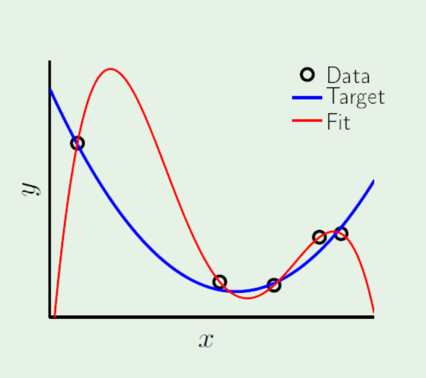
caicedo77@gmail.com

28 de octubre de 2016

Material adaptado a partir del curso de Caltech "Learning from Data", por Yaser S. Abu Mustafá.

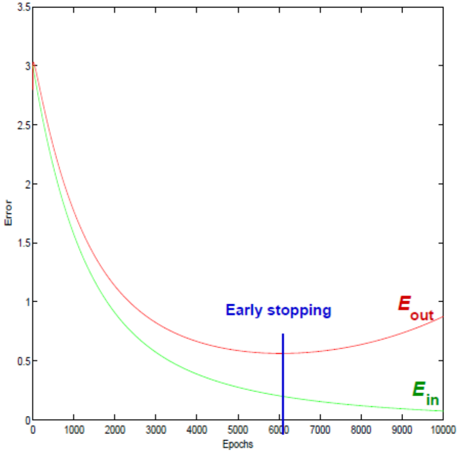
Overfitting: Un ejemplo ilustrativo

- Tenemos una función blanco sencilla.
- Solo tenemos 5 puntos de entrenamiento corruptos por ruido.
- Vamos a usar un modelo de regresión polinomial de 4to grado.
- $E_{in} = 0$, E_{out} elevado



Overfitting y mala generalización

- Ejemplo: Una Red Neuronal siendo entrenada con datos ruidosos.
- Overfitting: $E_{in} \downarrow, E_{out} \uparrow$

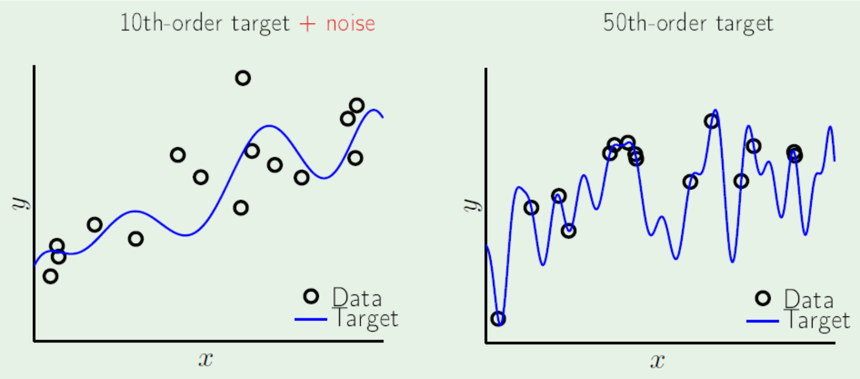


Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

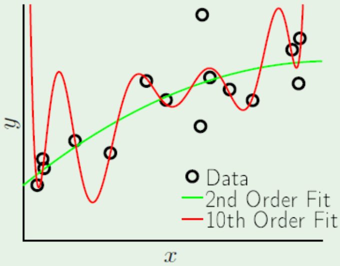
Quién tiene la culpa?

Overfitting: "Aprender de los datos más allá de lo justificado".
Culpable: Aprender del ruido: Nocivo!

Un caso de estudio

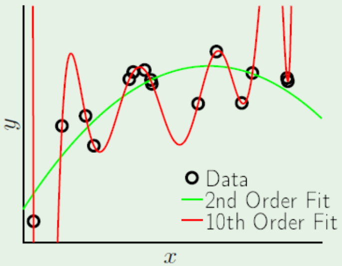


Dos hipótesis para cada blanco



Noisy low-order target

	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	9.00



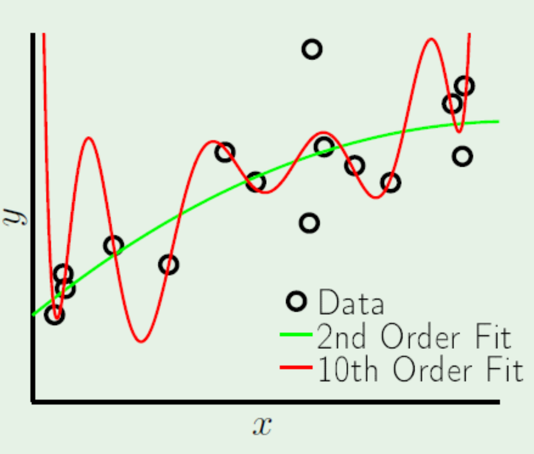
Noiseless high-order target

	2nd Order	10th Order
E_{in}	0.029	10^{-5}
E_{out}	0.120	7680

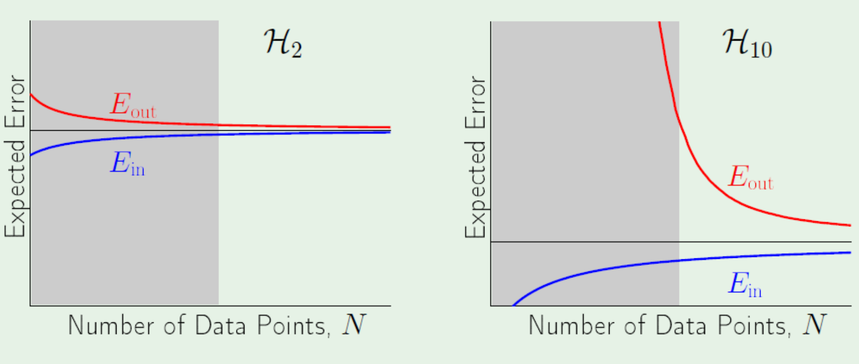
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

La ironía

- Tenemos 2 algoritmos de aprendizaje, O y R.
- Sabemos que la función blanco es de grado 10
- O escoge $\mathcal{H}_{10}$ y R $\mathcal{H}_2$



Curvas de Aprendizaje

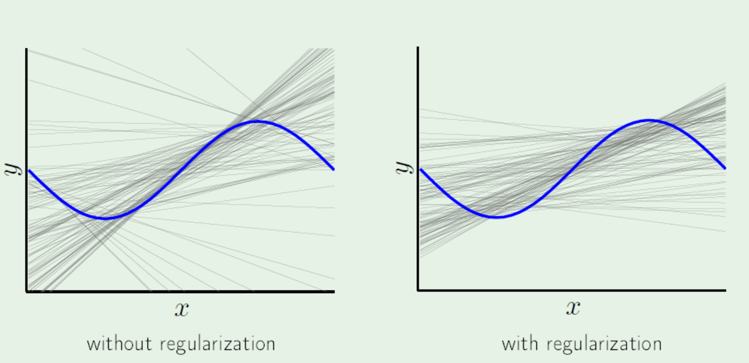


Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

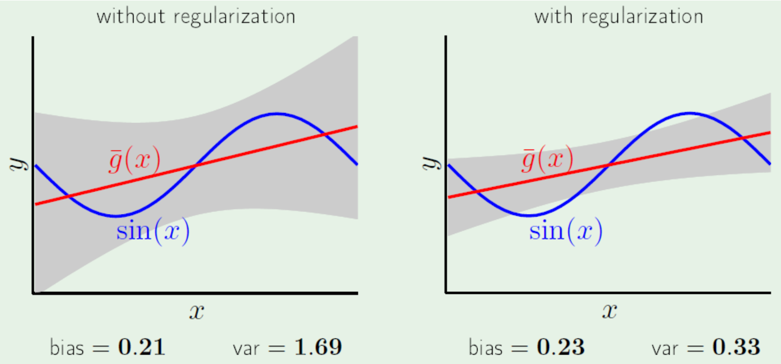
Dos armas contra el overfitting

- **Regularización:** Aplicar los frenos y usar la menor complejidad posible.
- **Validación:** Escoger adecuadamente entre varios modelos sin contaminar nuestro conjunto de pruebas.

Regularización: Un ejemplo de aplicación



Regularización: Un ejemplo de aplicación



Qué logramos con la Regularización

- Con la regularización le ponemos un freno a la varianza del conjunto de hipótesis utilizado.
- Le damos al algoritmo de aprendizaje un incentivo para preferir hipótesis más simples.
- Cómo? Introduciendo una penalidad adicional en función de la magnitud de los pesos utilizados.
- Por lo tanto el error de entrenamiento regularizado obedece a la siguiente estructura general:

$$E_{aug} = E_{in} + \frac{\lambda}{N}\theta^T\theta$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- Ejemplo: Recordemos la función de costo en la Regresión Lineal:

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n)^2$$

- La versión regularizada es:

$$J(\theta) = \frac{1}{N}\left[\sum_{n=1}^{N}(h(x_n) - y_n)^2 + \lambda \sum_{m=1}^{M} \theta^2\right]$$

- Vectorialmente:

$$J(\theta) = \frac{1}{N}\left[(X\theta - y)^T(X\theta - y) + \lambda\theta^T\theta\right]$$

- El parámetro lambda controla que tanta importancia dentro del costo total se le asigna a la penalización por la complejidad de la hipótesis seleccionada.

- La solución del problema de aprendizaje sería la siguiente:

$$\theta = (X^T X)^{-1} X^T y \text{ (Sin regularización)}$$

$$\theta = (X^T X + \lambda I)^{-1} X^T y \text{ (Con regularización)}$$

Regresión Logística Regularizada

- La función de costo a minimizar sería la siguiente:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i(\log h(x_i)) + (1-y_i)log(1-h(x_i)) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

- El gradiente de la función de costo ahora sería:

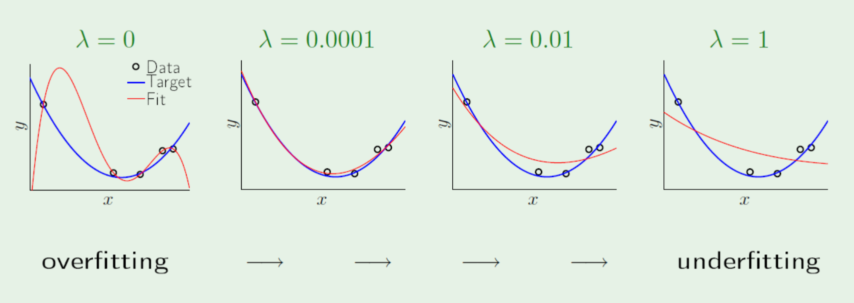
$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij} - \frac{\lambda}{m}\theta_j$$

- Y la regla de actualización del Gradiente Descendente quedaría así:

$$\theta j = \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)x_{ij} - \frac{\lambda}{m}\theta_j\right]$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Figura: Minimizando $E_{in}(\theta) + \frac{\lambda}{N}\theta^T\theta$ con diferentes valores de λ



- Si llamamos a nuestro regularizador $\Omega = \Omega(h)$, nuestro error a minimizar sería:

$$E_{aug} = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

- Si llamamos a nuestro regularizador $\Omega = \Omega(h)$, nuestro error a minimizar sería:

$$E_{aug} = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

- A quien se les parece esa expresión para el error?

La Regularización y la cota VC

- Si llamamos a nuestro regularizador $\Omega = \Omega(h)$, nuestro error a minimizar sería:

$$E_{aug} = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

- A quien se les parece esa expresión para el error?

$$E_{out}(h) \leq E_{in}(h) + \Omega(\mathcal{H})$$

- La medida de error regularizada es un mejor estimativo del valor esperado del error fuera del entrenamiento!

El valor óptimo para λ

- De acuerdo a lo que tenemos ahora, λ se convierte en un parámetro adicional de nuestro modelo de aprendizaje.
- Cómo hallamos el valor justo? Validación.

- Si utilizamos el conjunto de datos de prueba para probar diferentes modelos y/o establecer un valor adecuado para λ, E_{test} ya no será un estimador sin sesgo del verdadero E_{out}
- Entonces, cómo podemos hacer nuestra selección sin comprometer el conjunto de pruebas?
- La respuesta: Destinar un conjunto de datos específicamente para este fin: El conjunto de validación.
- La idea es poder tener una estimación del error fuera del entrenamiento sin comprometer el conjunto de pruebas, que debe ser reservado hasta el final.

Validación vs Regularización

- En nuestra discusión sobre el overfitting y la regularización, hemos llegado a la conclusión de que:

$$E_{out}(h) = E_{in}(h) + \text{penalidad por overfitting}$$

- Qué hace la regularización?

$$E_{out}(h) = E_{in}(h) + \underbrace{\text{penalidad por overfitting}}_{\text{La regularización estima esta cantidad}}$$

- Qué hace la validación?

$$\underbrace{E_{out}(h)}_{\text{La validación estima esta cantidad}} = E_{in}(h) + \text{penalidad por overfitting}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Error de validación

- Es posible probar que

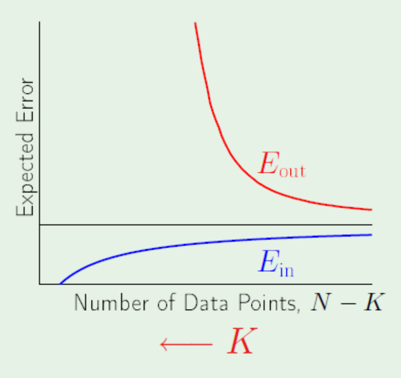
$$E_{val}(h) = E_{out}(h) \pm \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

- Donde K es el tamaño del conjunto de validación.
- Sin embargo, hay que recordar que los K puntos de validación salen del conjunto de entrenamiento!

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

El error en función de K

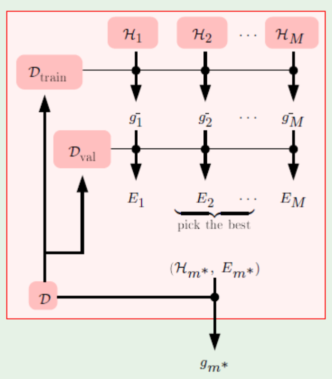
- Dado el siguiente conjunto de datos: $\mathcal{D} = (x_1, y_1), \ldots, (x_N, y_N)$
- Lo dividimos en conjunto de entrenamiento y conjunto de validación:
- Debido al término $\mathcal{O}(\frac{1}{\sqrt{K}})$, un K bajo llevará a un estimado de mala calidad del error real.
- Por otro lado, un K muy alto llevará a un entrenamiento de menor calidad, puesto que los datos de validación son datos que dejamos de usar en el entrenamiento.



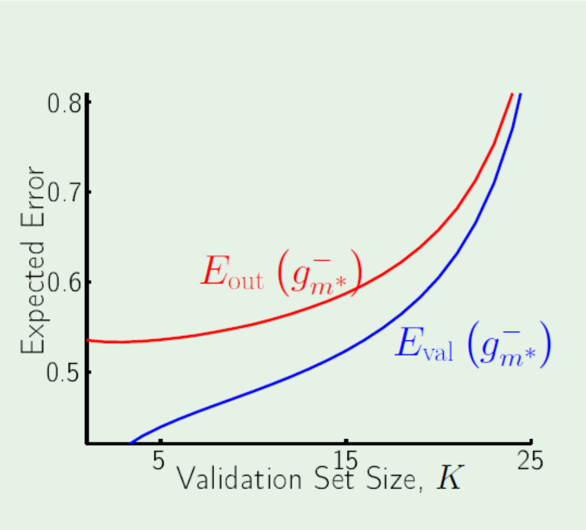
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Procedimiento de validación

- Se tienen M modelos de aprendizaje $\mathcal{H}_1, ..., \mathcal{H}_M$.
- Usamos $\mathcal{D}_{train}$ para entrenar cada modelo.
- Se evalua el rendimiento de cada modelo entrenado usando $\mathcal{D}_{val}$ item Se escoge como ganador al modelo con menor error de validación E_{val}.
- Se entrena el modelo ganador con $\mathcal{D}_{train} + \mathcal{D}_{val}$ y se prueba sobre $\mathcal{D}_{test}$



Sesgo de validación



Cuantos datos usar en validación?

- Heurísticamente, $K = \frac{N}{5}$
- Pero a veces no hay suficientes datos para darse ese lujo.
- En la situación anterior, nos enfrentamos a un dilema. Siendo g la hipótesis producto del entrenamiento con $\mathcal{D}_{train}$, y g^- la hipótesis producto del entrenamiento con $\mathcal{D}_{train} + \mathcal{D}_{val}$:

$$E_{out}(g) \underbrace{\approx}_{\text{K pequeño}} E_{out}(g^-) \underbrace{\approx}_{\text{K elevado}} E_{val}(g^-)$$

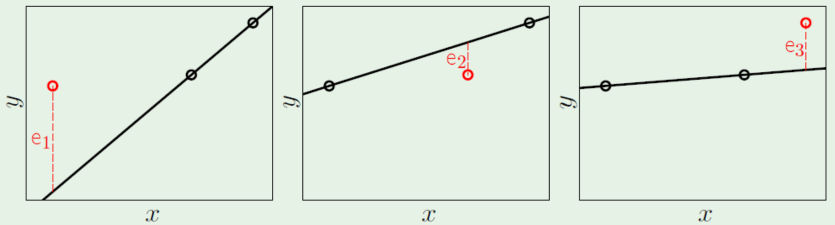
- Cómo podemos hacer al tiempo que K sea bajo y alto a la vez?
- La respuesta es Validación Cruzada: Usar un conjunto de datos tanto para el entrenamiento como para la validación.

- En este tipo de validación cruzada, se usan $N-1$ puntos para el entrenamiento, y un punto para validación.

$$\mathcal{D}_n = (x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n), (x_{n+1}, y_{n+1}), \ldots, (x_N, y_N)$$

- El error de validación cruzada sería el promedio del error cometido por la hipótesis producto del entrenamiento en cada uno de los ejemplos de entrenamiento.

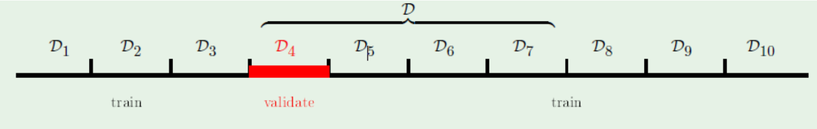
Ilustración Leave one out



$$E_{cv} = \frac{1}{3}\left(e_1 + e_2 + e_3 \right)$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- Dividimos el conjunto de entrenamiento en K pliegues (Usualmente $K = 10$).
- Se reserva uno de los pliegues para la validación, el resto se usa en el entrenamiento.
- El error de validación cruzada sería el promedio del error cometido por la hipótesis producto del entrenamiento en cada uno de los pliegues construidos.



Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS