

Regresión lineal, polinómica y sobreajuste

16/04/2021

Laboratorio de datos 1°C 2021

Repaso regresión lineal

El escenario más sencillo con el que nos podemos encontrar es el de una variable dependiente y una única variable independiente, que siguen una relación aproximadamente lineal, salvo ruido (típicamente, normalmente distribuido):

$$y \sim \beta_0 + \beta_1 x$$

En general, vamos a desconocer la relación de arriba, pero **con un conjunto de datos y mediante cuadrados mínimos** podemos estimar los coeficientes:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Valor que predecimos de y dado x

Todo lo que tenga sombrero, son **cantidades estimadas**

Repaso regresión lineal

Si tenemos múltiples variables independientes, estamos en un caso de **regresión lineal múltiple**:

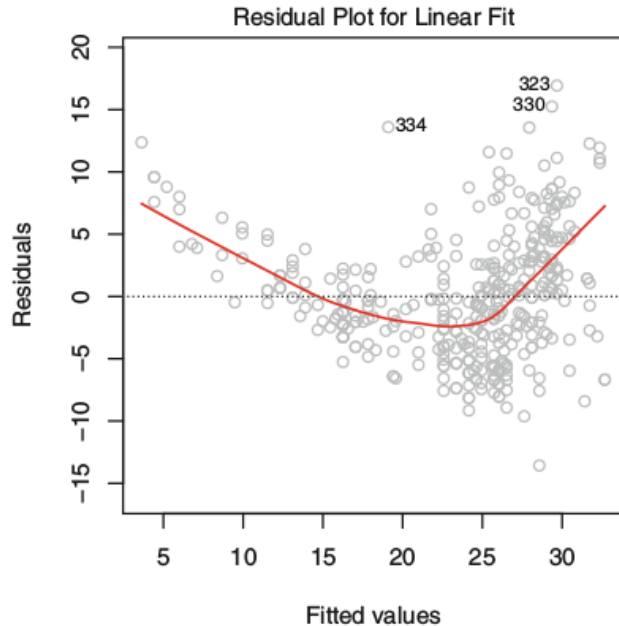
$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

Ejemplo: precio de un auto en términos de consumo de combustible, año de fabricación, etc.

Sin embargo, **este modelo nos impone una relación lineal** entre la variable dependiente y sus regresores.

Repaso regresión lineal

¿Cómo sabemos que necesitamos relaciones no-lineales?



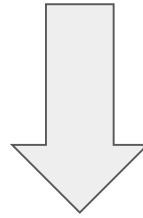
En el gráfico de los residuos **no deberíamos ver ningún patrón**. Si lo hubiera significa que el ruido depende de alguna de las variables independientes.

Gráfico de los residuos ($y - y_{\text{estimada}}$) en función de los valores estimados.




Repaso regresión lineal

¿Cómo podemos meter no linealidad en nuestro modelo?

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$



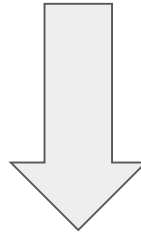
En ningún momento dijimos **quiénes son las x's**, en principio pueden ser cualquier cosa hasta... hasta cantantes de los 80's.

$$y \sim \beta_0 + \beta_1 \text{  } + \beta_2 \text{  } \dots + \beta_m \text{  }$$

Repaso regresión lineal

Independientemente quiénes son las x 's, el modelo es lineal respecto de los parámetros:

$$y \sim \beta_0 + \beta_1 \text{[img]} + \beta_2 \text{[img]} \dots + \beta_m \text{[img]}$$




Siempre que se cumplan las hipótesis del modelo lineal (como por ejemplo, no colinealidad entre los regresores), podemos incluir diferentes características.

$$y \sim \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m$$

Regresión de polinomios

La regresión polinómica sigue siendo lineal, ya que es **lineal respecto a los parámetros del modelo**.

$$y \sim \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_m x^m$$


Los parámetros se encuentran minimizando la suma del cuadrado de los residuos al igual que antes:

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$$

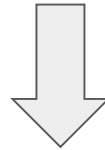
Grado del polinomio:
hiper-parámetro del modelo.

Regresión de polinomios

Este esquema funciona si nuestra variable dependiera de más de una variable independiente e incluso si uno de los coeficientes dependiera de una de dichas variables:

$$y \sim \beta_0 + (\beta_1 + \beta_3 z)x + \beta_2 z$$

Coeficiente dependiente de z



$$y \sim \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

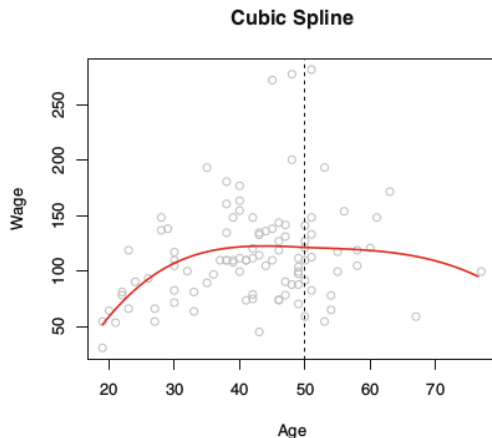
Términos de interacción

Regresión de funciones base

Este esquema es aún más general:

Las funciones pueden ser exponenciales, senos, cosenos, etc. (sin parámetros).

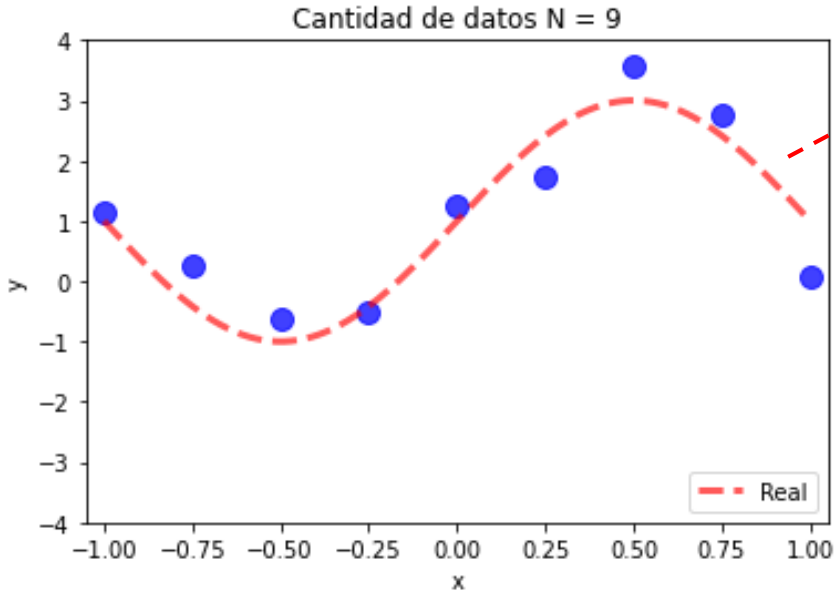
$$y \sim \beta_0 + \beta_1 \phi_1(\bar{x}) + \beta_2 \phi_2(\bar{x}) + \dots + \beta_m \phi_m(\bar{x})$$



Ejemplo muy utilizado: splines cúbicas.
Ver libro Introduction to Statistical Learning.

Modelo de juguete

La mayor cantidad de los ajustes que vamos a ver en las diapos siguientes se corresponden con este set de datos generados sintéticamente:



$$y = a * \sin(\pi x) + b.$$

$$a = 2 \text{ y } b = 1$$

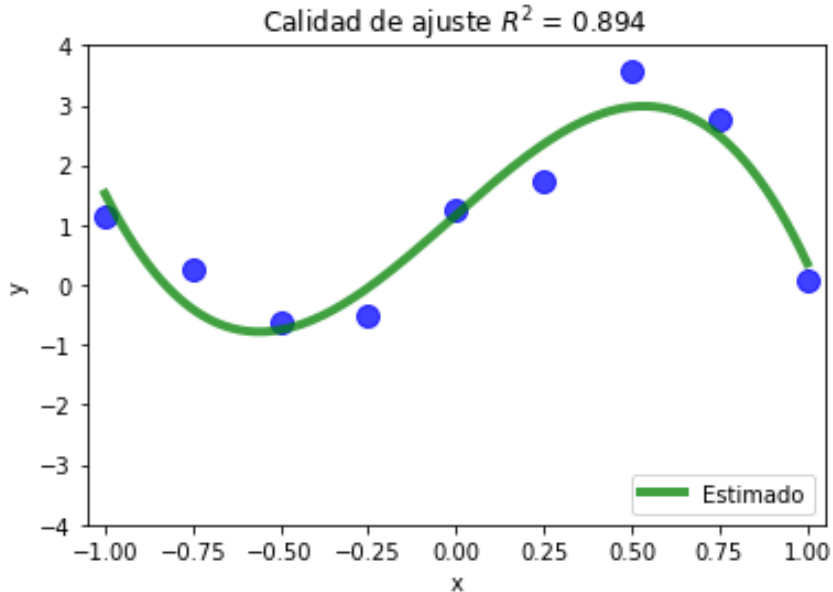
$$y + \epsilon$$

$N(0, \sigma = 0.5)$

Los datos son la relación de arriba más un ruido normal

Modelo de juguete

Inspeccionando un poco los datos, proponemos un polinomio de grado 3 (viendo por ejemplo que los datos presentan un mínimo y máximo, o un único punto de inflexión).



Modelo propuesto

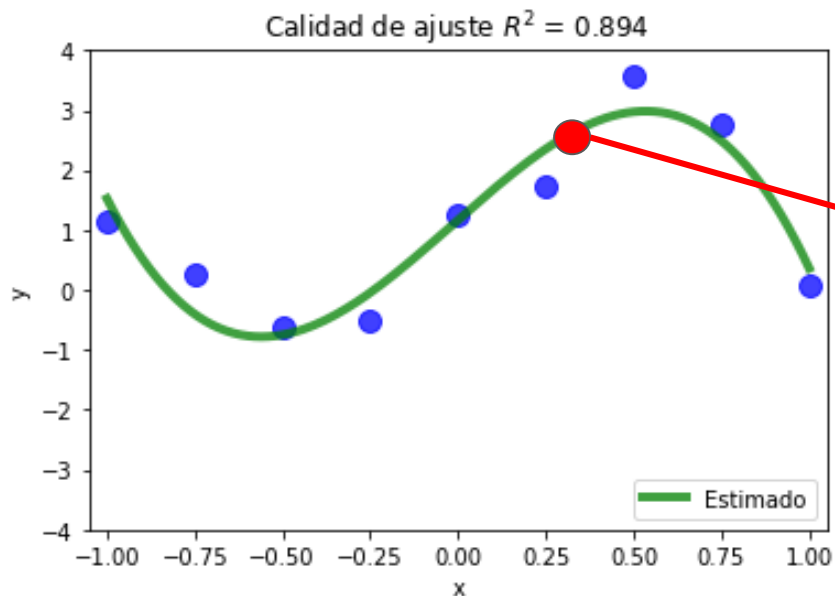
$$y \sim \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Coeficientes
estimados

Coefs.	d=3
0	1.17
1	5.16
2	-0.23
3	-5.74

Modelo de juguete

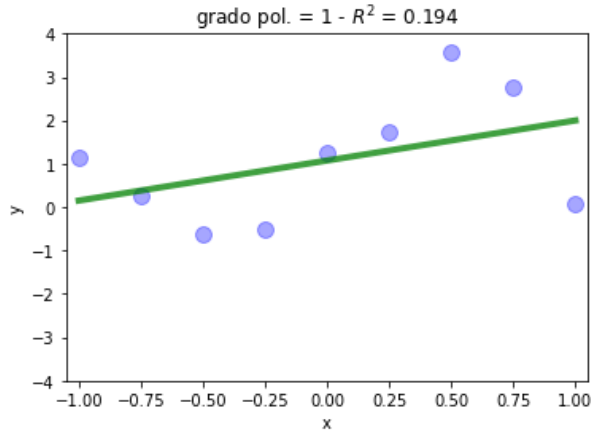
Inspeccionando un poco los datos, proponemos un polinomio de grado 3 (viendo por ejemplo que los datos presentan un mínimo y máximo, o un único punto de inflexión).



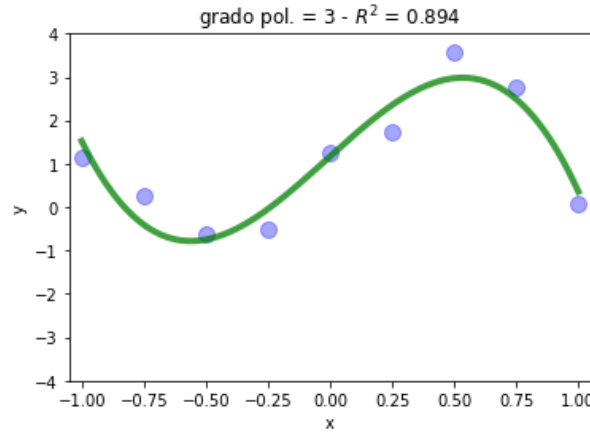
Con los coeficientes estimados, podemos hacer predicciones para nuevos valores de x . Por ejemplo: $x = 0.30$ y ~ 2.54 .

Pero, ¿cómo elegimos el grado del polinomio?

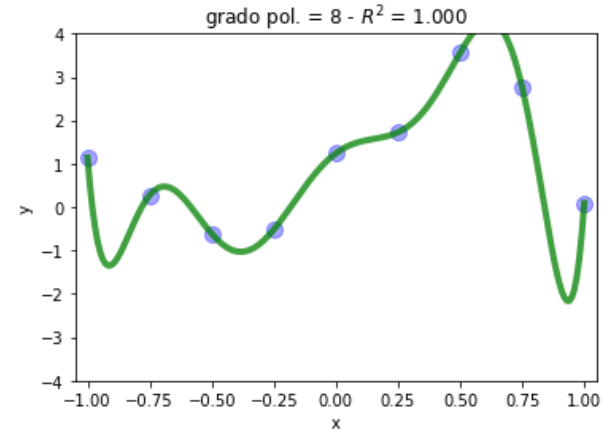
¿Qué pasa si probamos con polinomio de grado $m < N$ cualquiera?



underfitting

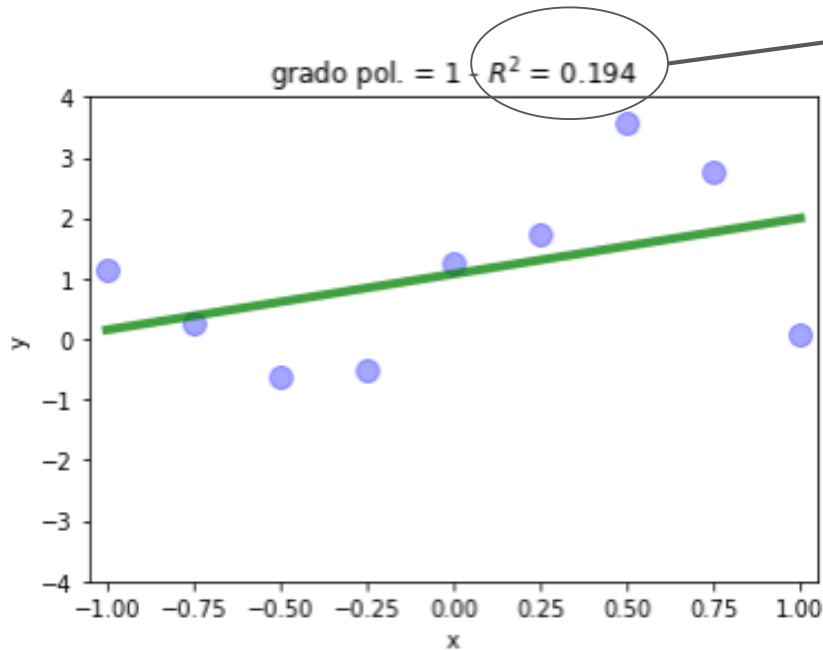


modelo más complejo



overfitting

Sub-ajuste (underfitting)

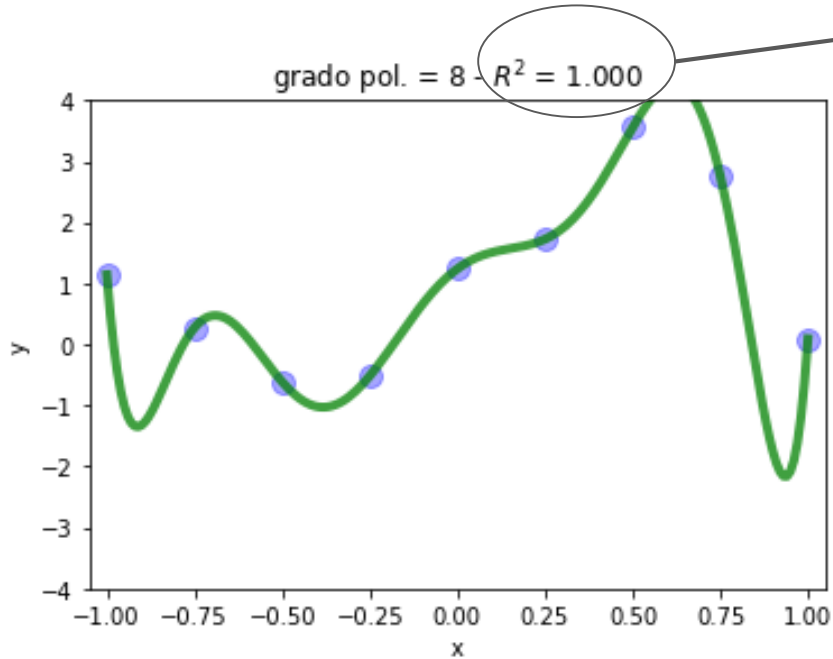


El ajuste es bastante malo (tenemos muy pocos grados de libertad).

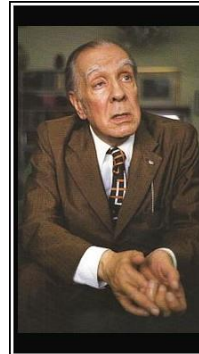
Los datos presentan una variabilidad, más allá del ruido intrínseco, que **un modelo simple no puede captar.**



Sobreajuste (overfitting)



El ajuste es perfecto (tengo tantos grados de libertad como datos en mi sistema). La curva pasa por todos los puntos (describe exactamente la variabilidad de los datos).

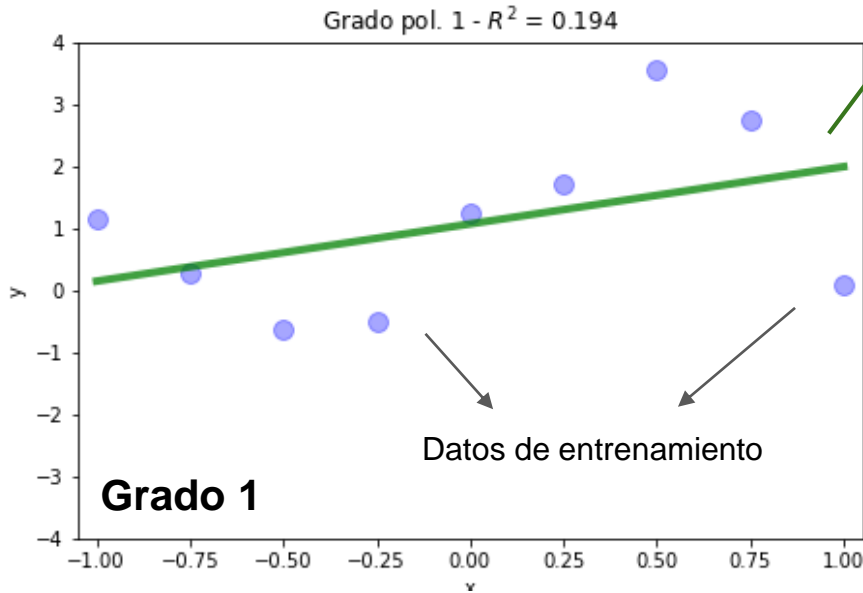


Sospecho, sin embargo que no era muy capaz de pensar. Pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos.

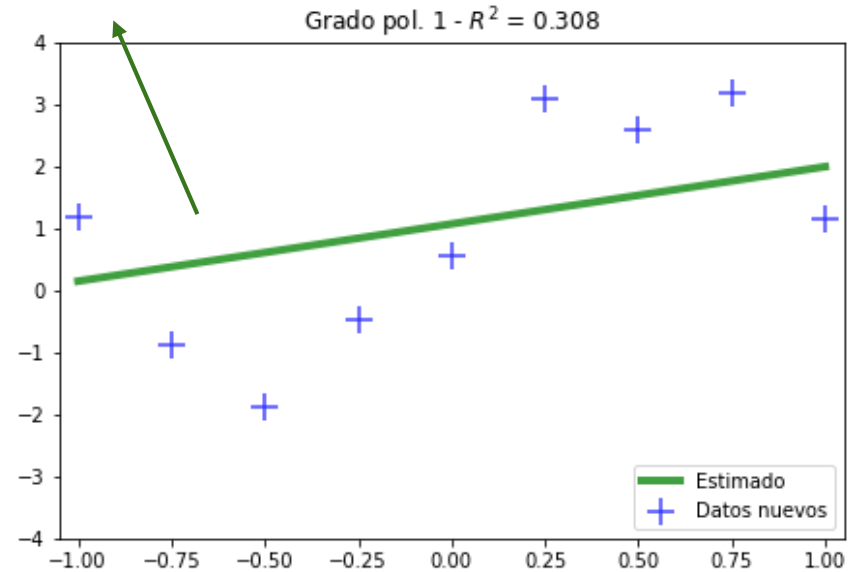
(Jorge Luis Borges)

¿Cuál es el problema de sub-ajustar o sobre-ajustar?

Hacer una mala predicción...



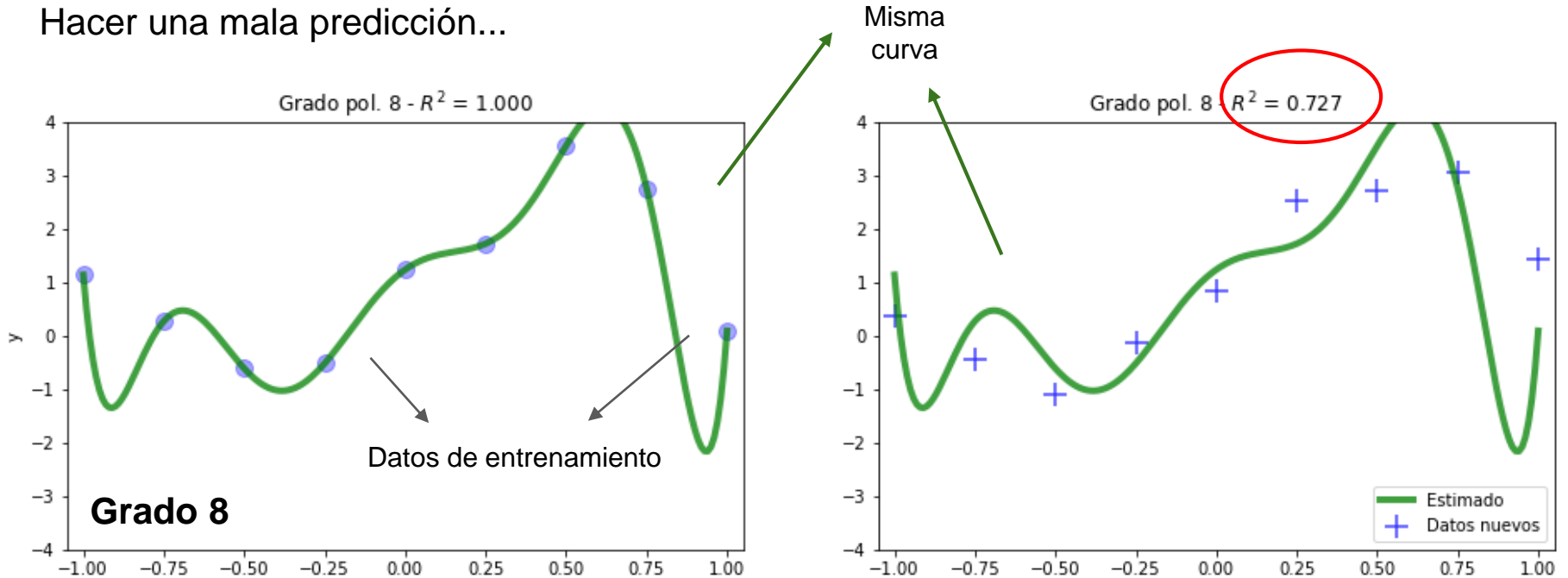
Misma curva



Si nuestro modelo sub-ajusta, **no logra captar la variabilidad de los datos** y por lo tanto comete sistemáticamente el mismo error.

¿Cuál es el problema de sub-ajustar o sobre-ajustar?

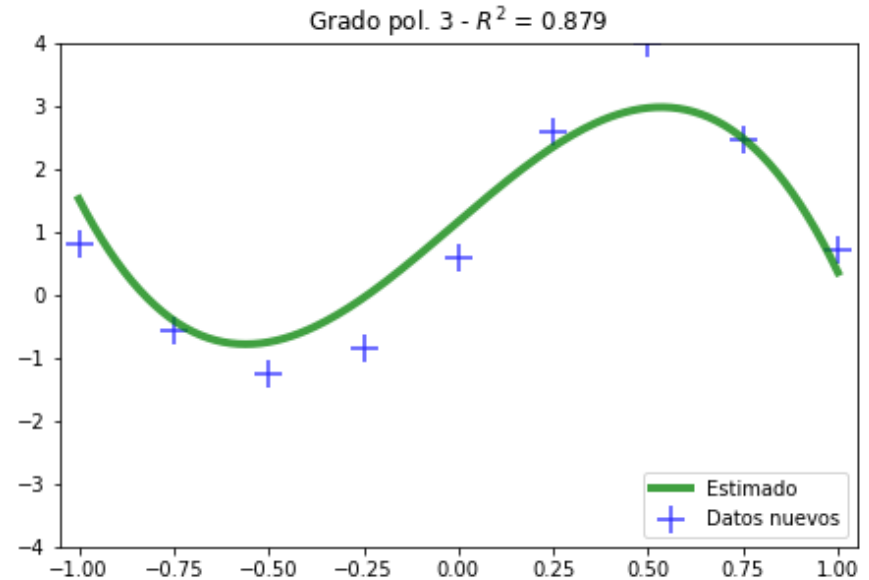
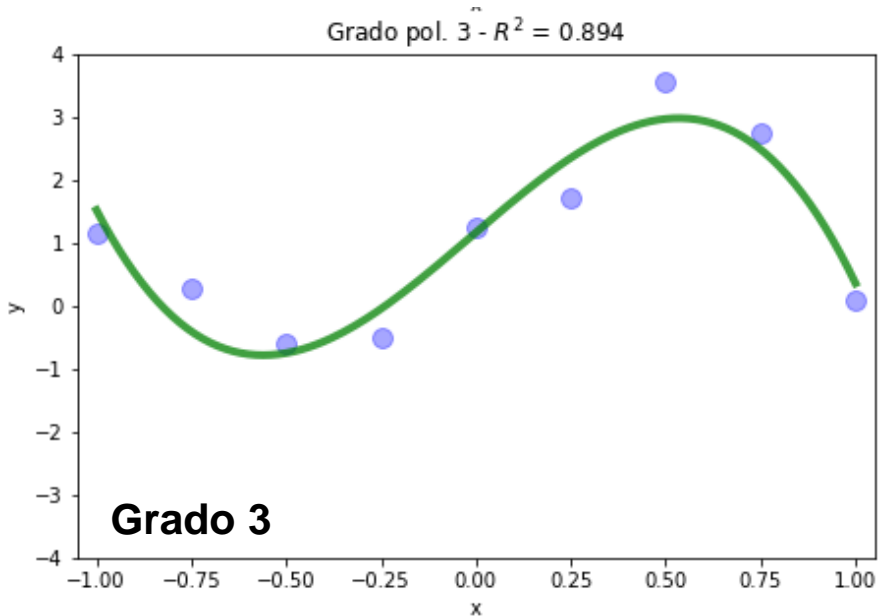
Hacer una mala predicción...



Si nuestro modelo sobre-ajusta, **la estimación perfecta en los datos de entrenamiento se pierde** cuando intentamos predecir nuevos datos.

¿Cuál es el problema de sub-ajustar o sobre-ajustar?

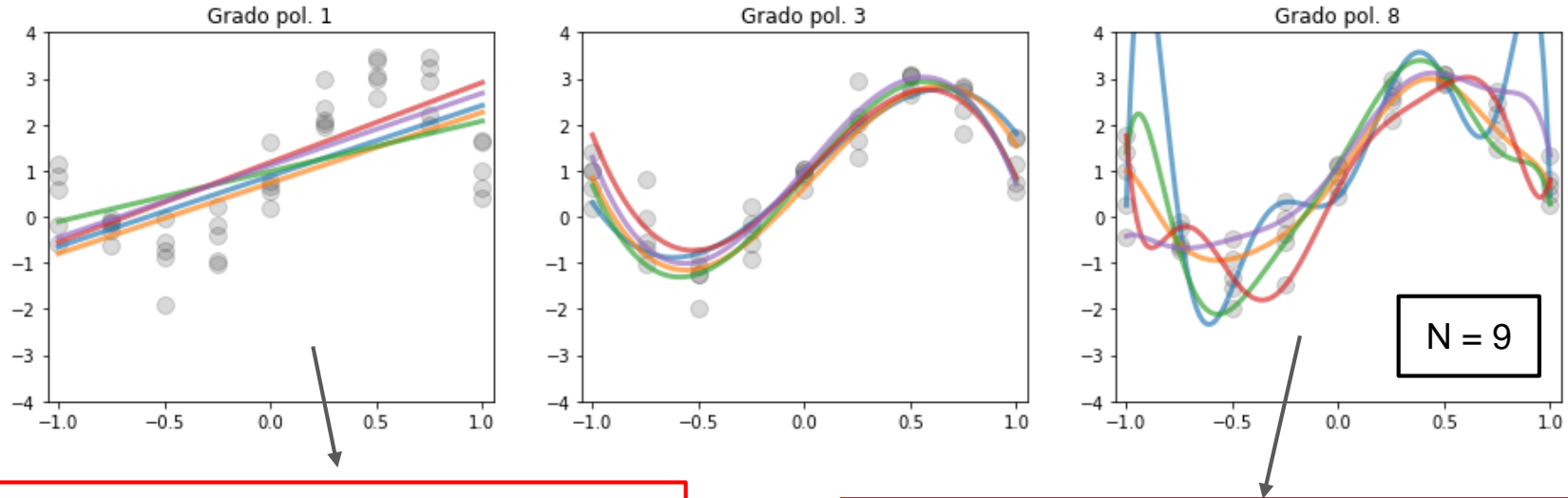
Si el modelo es el MODELO, se siente en la piel...



Un buen modelo **describe bien** datos distintos a los que fue entrenado.

¿Cuál es el problema de sub-ajustar o sobre-ajustar?

¿Qué pasa si cambian los datos sobre los ajustamos nuestro modelo? (Datos siempre provenientes de la misma población)



Un modelo que sub-ajusta es robusto ante diferentes datasets, pero es un mal predictor y comete errores sistemáticos.

Si el modelo sobre-ajusta es altamente sensible ante cambios en los datos de ajuste.

Sesgo y varianza

$$Y = f(X) + \epsilon$$

datos observados (pointing to Y)
función objetivo (pointing to f(X))
ruido en los datos (pointing to ϵ)

Cuentas
solamente a
modo informativo

Error esperado del modelo:

$$\begin{aligned} \mathbf{E}[y_0 - \hat{f}(x_0)]^2 &= \mathbf{E}[f(x_0) + \epsilon_0 - \hat{f}(x_0)]^2 \\ &= \mathbf{E}[f(x_0) - \hat{f}(x_0)]^2 + \text{Var}(\epsilon) \\ &= \dots \quad (\text{pasos engorrosos aquí}) \\ &= \underbrace{\left(\mathbf{E}[\hat{f}(x_0)] - y_0\right)^2}_{\text{Sesgo}} + \underbrace{\mathbf{E}\left[\hat{f}(x_0) - \mathbf{E}[\hat{f}(x_0)]\right]^2}_{\text{Varianza}} + \underbrace{\text{Var}(\epsilon)}_{\text{Error no reducible}} \end{aligned}$$

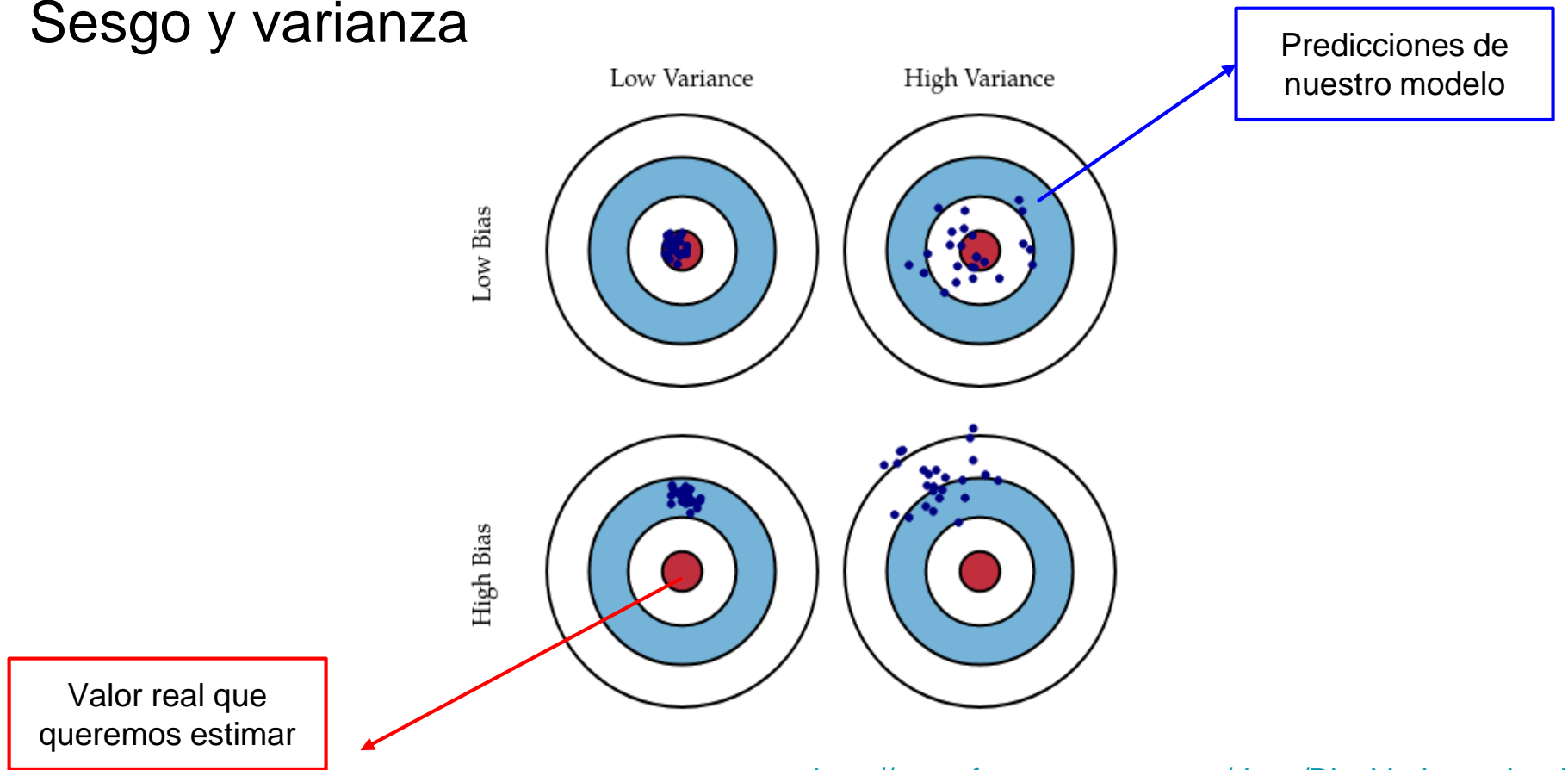
Esto es teórico, valor
esperado al
promediar sobre
diferentes datasets.

Sesgo y varianza

Interpretación:

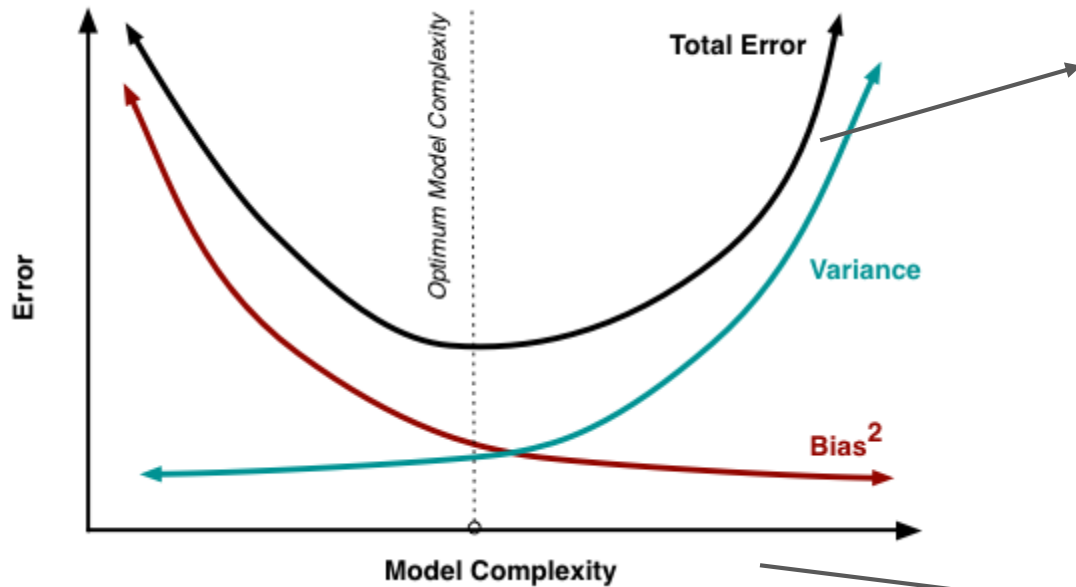
- **Sesgo:** es el error sistemático de nuestras predicciones. Si ajustamos nuestro modelo en diferentes datasets, ¿en cuánto difiere el valor medio de mis predicciones respecto del valor real?
- **Varianza:** si nos paramos en un punto y hacemos diferentes predicciones según diferentes datasets, ¿cuánto fluctúa mi predicción?

Sesgo y varianza



Sesgo y varianza

Lo importante es que el error que comete nuestro modelo viene de dos fuentes, sesgo y varianza, que compiten entre sí al variar la complejidad del modelo.

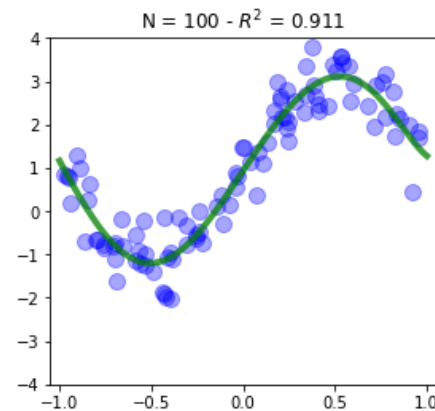
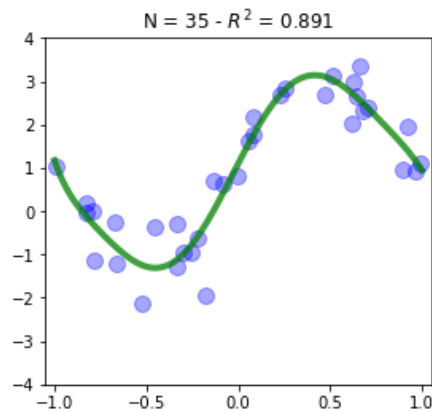
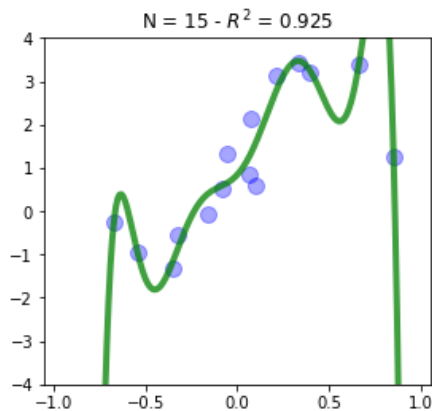
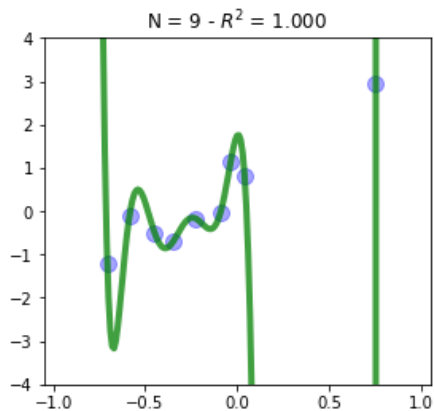


Trade-off sesgo-varianza: balance óptimo

En términos de polinomios, léase grado del mismo.

Otras características del sobreajuste

¿Qué pasa si dejamos fijo el grado y conseguimos más y más datos sobre los que ajustar?

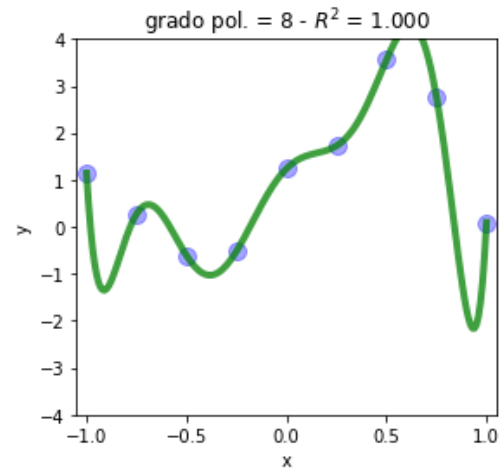
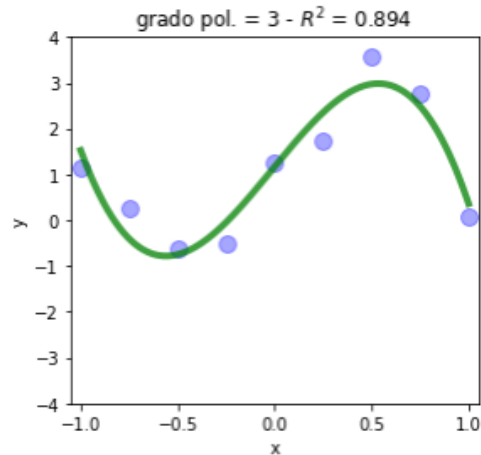


Polinomio de grado 8

El sobreajuste es un efecto de que la **cantidad de parámetros libres sea comparable con la cantidad de datos** con los que ajustamos.

Otras características del sobreajuste

Coefs.	d=1	d=3	d=5	d=8
0	1.07	1.17	0.76	1.24
1	0.92	5.16	5.70	4.23
2		-0.23	3.28	-16.90
3		-5.74	-7.97	5.80
4			-3.41	121.64
5			1.74	-28.54
6				-233.17
7				17.99
8				127.80



Al sobreajustar los coeficientes suelen tomar valores muy altos.

Regularización

La idea de regularizar el polinomio es prevenir que los coeficientes no adopten valores absolutos muy altos, asociados a cambios bruscos en la curva ajustada.

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^M |\hat{\beta}_j|^q$$

Próxima diapo...

Suma de los errores al cuadrado

Parámetro a tunear. Otro ejemplo de hiper-parámetro

Término de penalización: se suele excluir al parámetro beta0.

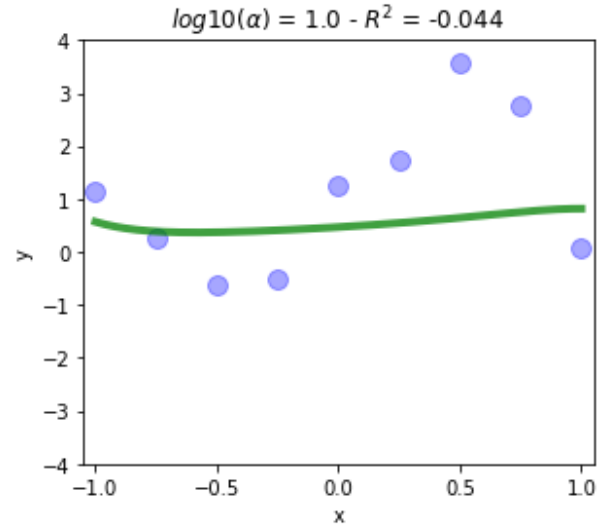
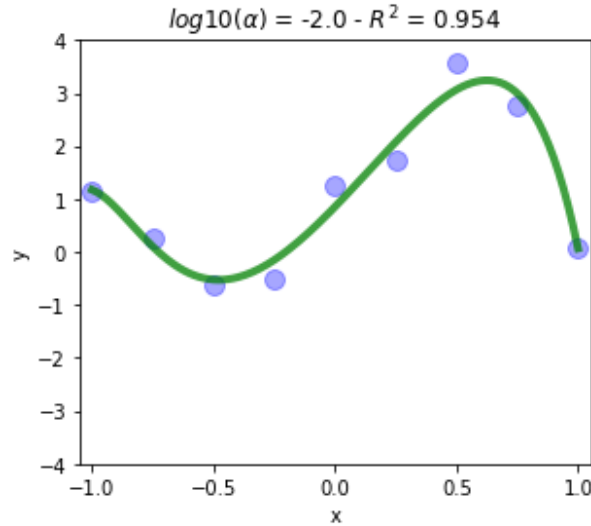
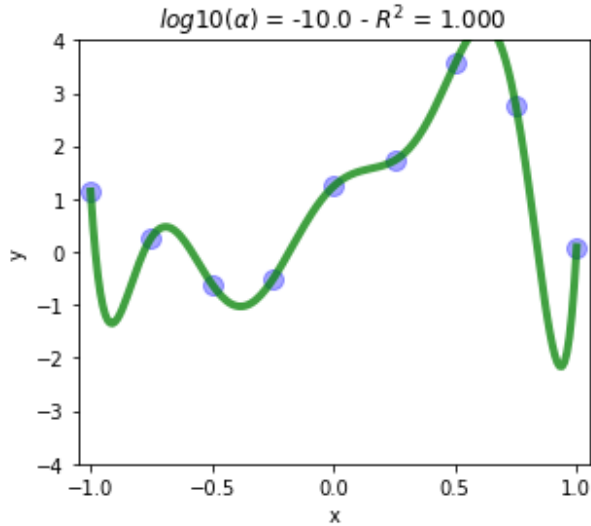
Regularización

La idea de regularizar el polinomio es prevenir que los coeficientes no adopten valores absolutos muy altos, asociados a cambios bruscos en la curva ajustada.

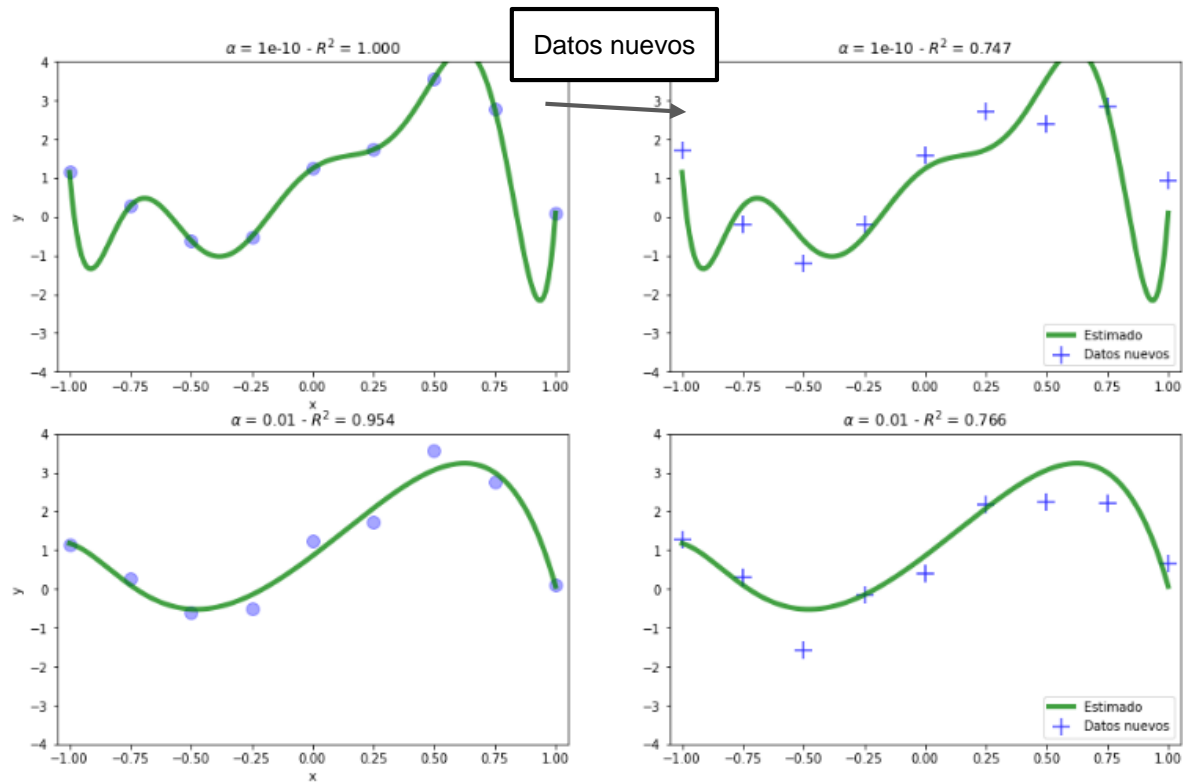
- **$q = 2$ (*ridge regression*):** deja a la función a minimizar cuadrática, con lo cual el proceso de minimización es muy parecido al de cuadrados mínimos.
- **$q = 1$ (*lasso regression*):** para valores de alfa altos, fuerza a que muchos coeficientes se vayan a 0, lo cual hace que el modelo se vuelva “esparso” (*sparse* en inglés, con muchos ceros). Ayuda a interpretar mejor modelo ya que **actúa como un selector de las variables importantes** (se queda con los términos dominantes y descarta los otros).

Regularización

+ alfa



Aumentar el término de penalización lleva a que el modelo sea más simple y previene el sobreajuste. Pero un valor muy alto, lleva a un polinomio de grado 0, que sabemos puede empezar a sub-ajustar.



Pasamos de la elección de un grado adecuado a la elección de un valor adecuado de alfa... y qué es un valor adecuado de alfa?



Con esto podemos ponerle un montón de parámetros en el modelo (un grado alto en el polinomio) y el término de regularización va a ponderar solo los soportados por los datos (+ sesgo y - varianza).

¿Cómo estimamos los hiper-parámetros, atacando el problema de sesgo-varianza?

- La elección del grado del polinomio y el parámetros se deben hacer con cuidado, buscando modelos que tengan bajo sesgo y baja varianza.
- Técnicas que vamos a ver en la materia:
 - Separación en datos de testeo y entrenamiento.
 - Validación cruzada.

Referencias

- Capítulo 3 y secciones 6.2 y 7.1 del libro “An Introduction to Statistical Learning”. James, Witten, Hastie & Tibshirani.
- Sección 1.1 de “Pattern Recognition and Machine Learning”. Bishop.
- <http://scott.fortmann-roe.com/docs/BiasVariance.html>