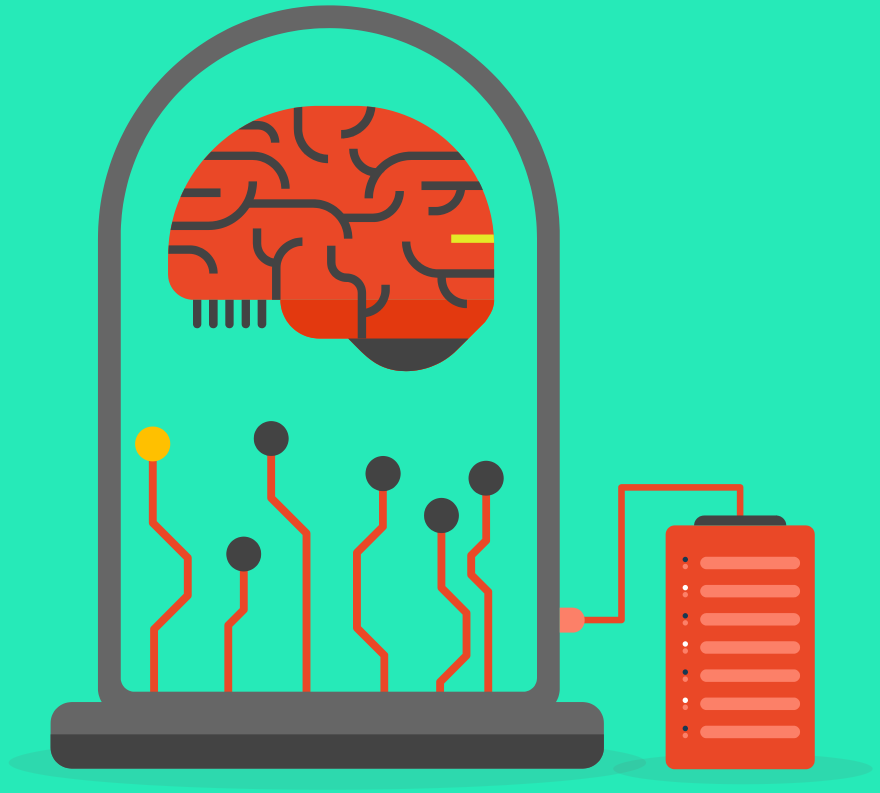
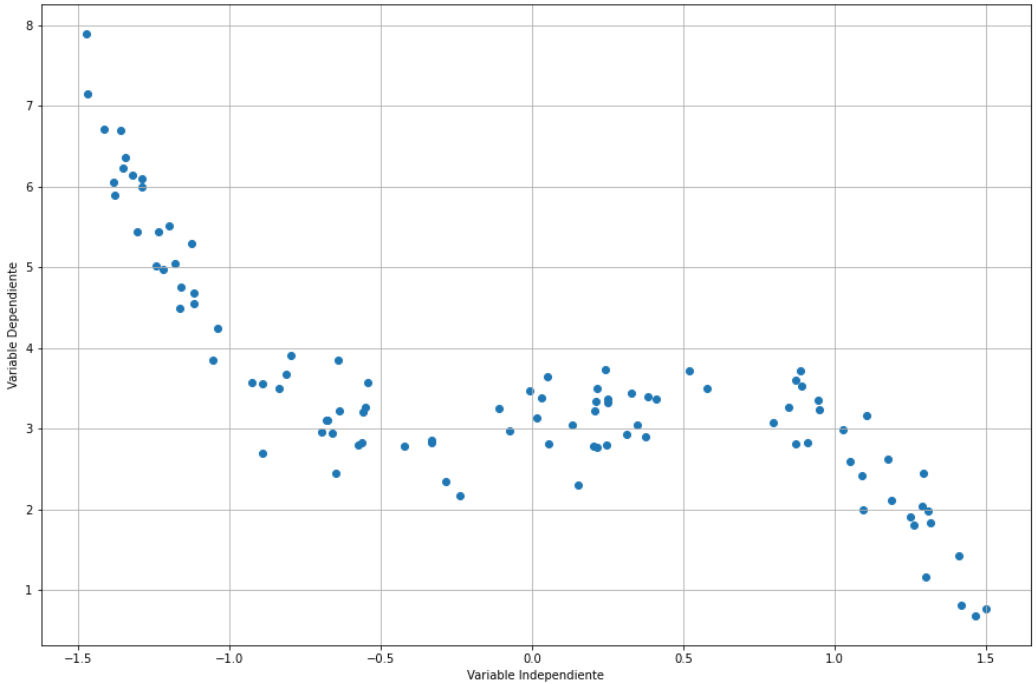
2. Regresión Polinómica



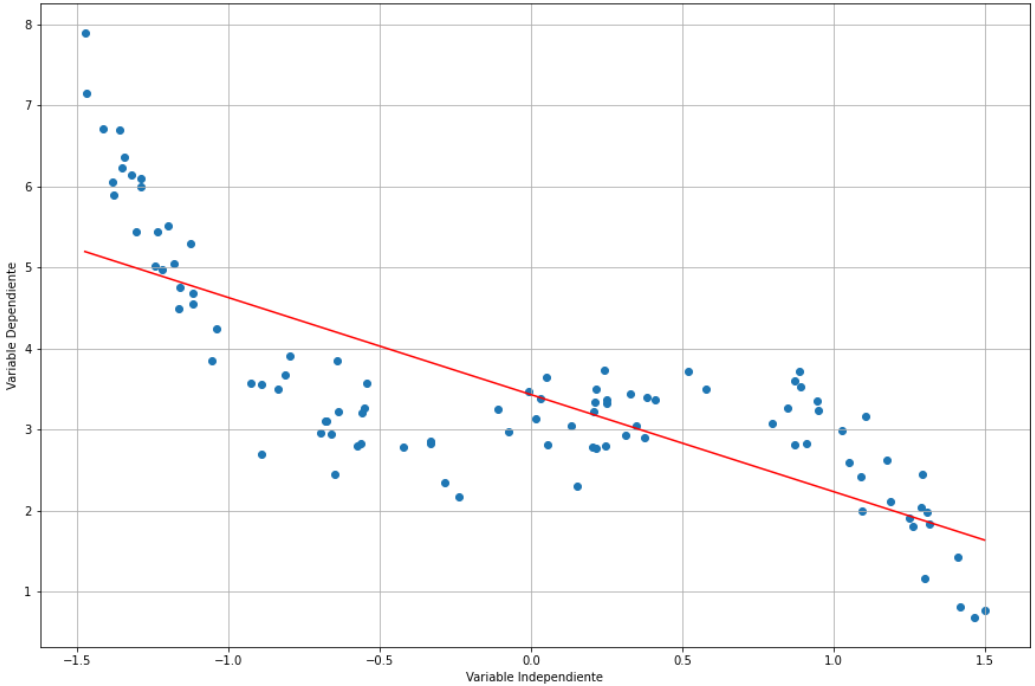
Planteamiento del Problema

Supongamos ahora que tenemos un conjunto de datos que, de un modo evidente, no presenta un comportamiento intrínsecamente lineal



Planteamiento del Problema

Al observar los datos, resulta evidente que una línea recta, aunque se ajuste a los datos, no representará un buen modelo que sirva para predecir nuevos datos similares a los observados



Regresión Polinómica

Una manera de resolver este problema es aplicar un modelo de regresión lineal similar a los vistos anteriormente, pero agregando variables de entrada que sean funciones polinómicas de la variable independiente original. Es decir:

x	y
-2	15
-1	4
0	3
1	3
2	-5
3	-30

x	x^2	x^3	y
-2	4	-8	15
-1	1	-1	4
0	0	0	3
1	1	1	3
2	4	8	-5
3	9	27	-30

Regresión Polinómica

Al realizar esta operación, estamos transformado la hipótesis lineal original:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

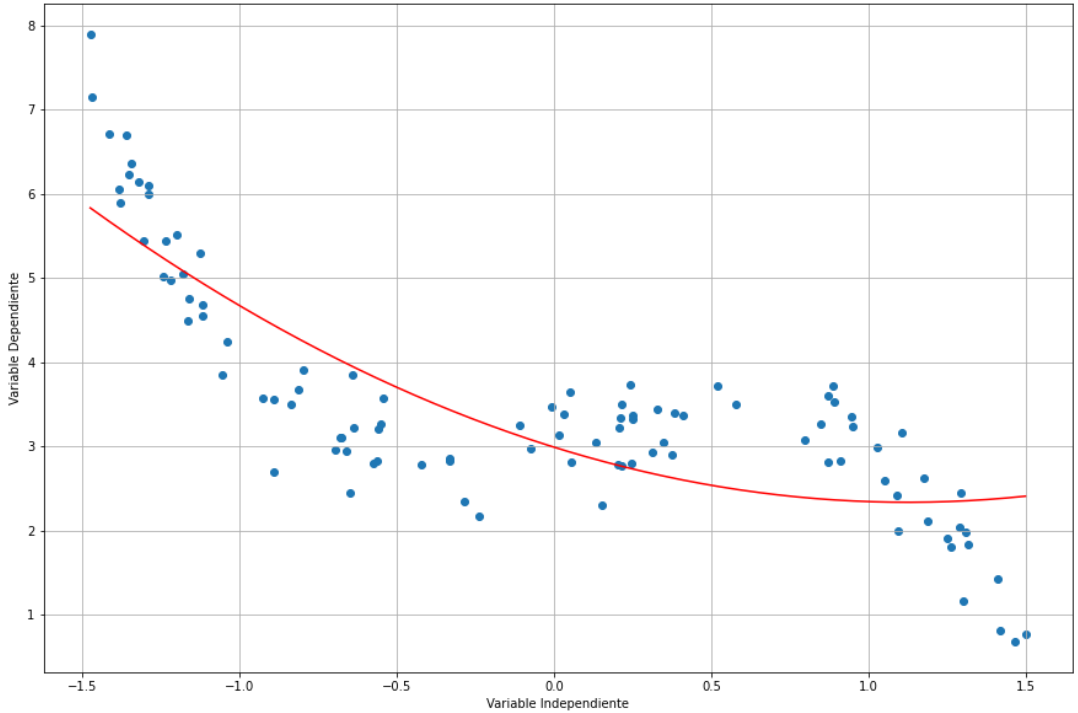
A una que tiene la forma:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

De modo que, si podemos encontrar el valor de los coeficientes que minimizan la misma función de costo, estaremos encontrando la mejor ecuación polinómica que ajusta los datos de entrenamiento.

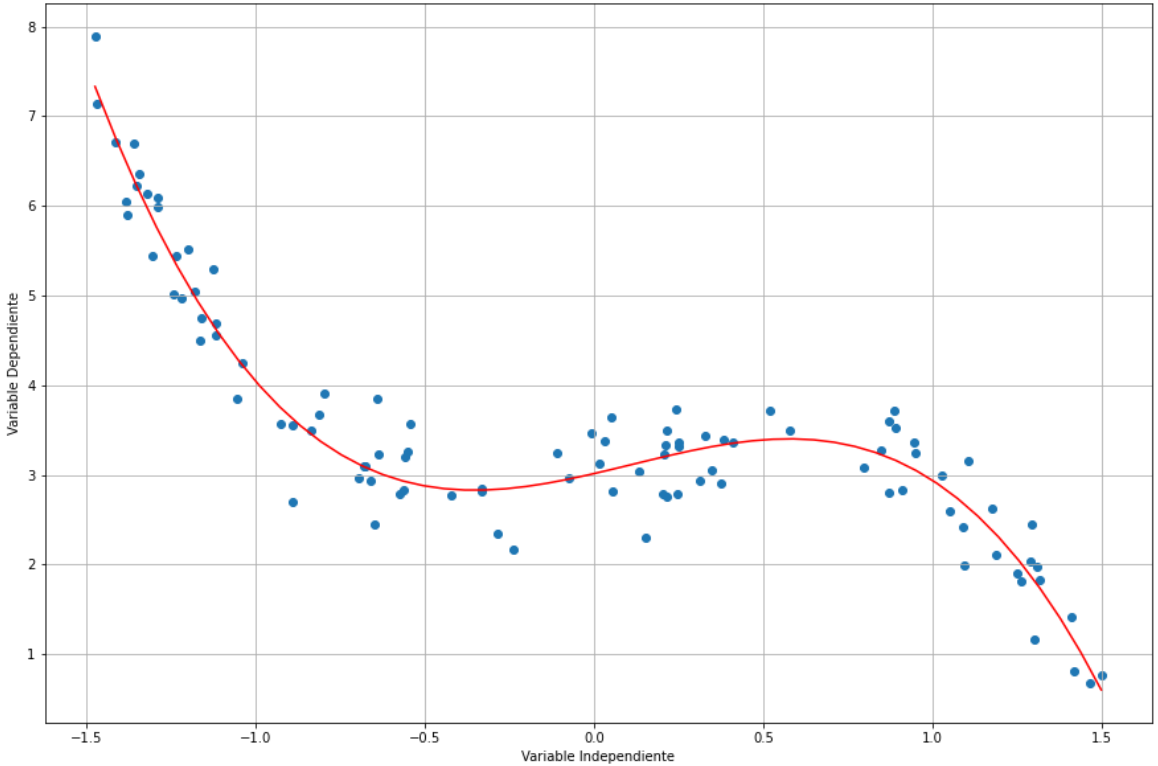
Regresión Polinómica

Por ejemplo, al aplicar a los datos de entrada originales una transformación polinómica de segundo orden, tenemos como ajuste a los datos lo siguiente:



Regresión Polinómica

Mientras que si la transformación es de tercer orden, obtenemos:



Regresión Polinómica

Este último caso demuestra que, en efecto, al incorporar potencias de la variable independiente original como nuevas variables a nuestros datos, podemos obtener un ajuste adecuado de los datos aplicando la misma teoría de regresión lineal vista antes. Si, para los casos mostrados, calculamos los valores del **RMSE** y **R^2** podremos observar el cómo los valores del error disminuyen cuando se aumenta el grado del polinomio, mientras que el **R^2** aumenta.

Ajuste	**RMSE**	**R^2**
Lineal	0.985	0.375
Polinómico Cuadrado	0.800	0.587
Polinómico Cúbico	0.374	0.909

En este caso, podemos decir que tanto el modelo lineal como el polinómico cuadrado están **subentrenados**. Es decir, los mismos no son capaces de reproducir el comportamiento general de los datos de entrenamiento.

Subentrenamiento y Sobreentrenamiento

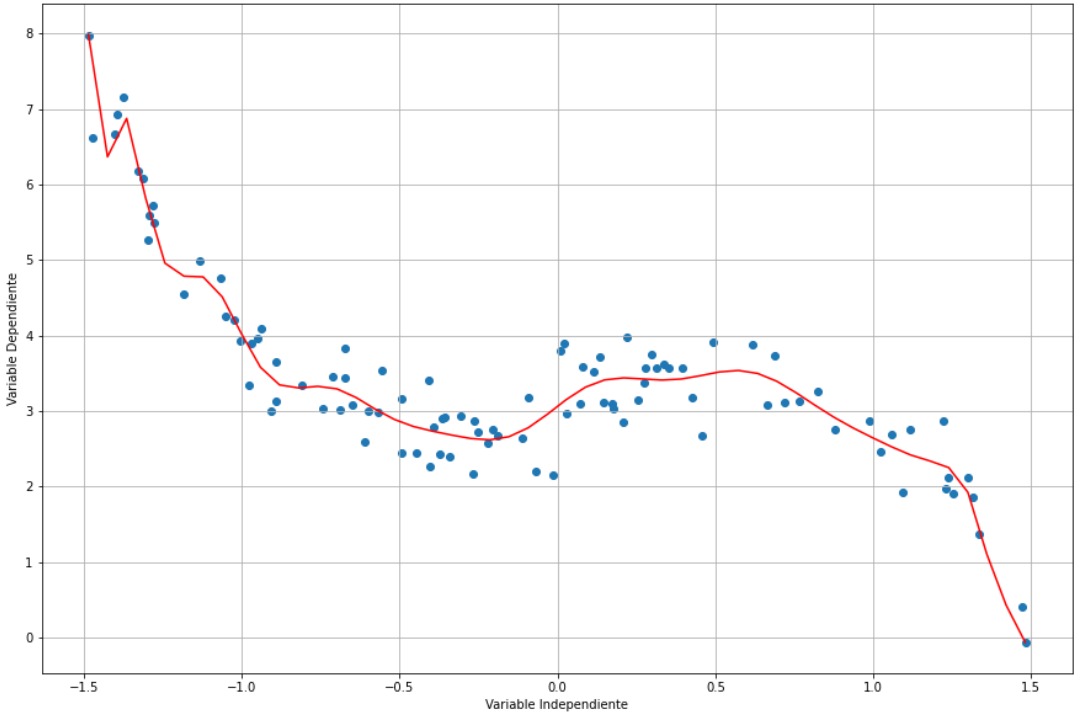
Se dice que un modelo tiene capacidad de **Generalización** cuando es capaz de ofrecer resultados de predicción con bajo error con datos distintos a los que se usaron durante el entrenamiento. Es decir, datos nuevos nunca antes vistos. Un modelo subentrenado, por lo tanto, presentará una mala capacidad de generalización.

Ahora bien, un modelo demasiado y grande y complejo puede sufrir del fenómeno opuesto, es decir, el **sobreentrenamiento**. En este caso, el modelo es tal que podría *memorizar* de manera 100% precisa los datos de entrenamiento, pero no es capaz de ofrecer buenos resultados con cualquier dato nuevo que no haya sido usado durante el entrenamiento.

Por ejemplo, veamos lo que ocurre si, para los datos mostrados en las gráficas anteriores, realizamos un ajuste polinómico de orden 25.

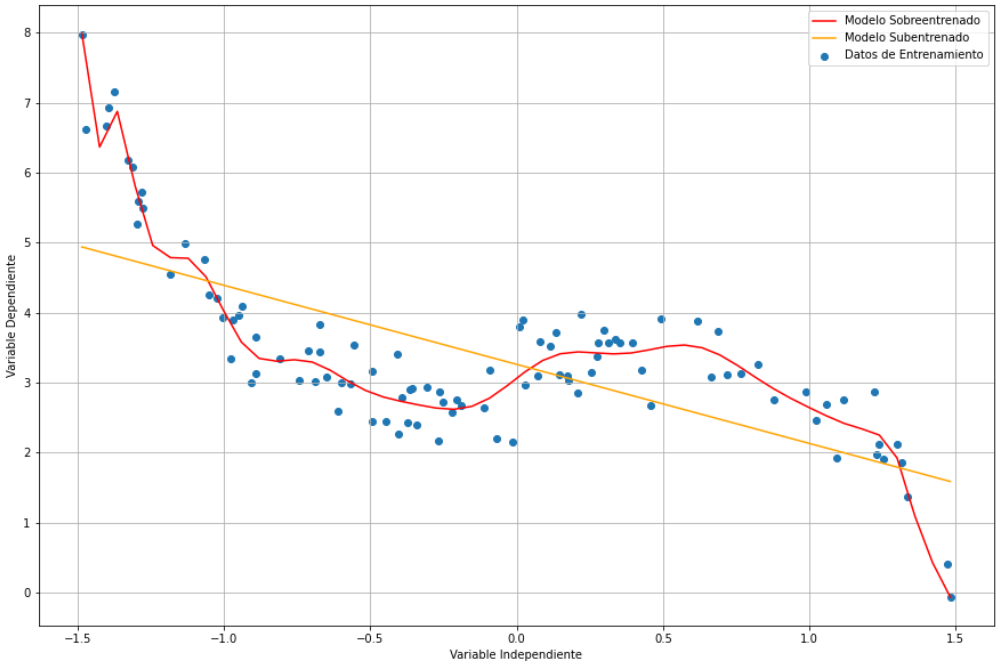
Subentrenamiento y Sobreentrenamiento

En este caso, el modelo ajusta los datos de una manera oscilante e intenta intersectar todos los puntos de entrenamiento.



Subentrenamiento y Sobreentrenamiento

Tenemos, finalmente, ejemplos de modelos **subentrenados** y **sobreentrenados**. Ninguno de los cuales serán capaces de generalizar los datos de entrenamiento a datos nuevos.



La compensación entre sesgo y varianza

Un resultado estadístico de los modelos de Machine Learning está en el hecho de que la capacidad de generalización de los mismos está determinada por:

Sesgos (Bias): cuando se asumen hipótesis incorrectas sobre los datos de entrenamiento. Por ejemplo, asumir que los datos son lineales cuando en realidad son cuadráticos. En consecuencia, un modelo con sesgo alto (high bias) es susceptible al subentrenamiento.

Varianza (variance): se refiere a la sensibilidad del modelo a pequeñas variaciones de los datos de entrenamiento. Un modelo con muchos grados de libertad tendrá alta varianza (high variance) y será susceptible al sobreentrenamiento.

Error irreducible: está determinado por los ruidos presentes en los datos, y solo se pueden mejorar limpiando los mismos.

De manera que, aumentar la complejidad de un modelo aumentará su varianza y reducirá sus sesgos, mientras que reducir la complejidad disminuirá la varianza pero aumentará el sesgo. A este concepto se le llama **compensación entre sesgo y varianza** (**bias/variance tradeoff**).

3er Notebook Práctico

Veamos ahora el cómo se implementan los modelos de Regresión Polinómica y entendamos mejor los conceptos de subentrenamiento y sobreentrenamiento.

¡Adelante con el Notebook!