Unidad 9. Introducción a Machine Learning

Por Sarahí Aguilar

Universidad Panamericana
ITISI 2016
BDP COM145
Primavera 2022

Agenda

- ¿Qué es el aprendizaje automático?
- Estimación de f
- Precisión contra interpretabilidad
- Aprendizaje supervisado contra no supervisado
- Problemas de regresión contra clasificación
- Calidad del ajuste
- Bias–variance tradeoff
- Introducción a scikit-learn

Agenda

¿Qué es el aprendizaje automático?

Estimación de f

Precisión contra interpretabilidad

Aprendizaje supervisado contra no supervisado

Problemas de regresión contra clasificación

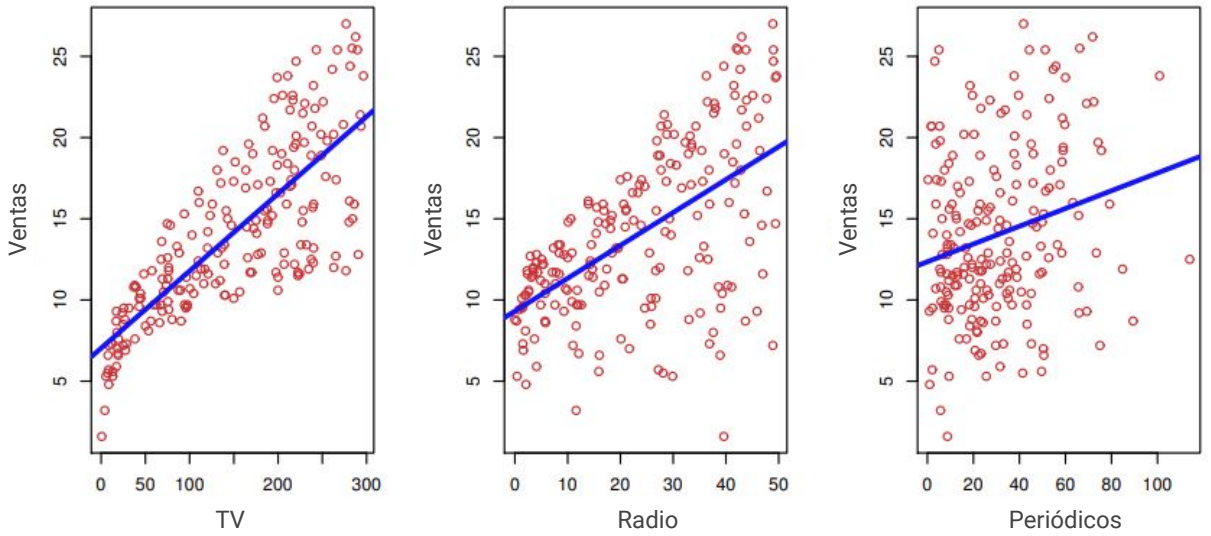
Calidad del ajuste

Bias–variance tradeoff

Introducción a scikit-learn

¿Cuál es la relación entre la publicidad y las ventas de un producto en particular?

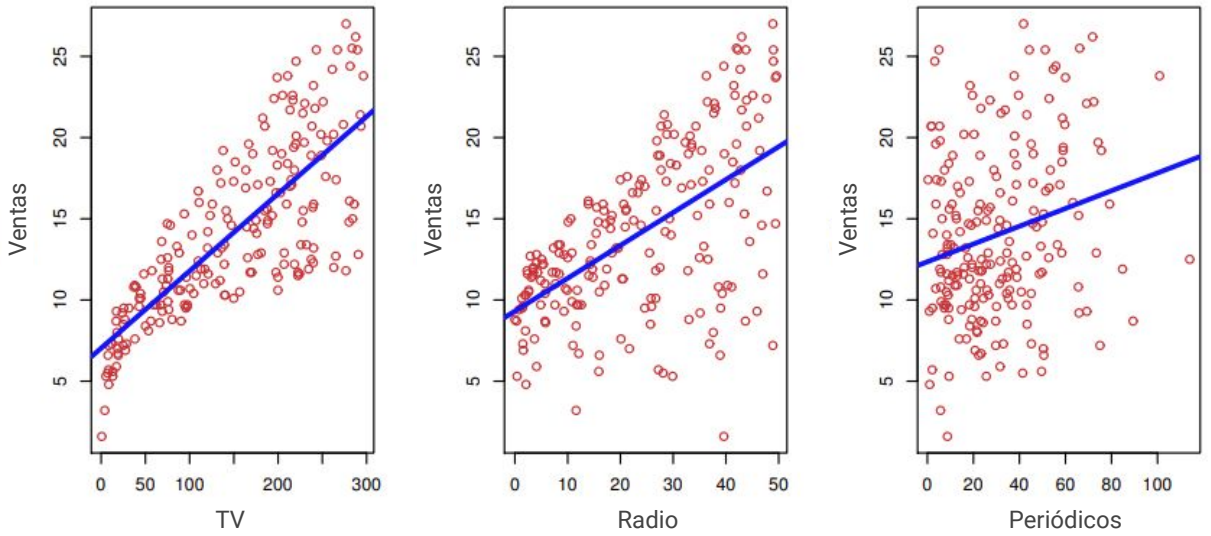
Si determinamos que existe una asociación entre la publicidad y las ventas,
entonces podemos ajustar los presupuestos de publicidad, aumentando así indirectamente las ventas.



Los gráficos muestran las ventas, en miles de unidades, en función de los presupuestos de televisión, radio y periódicos, en miles de dólares, para 200 mercados diferentes.
En cada gráfica se muestra el ajuste de mínimos cuadrados simple de las ventas a los presupuestos de cada medio.

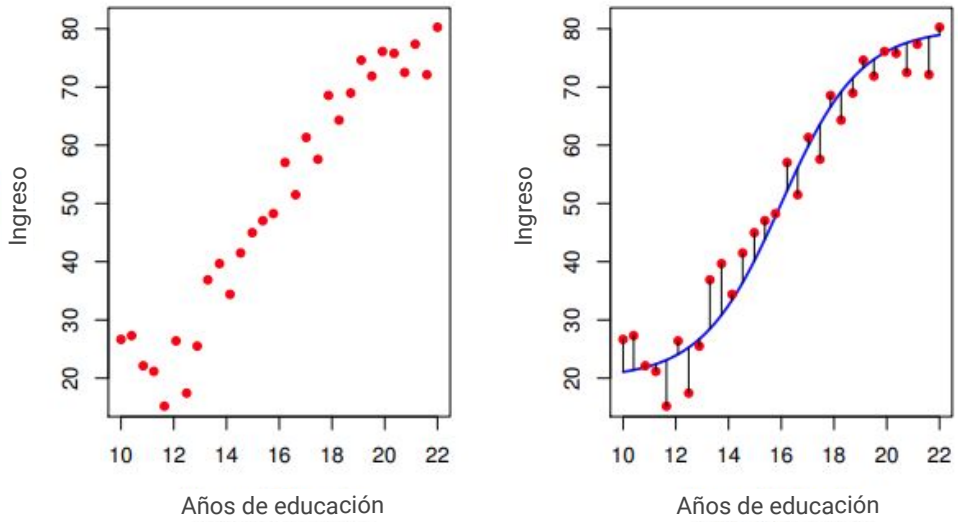
¿Cuál es la relación entre la publicidad y las ventas de un producto en particular?

De forma más general, supongamos que observamos una **respuesta cuantitativa Y y p diferentes predictores,** $X_1, X_2,...,X_p$.
Suponemos que existe alguna **relación entre** Y **y** $X = (X_1, X_2,...,X_p)$, que se puede escribir en la forma muy general $Y = f(x) + \varepsilon$.



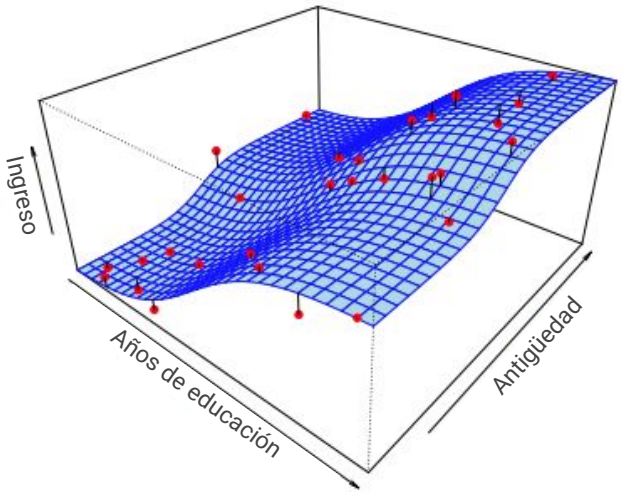
Los gráficos muestran las ventas, en miles de unidades, en función de los presupuestos de televisión, radio y periódicos, en miles de dólares, para 200 mercados diferentes.
En cada gráfica se muestra el ajuste de mínimos cuadrados simple de las ventas a los presupuestos de cada medio.

¿Cuál es la relación entre los años de educación y el ingreso?



Los puntos rojos son los valores observados de ingresos, en decenas de miles de dólares, y años de educación para 30 personas. La curva azul representa la verdadera relación subyacente entre los ingresos y los años de educación. Las líneas negras representan el error asociado con cada observación.

¿Cuál es la relación entre los años de educación y el ingreso?

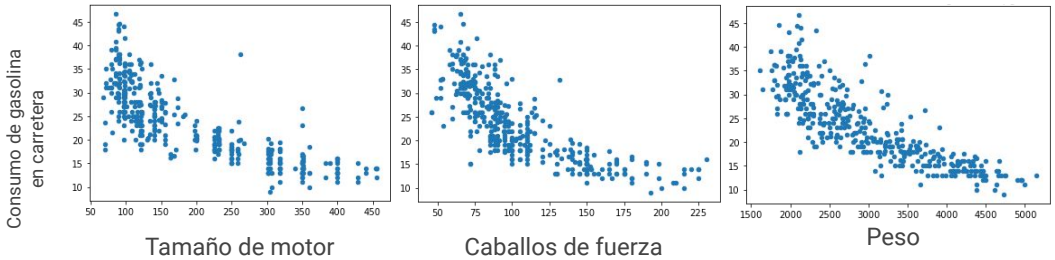


Los puntos rojos son los valores observados de ingresos, en decenas de miles de dólares, y años de educación y antigüedad para 30 personas. La curva azul representa la verdadera relación subyacente entre los ingresos, y los años de educación y la antigüedad. Las líneas negras representan el error asociado con cada observación.

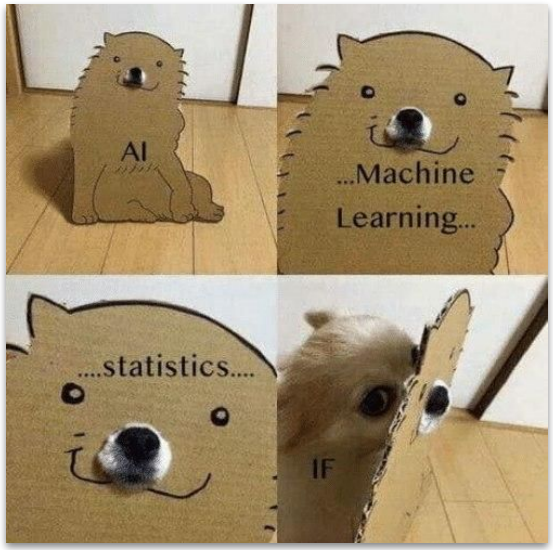
Dos tipos variables
(features, atributos o columnas)

En los vehículos con modelos que tuvieron un nuevo lanzamiento cada año entre 1999 y 2008,
¿es posible predecir el consumo de combustible de gasolina en carretera con otras de sus características físicas?

Vehículo Identificador único	**Consumo de gasolina en carretera (Millas por galón)** Variable dependiente	**Tamaño de motor (Litros)** Variable independiente	**Caballos de fuerza** Variable independiente	**Peso (Kg)** Variable independiente
0	18	307	130	3,504
1	15	150	165	3,693
2	18	318	150	3,436
...	...	...	...	...



En esencia, el aprendizaje automático se refiere a un conjunto de enfoques para **estimar** f .



Agenda

¿Qué es el aprendizaje automático?

Estimación de f

Precisión contra interpretabilidad

Aprendizaje supervisado contra no supervisado

Problemas de regresión contra clasificación

Calidad del ajuste

Bias–variance tradeoff

Introducción a scikit-learn

Why

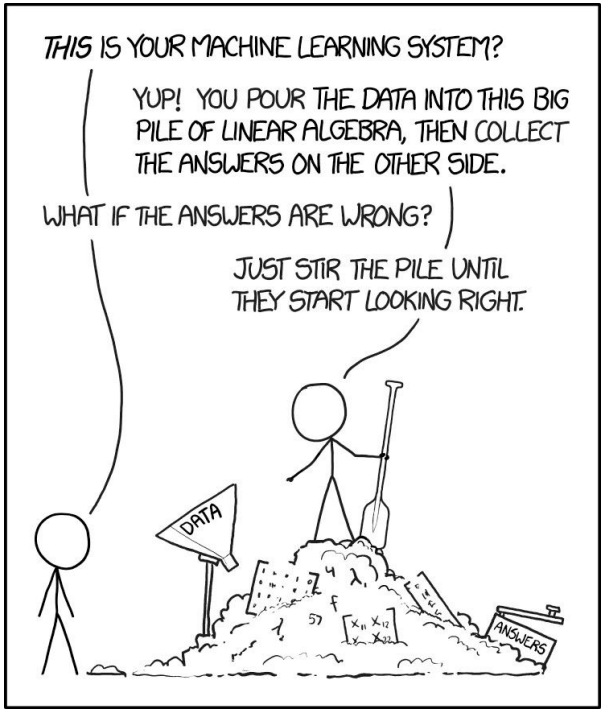
Predicción o inferencia

Predicción

$$\hat{Y} = \hat{f}(X)$$

No estamos necesariamente interesados en la forma exacta de $\hat{f}$.

La precisión de $\hat{Y}$ cómo predicción de Y depende de dos cantidades: el error reducible y error irreducible.



Why

Predicción o inferencia

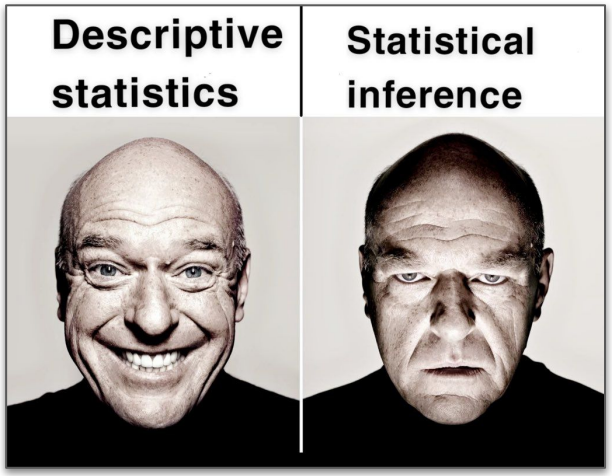
Inferencia

$$\hat{Y} = \hat{f}(X)$$

¿Qué predictores están asociados con la respuesta?

¿Cuál es la relación entre la respuesta y cada predictor?

¿Se puede resumir adecuadamente la relación entre Y y cada predictor usando una ecuación lineal, o la relación es más complicada?



What

Conjunto de datos de entrenamiento

Entrenamiento
Muestra aleatoria del 70-80%

Prueba
Muestra aleatoria del 20-30%

Validación
Muestra aleatoria del 10-20%

Identificador único	Variable dependiente (Y)	Variable independiente (X_1)	Variable independiente (X_2)	Variable independiente (X_3)

How

Modelos paramétricos y no paramétricos

- **Modelos paramétricos**

1. Asumimos la forma de la función f.

2. Utilizamos el conjunto de datos de entrenamiento para entrenar/ajustar al modelo.

Ejemplo

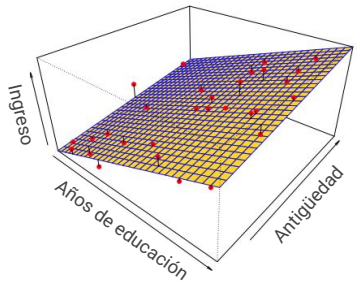
Asumimos que tiene una forma lineal:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Utilizamos el conjunto de datos de entrenamiento para estimar los parámetros β (coeficientes), tal que:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

mediante mínimos cuadrados ordinarios.



How

Modelos paramétricos y no paramétricos

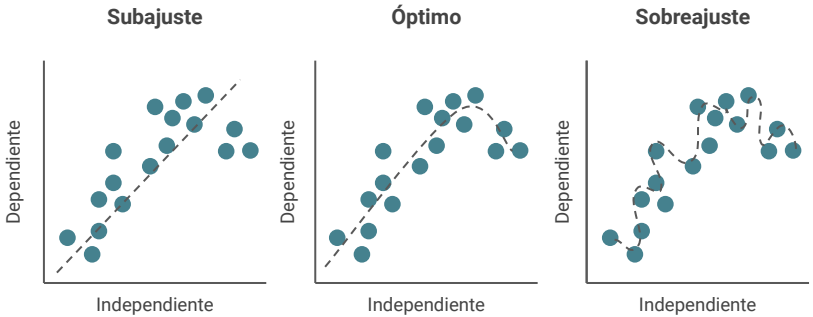
- **Modelos paramétricos**

Ventaja principal

Asumir una forma paramétrica para f **simplifica** el problema de estimar f porque generalmente es mucho más fácil estimar un conjunto de parámetros, que ajustar un función totalmente arbitraria f.

Desventaja principal

Si a forma de la función f elegida está **demasiado lejos de la verdadera** forma de la función f, entonces nuestra estimación será pobre. Podemos tratar de abordar este problema eligiendo modelos más flexibles que requieren de la **estimación un mayor número de parámetros**. No obstante, estos modelos más complejos pueden conducir al **sobreajuste**.



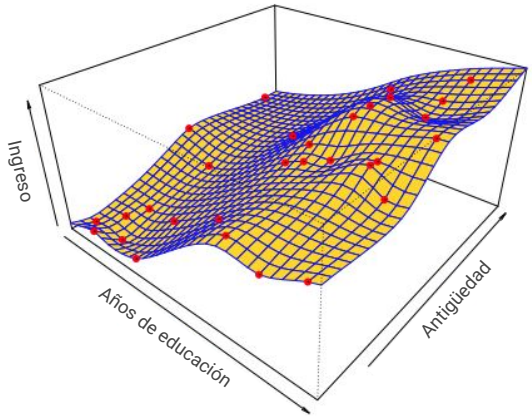
How

Modelos paramétricos y no paramétricos

- **Modelos no paramétricos**

1. **No** asumimos la forma de la función f.

2. Utilizamos el conjunto de datos de entrenamiento (con generalmente un **mayor número de observaciones** que el necesario para un modelo paramétrico) para entrenar al modelo (con generalmente un **mayor número de parámetros** que el necesario para un modelo paramétrico).



¿Por qué elegiríamos un **modelo más restrictivo** en lugar de un **modelo muy flexible**?

Agenda

¿Qué es el aprendizaje automático?

Estimación de f

Precisión contra interpretabilidad

Aprendizaje supervisado contra no supervisado

Problemas de regresión contra clasificación

Calidad del ajuste

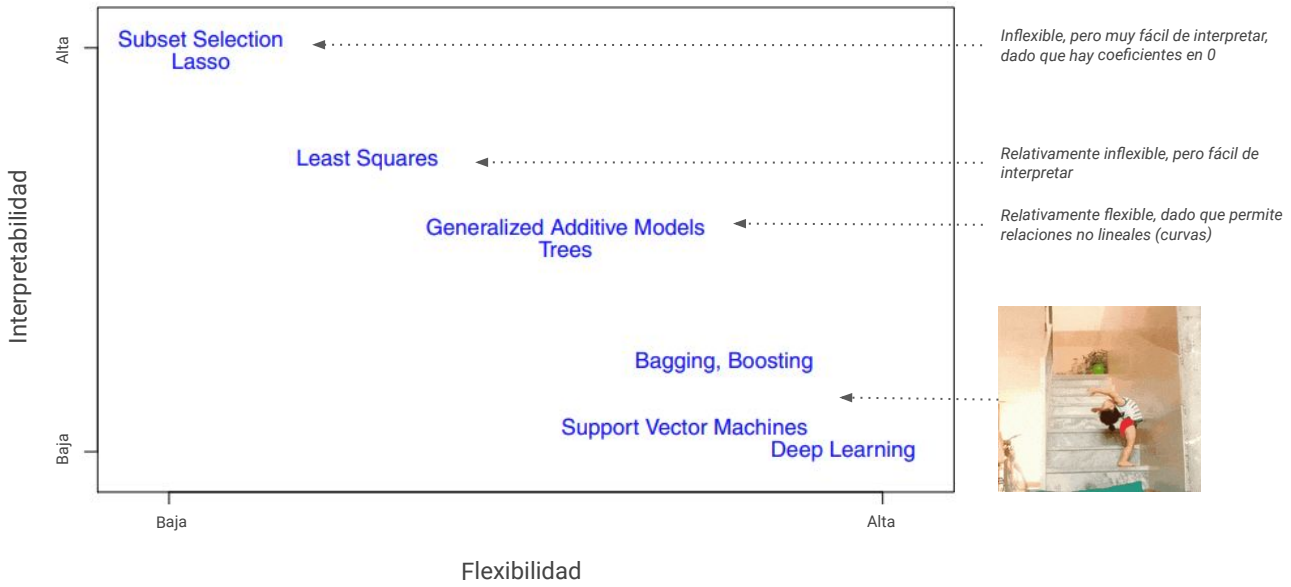
Bias–variance tradeoff

Introducción a scikit-learn

¿Por qué elegiríamos un modelo más restrictivo en lugar de un modelo muy flexible?

Si estamos interesados principalmente en la **inferencia**, entonces los **modelos más restrictivos** son mucho **más interpretables**.

Los **modelos muy flexibles**, pueden conducir a estimaciones tan complicadas de f que es **difícil entender cómo se asocia cada predictor con la respuesta**.



Agenda

¿Qué es el aprendizaje automático?

Estimación de f

Precisión contra interpretabilidad

Aprendizaje supervisado contra no supervisado

Problemas de regresión contra clasificación

Calidad del ajuste

Bias–variance tradeoff

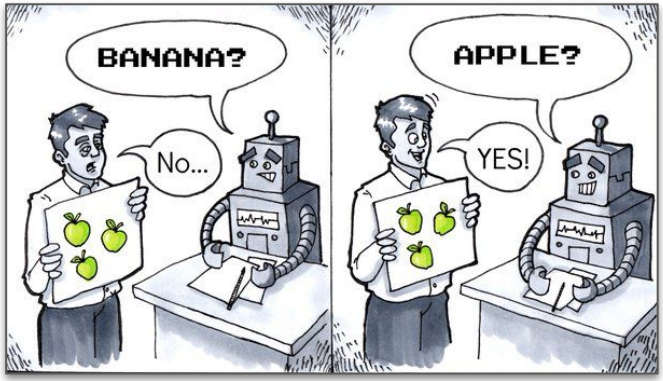
Introducción a scikit-learn

Dos tipos aprendizaje

Supervisado

Para cada observación de la(s) variable(s) predictora(s) x_i, $i = 1,...,n$, hay una variable de respuesta asociada y_i.

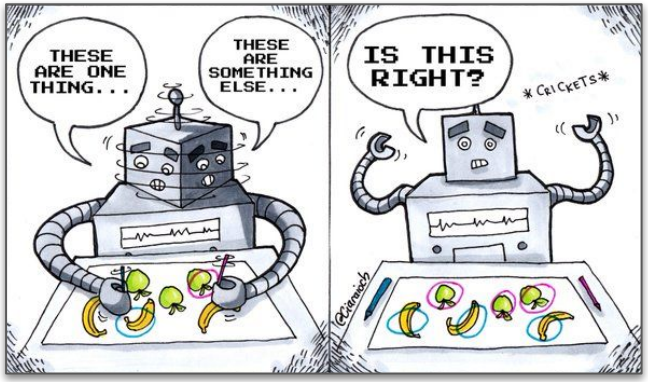
Queremos entrenar/ajustar un modelo que relacione las variables predictoras a la variable de respuesta.



No supervisado

Para cada observación de la(s) variable(s) predictora(s) x_i, $i = 1,...,n$, **no** hay una variable de respuesta asociada y_i.

Queremos entrenar/ajustar un modelo que relacione las variables predictoras a una **nueva** variable de respuesta.



Agenda

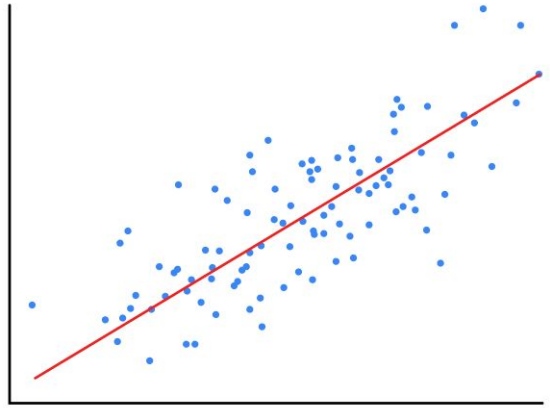
- ¿Qué es el aprendizaje automático?
- Estimación de f
- Precisión contra interpretabilidad
- Aprendizaje supervisado contra no supervisado
- **Problemas de regresión contra clasificación**
- Calidad del ajuste
- Bias–variance tradeoff
- Introducción a scikit-learn

Dos tipos de datos

Dos tipos de problemas

Regresión

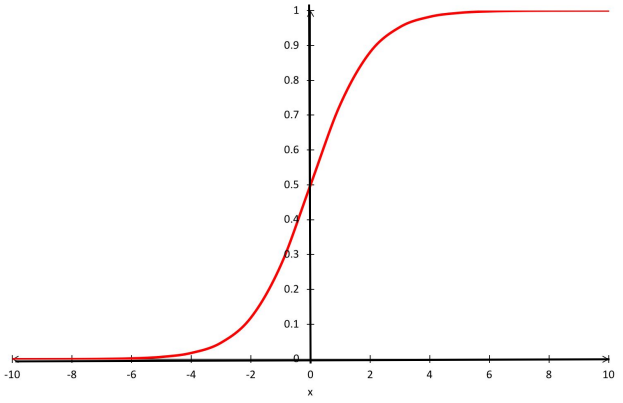
Con una variable de respuesta **numérica**.



Regresión lineal

Clasificación

Con una variable de respuesta **categórica**.



Regresión logística

Agenda

¿Qué es el aprendizaje automático?

Estimación de f

Precisión contra interpretabilidad

Aprendizaje supervisado contra no supervisado

Problemas de regresión contra clasificación

Calidad del ajuste

Bias–variance tradeoff

Introducción a scikit-learn

Calidad del ajuste

No existe un modelo universal que resuelva todos los problemas de Machine Learning dado cualquier conjunto de datos.

Resulta importante decidir **qué modelo produce el mejor ajuste** para conjunto de datos dado.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Para evaluar la calidad del ajuste de un modelo en un conjunto de datos dado, necesitamos una forma de **cuantificar** hasta qué punto el **valor de respuesta pronosticado** para una observación determinada se acerca al **valor de respuesta real** para esa observación.

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$
$$\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$$
$$\hat{f}(x_i) \approx y_i$$

Entrenamiento
Muestra aleatoria del 70-80%

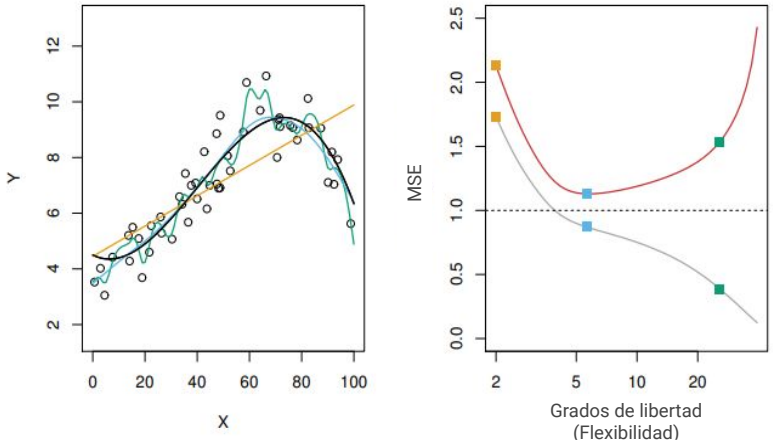
Prueba
Muestra aleatoria del 20-30%

Validación
Muestra aleatoria del 10-20%

Identificador único	Variable dependiente (Y)	Variable independiente (X_1)	Variable independiente (X_2)	Variable independiente (X_3)
		Evaluación de la calidad del ajuste		

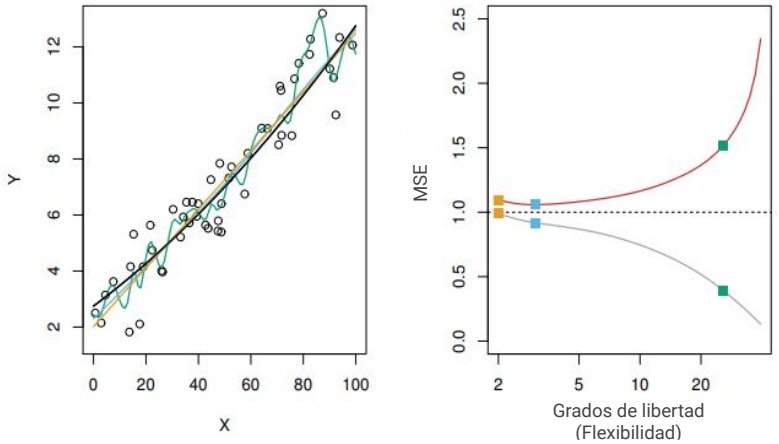
Calidad del ajuste

Ejemplo 1



MSE en conjunto de datos de prueba
MSE en conjunto de datos de entrenamiento

Ejemplo 2



MSE en conjunto de datos de prueba
MSE en conjunto de datos de entrenamiento

Agenda

- ¿Qué es el aprendizaje automático?
- Estimación de f
- Precisión contra interpretabilidad
- Aprendizaje supervisado contra no supervisado
- Problemas de regresión contra clasificación
- Calidad del ajuste
- **Bias–variance tradeoff**
- Introducción a scikit-learn

Bias

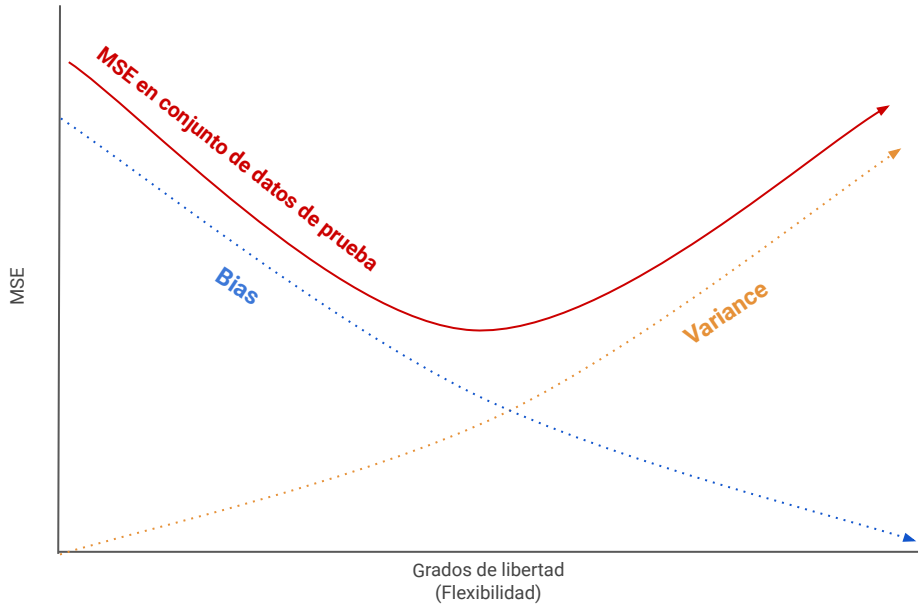
Error de suposiciones erróneas en el modelo.

Un alto bias puede hacer que un modelo **omita las relaciones relevantes** entre las variables predictoras y la variable respuesta (subajuste).

Variance

Error de sensibilidad a pequeñas fluctuaciones en el conjunto de datos de entrenamiento.

Una variación alta puede resultar de un algoritmo que **modela el ruido aleatorio** en el conjunto de datos de entrenamiento (sobreajuste).



Agenda

- ¿Qué es el aprendizaje automático?
- Estimación de f
- Precisión contra interpretabilidad
- Aprendizaje supervisado contra no supervisado
- Problemas de regresión contra clasificación
- Calidad del ajuste
- Bias–variance tradeoff
- **Introducción a scikit-learn**

scikit-learn es una biblioteca para aprendizaje automático de software libre para el lenguaje de programación Python.

- Herramientas simples y eficientes para el análisis predictivo de datos.
- Basado en NumPy, SciPy y matplotlib.
- Accesible para todos y reutilizable en varios contextos.
- Código abierto, utilizable comercialmente.

¿Cómo seleccionar el estimador *correcto*?