Aplicando Modelos de Machine Learning

Luis Zúñiga

Agenda

1. Evaluación de Modelos de ML
2. Bias-Variance Tradeoff
3. Curvas de Aprendizaje
4. Juntando todo…

Evaluación de Modelos de ML

¿Qué modelo elegir?

Redes Neuronales

Regresión Logística

Regresión lineal

Naïve Bayes

Regresión polinomial

¿Qué modelo elegir?

Redes Neuronales

Regresión Logística

Regresión lineal

Naïve Bayes

Regresión polinomial

- ¿Es buena idea probar todos los modelos?
- ¿Cómo elegir un buen modelo candidato?
- Si obtengo malos resultados, ¿qué puedo hacer sin perder mucho tiempo?

¿Qué modelo elegir?

Supongamos que entrenamos un modelo para predecir **el precio de una casa con un modelo de regresión lineal con regularización**:

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

Desafortunadamente, al probar nuestro modelo entrenado (la hipótesis) con datos que jamás llegó a ver, nos damos cuenta que **las predicciones realizadas son malas**.

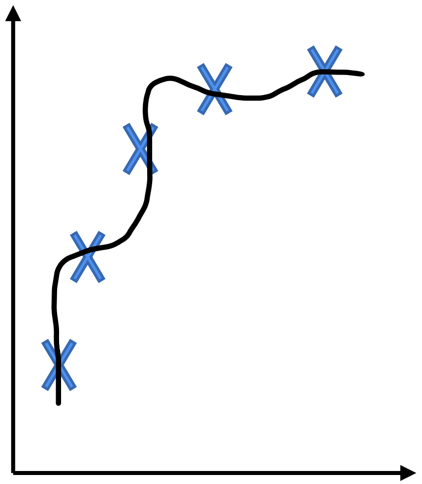
¿Qué se puede hacer?

¿Qué modelo elegir?

Después de mucho pensar, llegan a las siguientes opciones:

- Tener más datos para el entrenamiento.
- Considerar menos características.
- Obtener más características.
- Considerar combinaciones de características más complejas.
- Aumentar o reducir el valor de λ.

Evaluación del modelo (hipótesis)



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

Uno de los problemas que puede surgir al momento de entrenar los modelos es que exista sobreajuste o desajuste.

¿Cómo se detectaba?

Al probar el modelo con datos que jamás vio en el entrenamiento, tiene un pésimo rendimiento.

Evaluación del modelo (hipótesis)

Tamaño (m²)	Precio (m²)
150	2500
60	3230
60	3150
80	2120
60	1500
75	1670
300	4350
250	3750
150	2550
60	4500

Conjunto de entrenamiento (70%-80%) al azar

Se entrena con la partición que corresponde al conjunto de entrenamiento.

Conjunto de prueba (20%-30%) al azar

Se evalúa con la partición que corresponde al conjunto de prueba.

Train/Test - Procedimiento

Para el caso de regresión:

- Obtener los parámetros $\boldsymbol{\theta}$ mediante el entrenamiento con el conjunto de entrenamiento (minimizando el error mediante gradiente descendiente $J(\boldsymbol{\theta})$).

- Con los parámetros $\boldsymbol{\theta}$ encontrados, determinar el error con el conjunto de prueba:

$$J_{test}(\theta) = \frac{1}{2n_{test}} \sum_{i=1}^{n_{test}} \left(h_\theta\left(x_{test}^{(i)}\right) - y_{test}^{(i)} \right)^2$$

Train/Test - Procedimiento

Para el caso de clasificación:

- Obtener los parámetros $\boldsymbol{\theta}$ mediante el entrenamiento con el conjunto de entrenamiento

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

- Con los parámetros $\boldsymbol{\theta}$ encontrados, determinar el error con el conjunto de prueba:

$$Test\ error = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} err\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{test}^{(i)}\right), y_{test}^{(i)}\right)$$

$$err(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = \begin{cases} 1 \quad \text{si} \quad h_{\theta}(\boldsymbol{x}) \geq 0.5, y = 0 \\ 1 \quad \text{si} \quad h_{\theta}(\boldsymbol{x}) < 0.5, y = 1 \\ \ 0 \ \text{ en cualquier otro caso} \end{cases}$$

Accuracy, Recall, Precision o F1

Selección de Modelos – Conjunto de Validación

Supongamos que tenemos los siguientes modelos:

1. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$
2. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

$$\vdots$$

10. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$

Existe un nuevo parámetro d que rige el grado del polinomio de regresión lineal.

¿Cómo elegir el mejor?

Selección de Modelos – Conjunto de Validación

Supongamos que tenemos los siguientes modelos:

	Entrenamiento	Prueba
1. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$	$\boldsymbol{\theta}_1$	
2. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x$	$\boldsymbol{\theta}_2$	
3. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x + \theta_3 x$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_5$
$\vdots$	$\vdots$	
10. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x$	$\boldsymbol{\theta}_{10}$	

¿Ven algo mal?

Selección de Modelos – Conjunto de Validación

Supongamos que tenemos los siguientes modelos:

1. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$
2. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x$
3. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x + \theta_3 x$

$$\vdots$$

10. $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x$

Al elegir $\boldsymbol{\theta}_5$ usamos **el conjunto de prueba para comparar** los modelos entre sí y elegir el mejor.

Es una **estimación optimista del error de generalización** por el parámetro adicional d.

Selección de Modelos – Conjunto de Validación

Tamaño (m^2)	Precio (m^2)
150	2500
60	3230
60	3150
80	2120
60	1500
75	1670
300	4350
250	3750
150	2550
60	4500

Conjunto de entrenamiento (60%

Conjunto de prueba (20%)

Conjunto de validación (20%)

Se entrena con la partición que corresponde al conjunto de entrenamiento.

Se evalúa con la partición que corresponde al conjunto de prueba.

Se elige la que menor error tenga en el conjunto de validación como medida de generalización.

Selección de Modelos – Conjunto de Validación

Tamaño (m²)	Precio (m²)
150	2500
60	3230
60	3150
80	2120
60	1500
75	1670
300	4350
250	3750
150	2550
60	4500

Conjunto de entrenamiento (60%

Conjunto de prueba (20%)

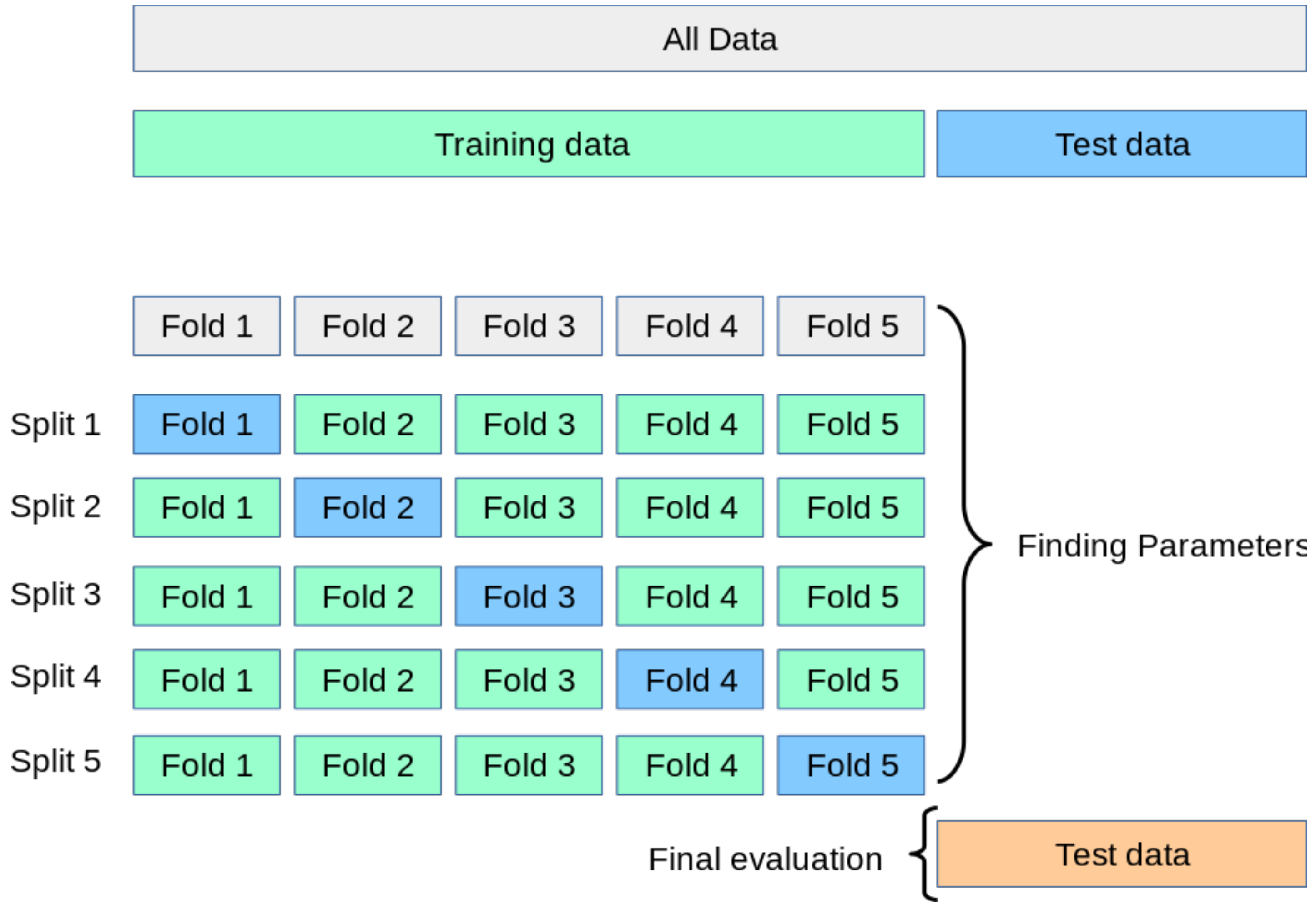
Conjunto de validación (20%)

Se entrena con la partición que corresponde al conjunto de entrenamiento.

Se evalúa con la partición que corresponde al conjunto de prueba.

Se elige la que menor error tenga en el conjunto de validación como medida de generalización.

Selección de Modelos – Validación Cruzada



Selección de Modelos – Validación Cruzada

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

Simplification is based on assumptions (model bias)

Assumptions fail in certain situations

Roughly speaking:

"No one model works best for all possible situations."

Bias - Variance

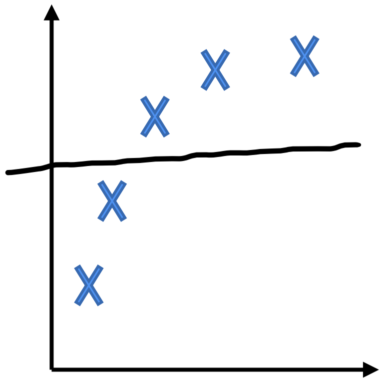
Bias - Variance

Como habíamos comentado hace unas semanas, el problema del *bias – variance tradeoff* es importante determinarlo. Si nuestro modelo no es bueno, puede:

- Tener un alto sesgo (bias).
- Tener alta varianza (variance).

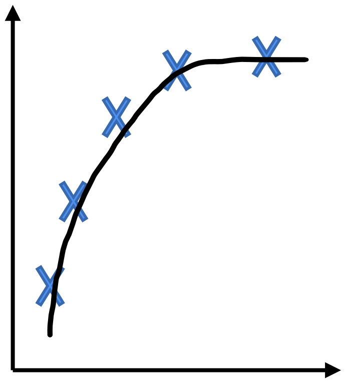
¿Cómo identificarlo?

Bias - Variance



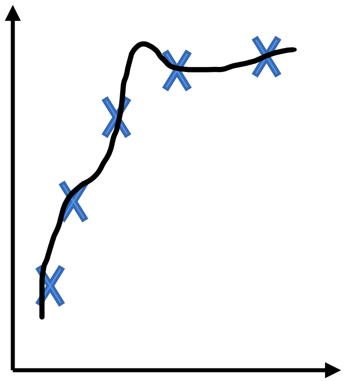
$\theta_0 + \theta_1 x$

Underfitting
(high bias)



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

Just Right



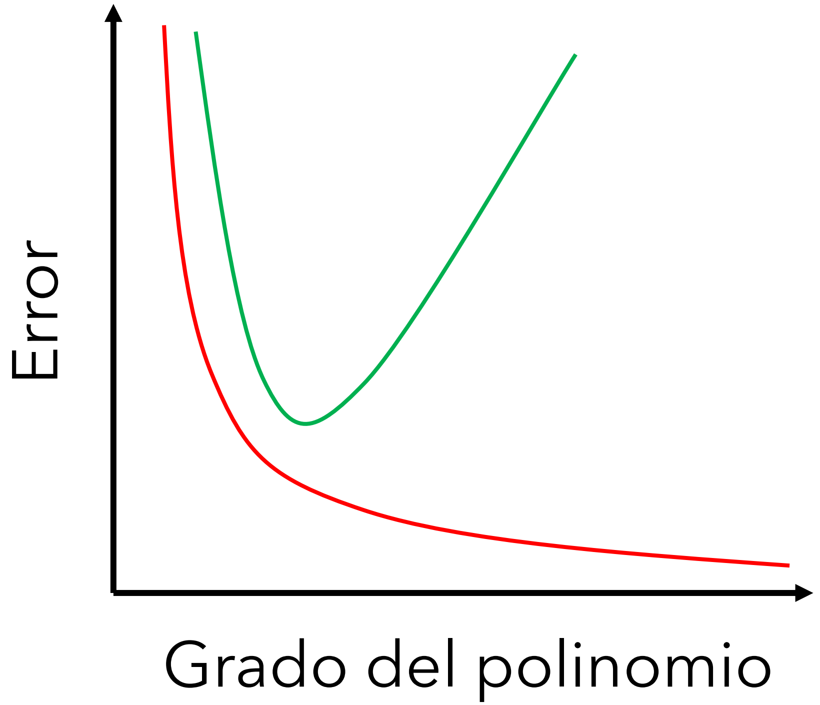
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Overfitting
(high variance)

Bias - Variance

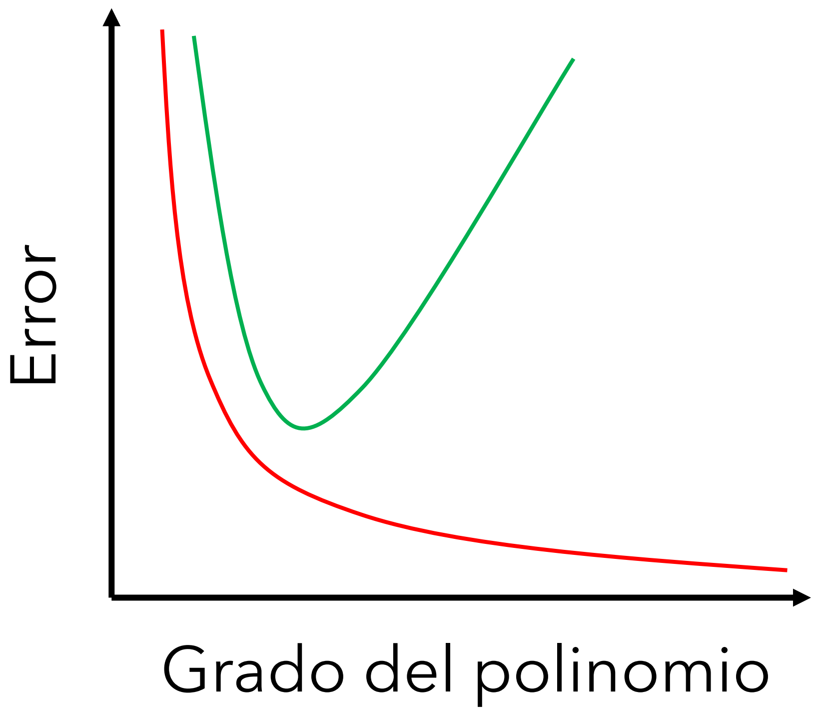
Error de entrenamiento: $J_{train}(\boldsymbol{\theta}) = \frac{1}{2n_{train}} \sum_{i=1}^{n_{train}} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{train}^{(i)} \right) - y_{train}^{(i)} \right)^2$

Error de validación: $J_{val}(\boldsymbol{\theta}) = \frac{1}{2n_{val}} \sum_{i=1}^{n_{val}} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{val}^{(i)} \right) - y_{val}^{(i)} \right)^2$



Bias - Variance

Considerando lo anterior, el error de prueba/validación $J(\boldsymbol{\theta})$ es alto, lo cual indica que algo anda mal. Pero, ¿qué es? ¿Sesgo o varianza?



Sesgo (*underfit*)

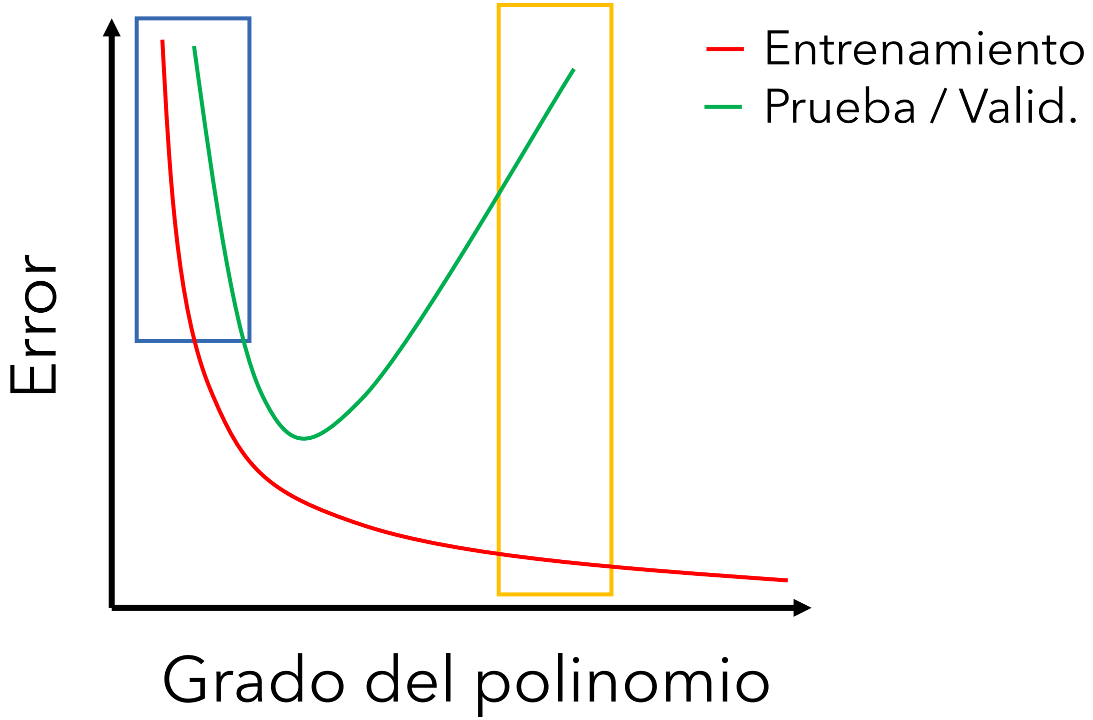
$J_{train}(\boldsymbol{\theta})$ alto
$J_{test}(\boldsymbol{\theta}) \approx J_{train}(\boldsymbol{\theta})$

Varianza (*overfit*)

$J_{train}(\boldsymbol{\theta})$ bajo
$J_{test}(\boldsymbol{\theta}) \gg J_{train}(\boldsymbol{\theta})$

Bias - Variance

Considerando lo anterior, el error de prueba/validación $J(\boldsymbol{\theta})$ es alto, lo cual indica que algo anda mal. Pero, ¿qué es? ¿Sesgo o varianza?



Sesgo (*underfit*)

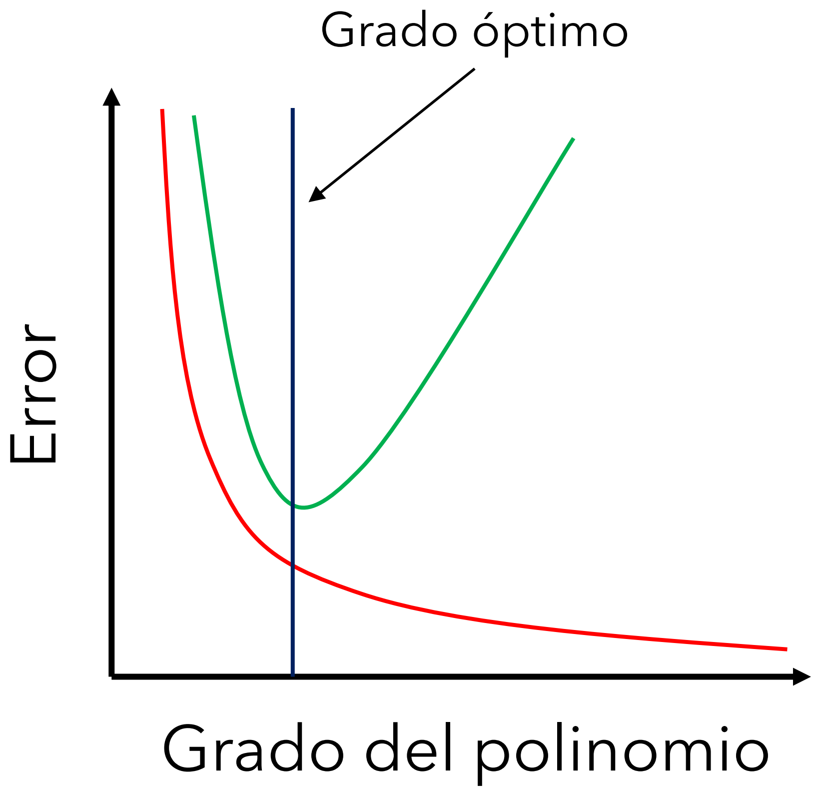
$J_{train}(\boldsymbol{\theta})$ alto
$J_{test}(\boldsymbol{\theta}) \approx J_{train}(\boldsymbol{\theta})$

Varianza (*overfit*)

$J_{train}(\boldsymbol{\theta})$ bajo
$J_{test}(\boldsymbol{\theta}) \gg J_{train}(\boldsymbol{\theta})$

Bias - Variance

Considerando lo anterior, el error de prueba/validación $J(\boldsymbol{\theta})$ es alto, lo cual indica que algo anda mal. Pero, ¿qué es? ¿Sesgo o varianza?



Sesgo (*underfit*)

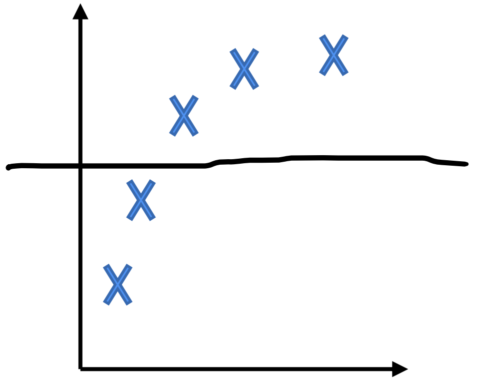
$J_{train}(\boldsymbol{\theta})$ alto
$J_{test}(\boldsymbol{\theta}) \approx J_{train}(\boldsymbol{\theta})$

Varianza (*overfit*)

$J_{train}(\boldsymbol{\theta})$ bajo
$J_{test}(\boldsymbol{\theta}) \gg J_{train}(\boldsymbol{\theta})$

Bias - Variance

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\left[\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{p}\theta_p^2\right] \qquad h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

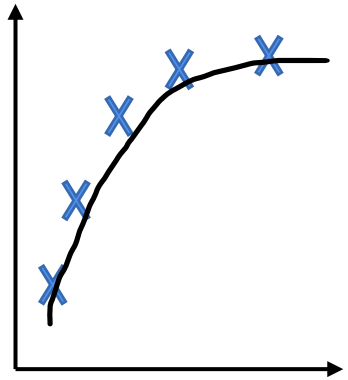


λ grande

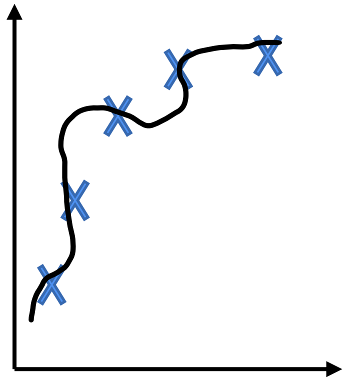
Alto sesgo (Underfitting)

$\lambda = 10000 \rightarrow \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$



λ intermedia



λ pequeña

Alto varianza (Overfitting)

$\lambda = .0001 \rightarrow \theta_1 \approx \theta_1, \theta_2$
$\approx \theta_2, \ldots$

Bias – Variance y Regularización

¡Se determinan los parámetros con **regularización**!

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

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

$$J_{train}(\theta) = \frac{1}{2n_{train}} \sum_{i=1}^{n_{train}} \left(h_\theta\left(x_{train}^{(i)}\right) - y_{train}^{(i)}\right)^2 \qquad J_{test}(\theta) = \frac{1}{2n_{test}} \sum_{i=1}^{n_{test}} \left(h_\theta\left(x_{test}^{(i)}\right) - y_{test}^{(i)}\right)^2$$

$$J_{val}(\theta) = \frac{1}{2n_{val}} \sum_{i=1}^{n_{val}} \left(h_\theta\left(x_{val}^{(i)}\right) - y_{val}^{(i)}\right)^2$$

¡Se evalúa el error **sin regularizar**!

Bias – Variance y Regularización

Modelo:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

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

1. Probar con $\lambda = 0$ $\quad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \to \theta^{(1)} \to J_{val}\left(\theta^{(1)}\right)$
2. Probar con $\lambda = 0.01$ $\quad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \to \theta^{(2)} \to J_{val}\left(\theta^{(2)}\right)$
3. Probar con $\lambda = 0.02$
4. Probar con $\lambda = 0.04$ $\quad \vdots$
5. Probar con $\lambda = 0.08$
6. Probar con $\lambda = 0.16$ $\quad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \to \theta^{(6)} \to J_{val}\left(\theta^{(6)}\right)$

$\vdots \qquad \vdots$

Entrenar el modelo con el conjunto de entrenamiento CON regularización y obtener nuestra hipótesis.

Evaluamos con el conjunto de validación/prueba con las definiciones sin regularización.

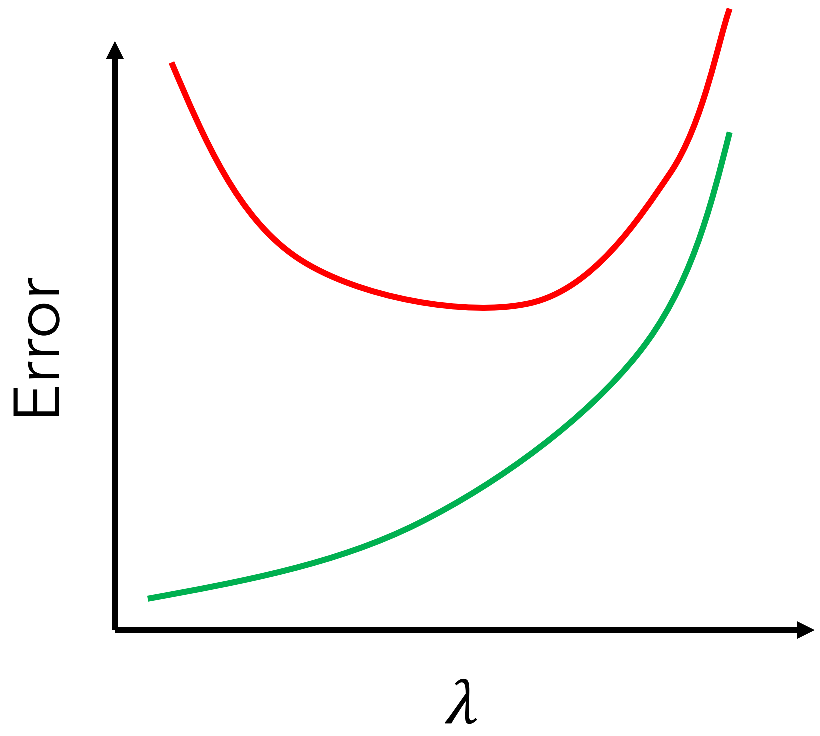
Se elige el que tenga menor error. Supongamos es θ_3.

Bias – Variance y Regularización

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

$$\boxed{J_{train}(\theta) = \frac{1}{2n_{train}}\sum_{i=1}^{n_{train}}\left(h_\theta\left(x_{train}^{(i)}\right) - y_{train}^{(i)}\right)^2}$$

$$\boxed{J_{val}(\theta) = \frac{1}{2n_{val}}\sum_{i=1}^{n_{val}}\left(h_\theta\left(x_{val}^{(i)}\right) - y_{val}^{(i)}\right)^2}$$

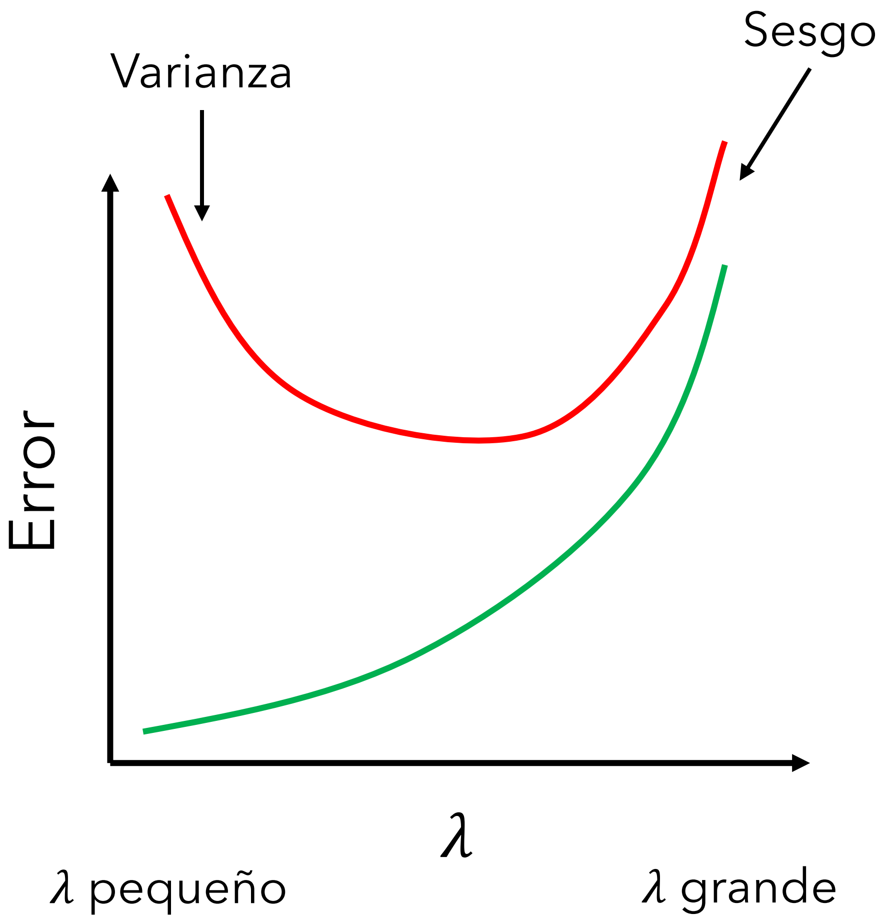


Bias – Variance y Regularización

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

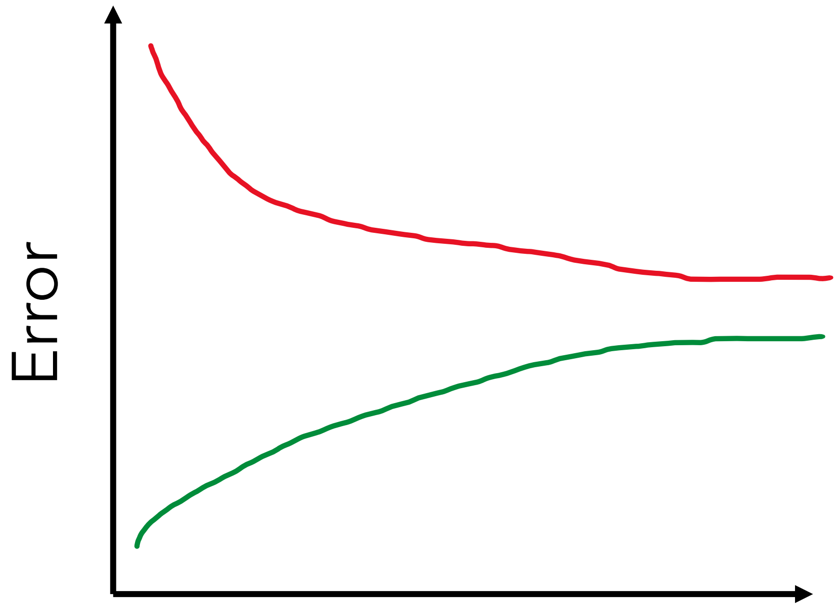
$$J_{train}(\theta) = \frac{1}{2n_{train}}\sum_{i=1}^{n_{train}}\left(h_\theta\left(x_{train}^{(i)}\right) - y_{train}^{(i)}\right)^2$$

$$J_{val}(\theta) = \frac{1}{2n_{val}}\sum_{i=1}^{n_{val}}\left(h_\theta\left(x_{val}^{(i)}\right) - y_{val}^{(i)}\right)^2$$



Curvas de Aprendizaje

Curvas de Aprendizaje

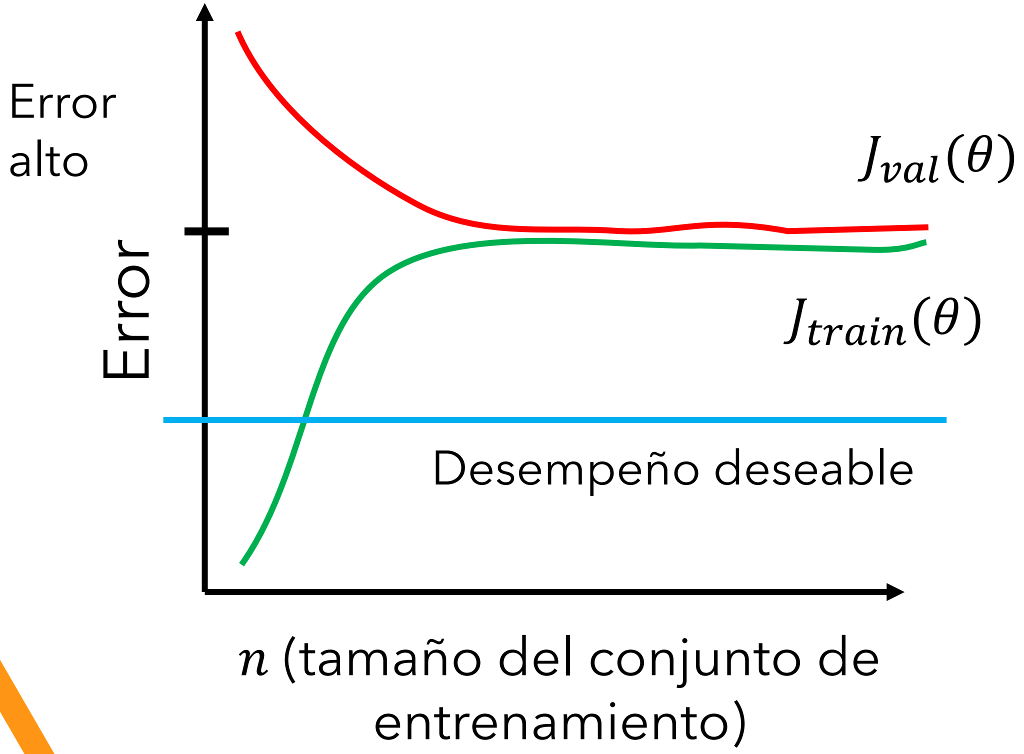


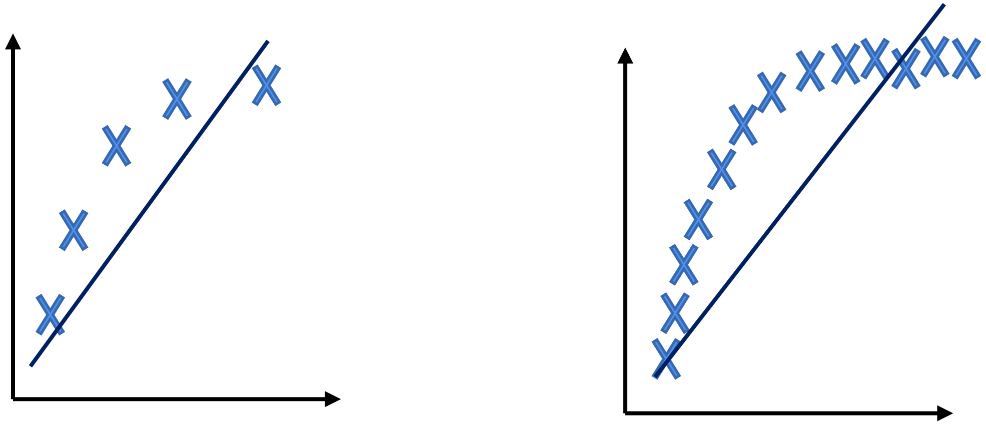
n (tamaño del conjunto de entrenamiento)

$$J_{train}(\theta) = \frac{1}{2n_{train}} \sum_{i=1}^{n_{train}} \left(h_\theta\left(x_{train}^{(i)}\right) - y_{train}^{(i)}\right)^2$$

$$J_{val}(\theta) = \frac{1}{2n_{val}} \sum_{i=1}^{n_{val}} \left(h_\theta\left(x_{val}^{(i)}\right) - y_{val}^{(i)}\right)^2$$

Curvas de Aprendizaje – Sesgo Alto



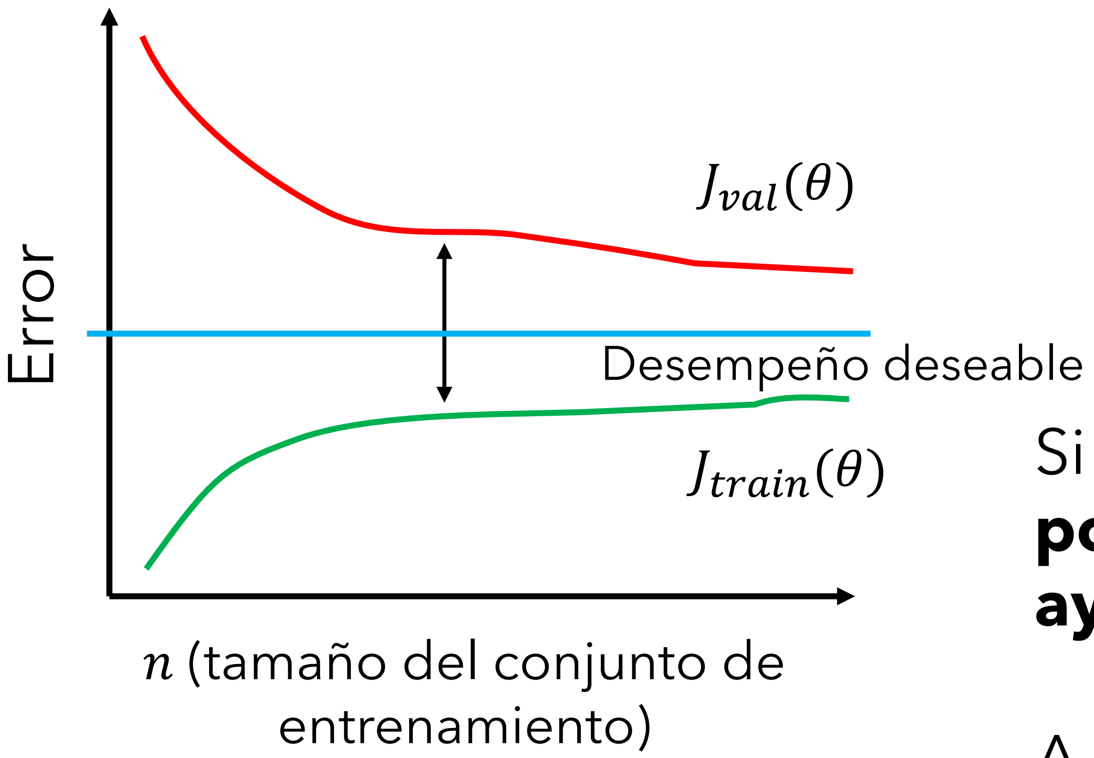


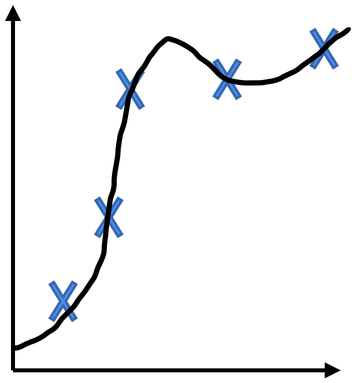
$$h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$

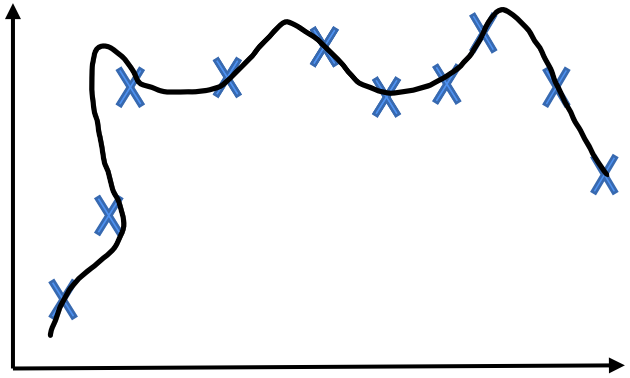
Si un algoritmo de aprendizaje sufre de alto sesgo, **obtener más información no ayuda a mejorar su rendimiento**.

Al ser un modelo muy simple, no logrará capturar la complejidad de la información, sin importar cuanta sea.

Curvas de Aprendizaje – Varianza Alta







$$h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \ldots + \theta_{100} x^{100}$$

Si un algoritmo tiene alta varianza, **es posible que añadir más información ayude a su desempeño**.

A largo plazo, ambas curvas pueden converger a una asíntota.

Juntemos todo…

¿Qué modelo elegir?

Considerando las opciones iniciales, tenemos que:

- Tener más datos para el entrenamiento. → *resuelve* varianza alta
- Considerar menos características. → *resuelve* varianza alta
- Obtener más características. → *resuelve* sesgo alto
- Considerar combinaciones de características más complejas. → *resuelve* sesgo alto
- Aumentar el valor de λ. → *resuelve* varianza alto
- Reducir el valor de λ → *resuelve* sesgo alto

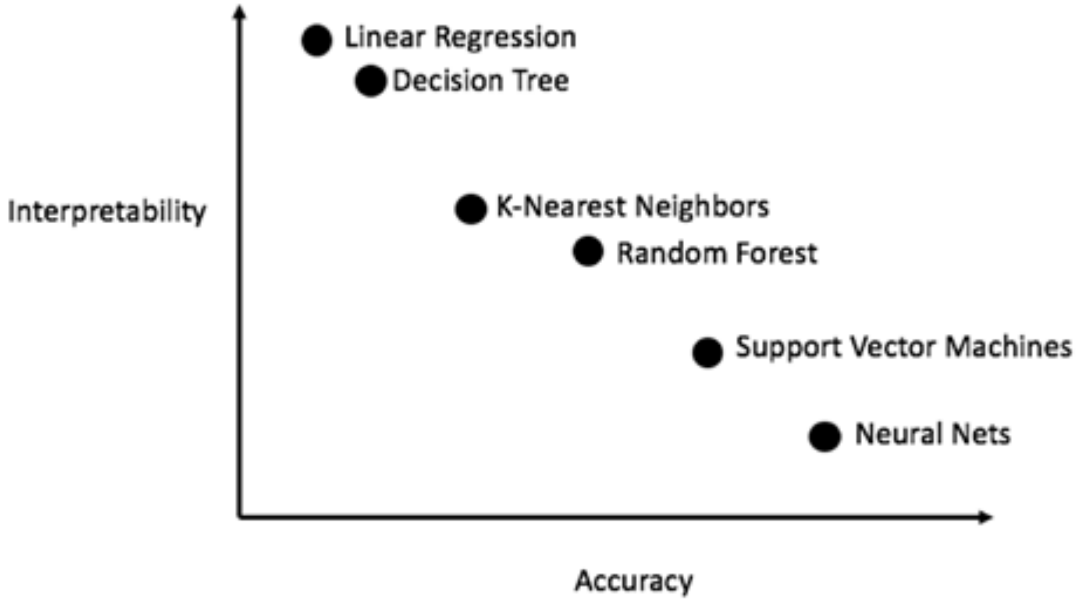
Conclusiones y Notas Adicionales

Conclusiones y Notas Adicionales

- Debemos elegir un modelo con un error aceptable para que sea capaz de generalizar.
- En el caso de regresión, esto implica elegir un polinomio de grado medio.
  - Muy bajo nos lleva al underfitting.
  - Muy alto nos lleva al overfitting.
- En el caso de clasificación, se debe tener cuidad con las funciones que activan umbrales y los métodos de clasificación.

Conclusiones y Notas Adicionales

- En todos los casos anteriores, alto sesgo implica baja varianza y viceversa.
- También se debe considerar el aspecto de la interpretabilidad de los modelos.



¡Gracias!

Luis Zúñiga

p40887@correo.uia.mx

Website