Regresión lineal

INF-396

Prof: Juan G. Pavez S.

Aprendizaje Supervisado

Dado un conjunto de datos de entrenamiento de tamaño n compuesto de pares de entrada-salida $\{x_{1:n}, y_{1:n}\}$, donde cada entrada $x_i \in R^{1xd}$ es un vector con d-atributos (features, características). La entrada también se conoce como predictor o covariantes. La salida referida, como target, muchas veces unidimensional $x \in R$.

Aprendizaje Supervisado

Target variable

$$(x_i, y_i)$$

Input variables

$$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})$$

Features

Aprendizaje Supervisado

Caso de entrenamiento (ejemplo)

Target variable

$$(x_i, y_i)$$

Input variables

$$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})$$

Features

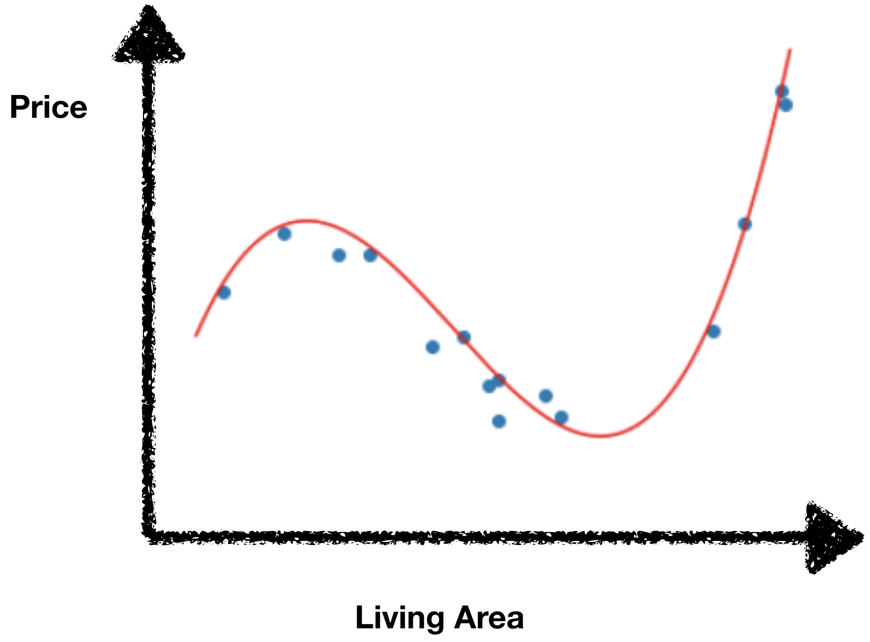
Dataset

$$D = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ \vdots & \\ x_N & y_N \end{bmatrix}$$

Aprendizaje Supervisado

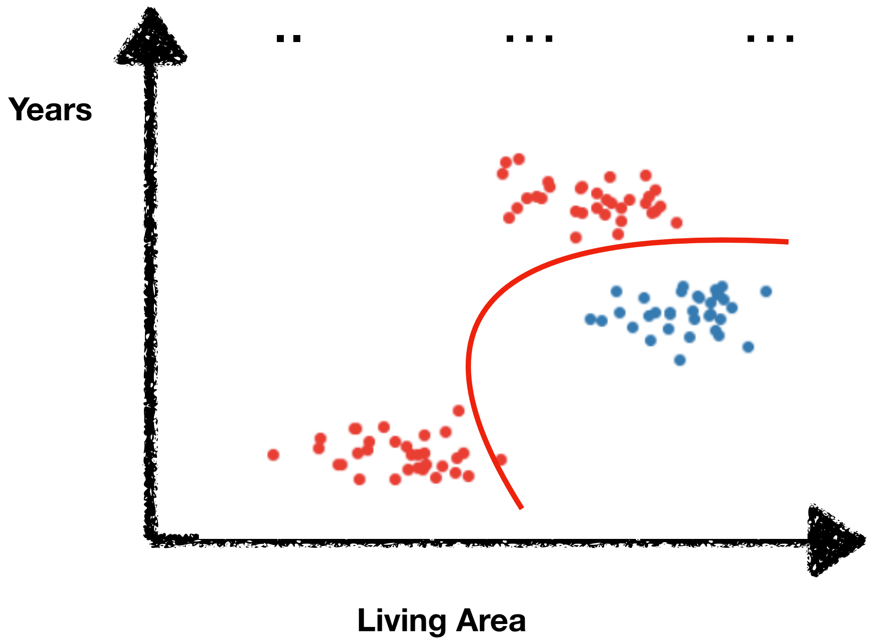
Regresión (Price prediction)

Living Area (mt2)	Price(100$)
2104	400
1600	330
1414	232
…	….



Clasificación (Insurance classification)

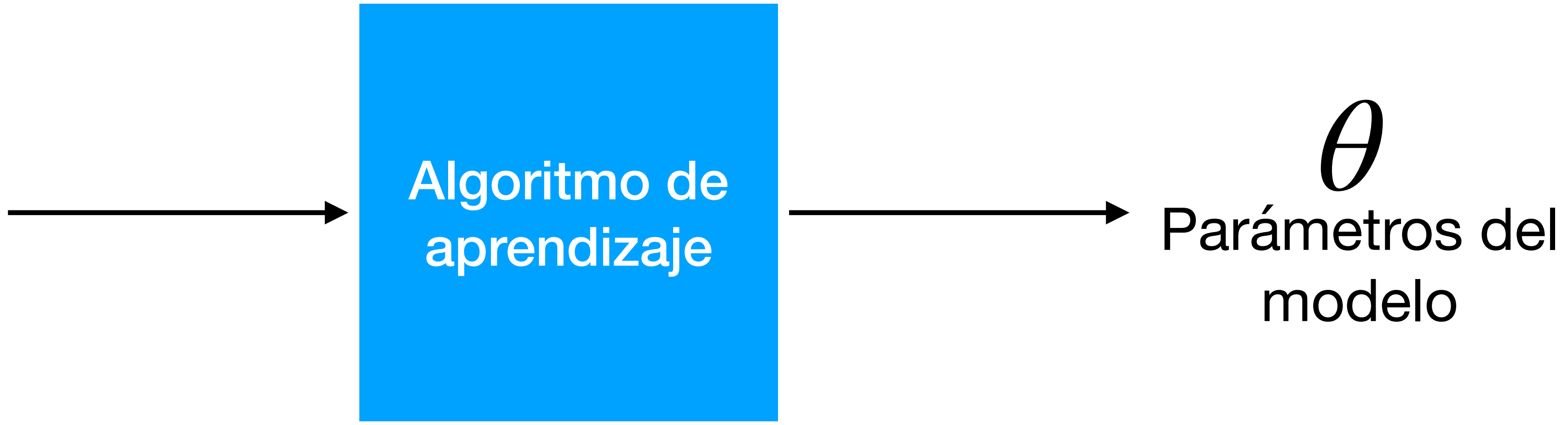
Living Area (mt2)	Years	Home insurance classification
2104	40	B
1600	3	A
1414	23	B
1344	40	B
..	…	…



Aprendizaje Supervisado

Algoritmo de Aprendezije

Living Area (mt2)	Price(100$)
2104	400
1600	330
1414	232
…	….



Aprendizaje Supervisado

Caso de entrenamiento (ejemplo)

Dataset

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

$$y = \begin{bmatrix} 5 \\ 25 \\ 22 \\ 18 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Aprendizaje Supervisado

Caso de entrenamiento (ejemplo) **Dataset**

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

$$y = \begin{bmatrix} 5 \\ 25 \\ 22 \\ 18 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} 2 \\ 22 \\ 16.5 \\ 18.5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix} \qquad \theta = \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}$$

Regresión

- **Regresión Lineal**
  La respuesta es una función lineal de la entrada

$$y(x) = w^T x + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

  Normalmente se asume

$$\epsilon \sim N(0, \sigma^2)$$

  Si ese es el caso entonces

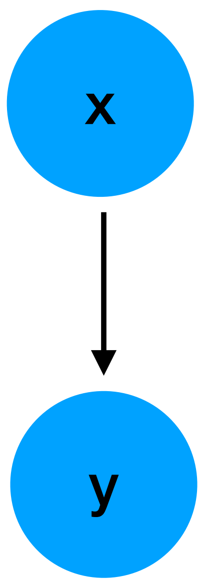
$$p(y \mid x, w) = N(\mu(x), \sigma^2(x))$$

  donde

$$\mu(x) = w^T x \qquad \sigma^2(x) = \sigma^2 \qquad \theta = (w, \sigma^2)$$

Regresión

- **Regresión Lineal**
  - Queremos modelar la probabilidad condicional $p(y\,|\,x, \theta)$



$$X = \sigma Z + m, Z \sim N(0,1)$$

$$P(a < X < b) = P(a < aZ + m < b)$$

$$= P\left(\frac{a-m}{\sigma} < Z < \frac{b-m}{\sigma}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{(a-m)/\sigma}^{(b-m)/\sigma} e^{-z^2} dz$$

$$x = \sigma z + m$$

$$P(a < X < b) = \frac{1}{\sqrt{2\pi\sigma}} \int_a^b e^{-(x-m)^2/\sigma^2} dx$$

$$y = w^T x + \epsilon$$

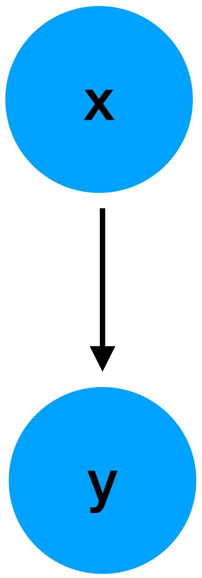
$$\epsilon \sim N(0,\sigma^2)$$

$$p(y\,|\,x, \theta) \sim N(w^T x, \sigma^2)$$

Regresión

- **Regresión Lineal**

  - Queremos modelar la probabilidad condicional $p(y \mid x, \theta)$



$$p(y \mid x, \theta) = w^T x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$p(y \mid x, \theta) \sim N(w^T x, \sigma^2)$$

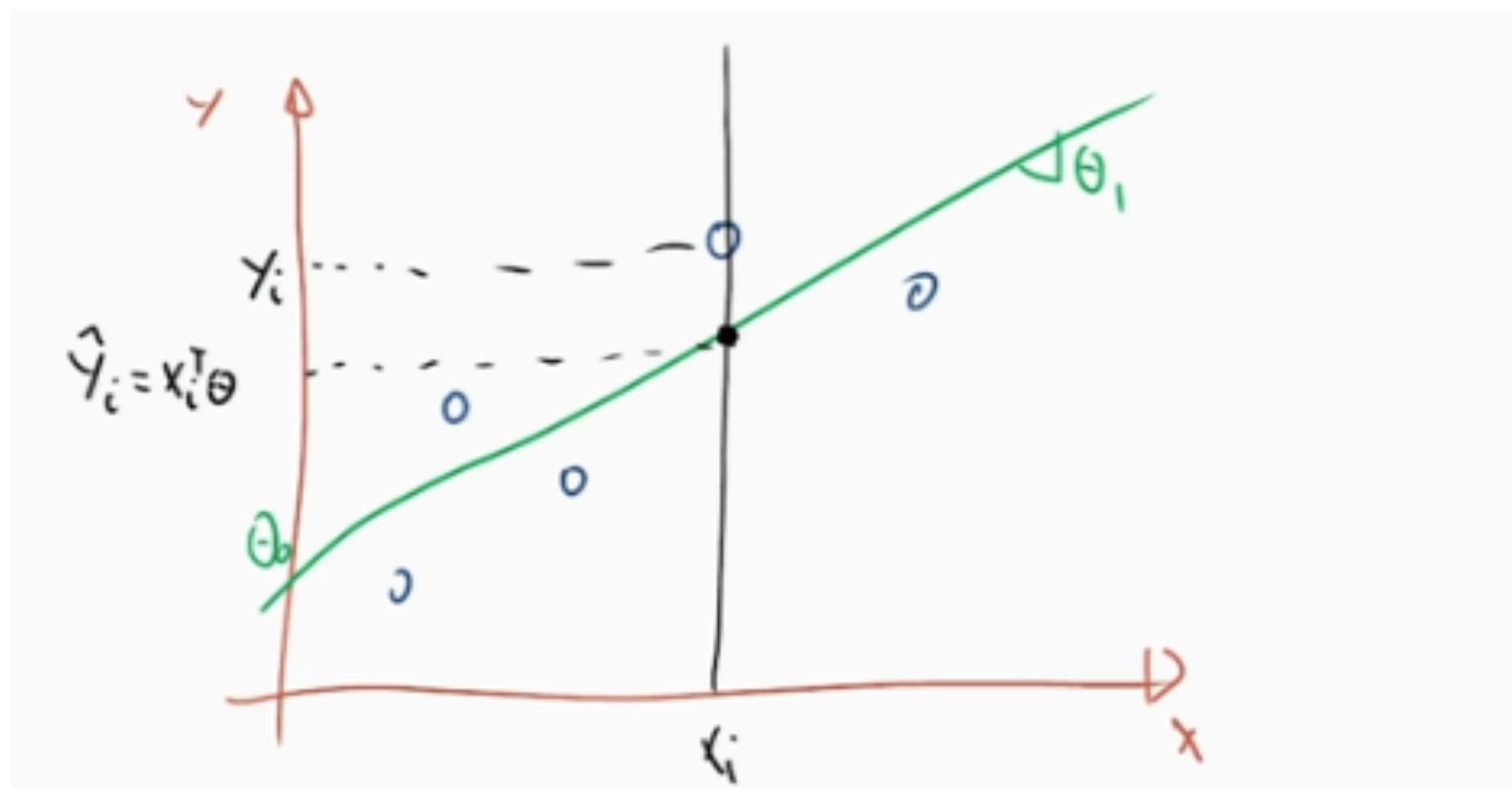
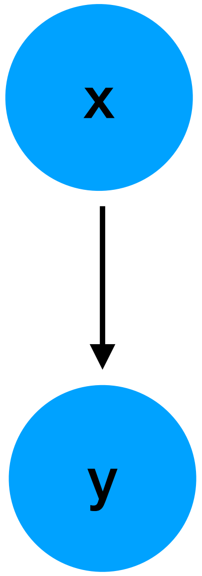


Imagen clase curso Nando de Freitas

Regresión

- **Regresión Lineal**
  - Queremos modelar la probabilidad condicional $p(y\,|\,x,\theta)$



$$p(y\,|\,x,\theta) = w^T x + \epsilon$$

$$\epsilon \sim N(0,\sigma^2)$$

$$p(y\,|\,x,\theta) \sim N(w^T x, \sigma^2)$$

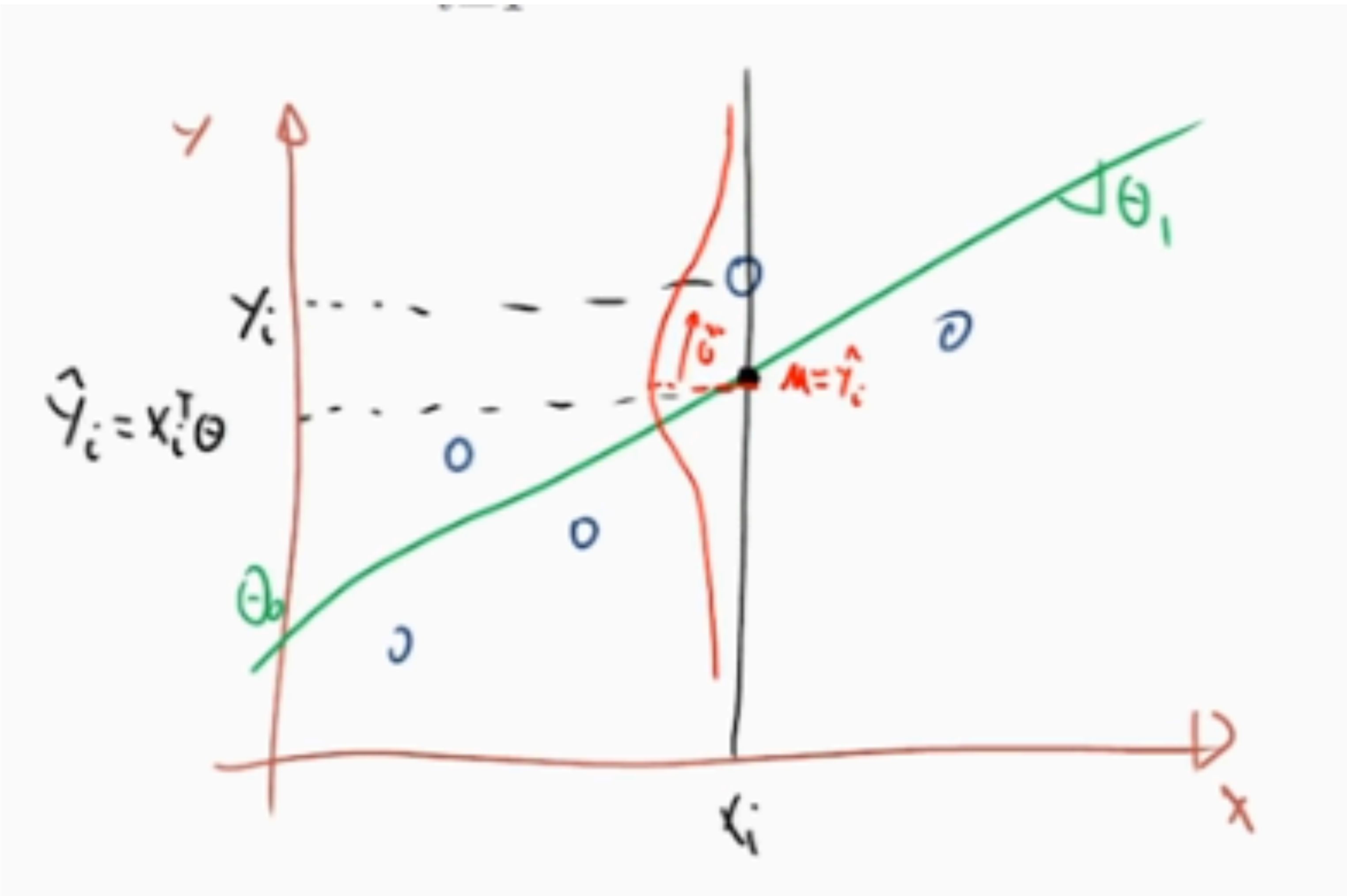


Imagen clase curso Nando de Freitas

Regresión

- **Regresión Lineal**
  La respuesta es una función lineal de la entrada

$$p(y \mid x, \theta) = N(\mu(x), \sigma^2(x))$$

Considerar el caso 1-dimensional:

$$\mu(x) = w_0 + w_1 x = w^T x \qquad x = (1, x)$$

bias intercept

Regresión

- **Regresión Lineal**
  Podemos estimar $\theta = w$ usando Maximum Likelihood.

$$l(\theta) = \sum_{i=1}^{N} \log p(y_i \,|\, x_i, \theta)$$

$$= \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2 \right] \right]$$

Regresión

- **Regresión Lineal**
  Podemos estimar $\theta = w$ usando Maximum Likelihood.

$$l(\theta) = \sum_{i=1}^{N} \log p(y_i \,|\, x_i, \theta)$$

$$= \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right] \right]$$

$$= -\sum_{i=1}^{N} \frac{1}{2\sigma^2} (y_i - w^T x_i)^2 + \sum_{i=1}^{N} \log \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}}$$

Regresión

- **Regresión Lineal**
  Podemos estimar $\theta = w$ usando Maximum Likelihood.

$$l(\theta) = \sum_{i=1}^{N} \log p(y_i \,|\, x_i, \theta)$$

$$= \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2 \right] \right]$$

$$= -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2} \log(2\pi\sigma^2)$$

Regresión

- **Regresión Lineal**
  Podemos estimar $\theta = w$ usando Maximum Likelihood.

$$l(\theta) = -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2}\log(2\pi\sigma^2)$$

- RSS es la **residual sum of squares o suma de errores cuadráticos**

$$RSS(w) = ||\epsilon||_2^2 = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

- Conocido como SSE (sum of squared errors) y
  SSE/N = MSE

Regresión

- **Regresión Lineal**
  MLE para $\theta = w$ es obtenido minimizando el RSS, por eso se conoce como **mínimos cuadrados** (el modelo puede ser derivado completamente sin interpretación probabilística).

- Para derivar el MLE

$$NLL(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$
$$= \frac{1}{2}(y - Xw)^T(y - Xw)$$

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

Regresión

$$NLL(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

$$X \in R^{Nxm}$$

$$w \in R^{mx1}$$

$$y \in R^{Nx1}$$

$$= \frac{1}{2}(y - Xw)^T(y - Xw)$$

$$= \frac{1}{2}(y^T - (Xw)^T)(y - Xw)$$

$$= \frac{1}{2}(y^T y - y^T Xw - (Xw)^T y + w^T X^T Xw)$$

Regresión

$$NLL(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

$$= \frac{1}{2}(y - Xw)^T(y - Xw)$$

$$= \frac{1}{2}(y^T - (Xw)^T)(y - Xw)$$

$$= \frac{1}{2}(y^T y - y^T Xw - (Xw)^T y + w^T X^T Xw)$$

$$= \frac{1}{2}w^T(X^T X)w - y^T(Xw)$$

$$= \frac{1}{2}w^T(X^T X)w - w^T(X^T y)$$

Regresión

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - w^T (X^T y)$$

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

$$\frac{\partial NLL(w)}{\partial w} = ?$$

Regresión

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - (X^T y)^T w$$

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

$$\frac{\partial NLL(w)}{\partial w} = ?$$

$$\frac{\partial (b^T a)}{\partial a} = \frac{\partial (b_1 a_1 + \ldots b_N a_N)}{\partial a}$$

$$= \begin{bmatrix} \frac{\partial b^T a}{\partial a_1} \\ \cdots \\ \frac{\partial b^T a}{\partial a_N} \end{bmatrix} = b$$

Regresión

$$X \in R^{Nxm}$$

$$w \in R^{mx1}$$

$$y \in R^{Nx1}$$

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - (X^T y)^T w$$

$$\frac{\partial NLL(w)}{\partial w} = \ldots + X^T y$$

$$\frac{\partial (a^T A a)}{\partial a} = (A + A^T) a$$

Regresión

$$X \in R^{Nxm}$$

$$w \in R^{mx1}$$

$$y \in R^{Nx1}$$

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - (X^T y)^T w$$

$$\frac{\partial NLL(w)}{\partial w} = \frac{1}{2} (X^T X + X^T X) w + X^T y$$

$$\frac{\partial (a^T A a)}{\partial a} = (A + A^T) a$$

Regresión

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - (X^T y)^T w$$

$$\frac{\partial NLL(w)}{\partial w} = \frac{1}{2}(X^T X + X^T X) w + X^T y$$

$$\frac{\partial (a^T A a)}{\partial a} = (A + A^T) a$$

5 Puntos Tarea 2

Regresión

$$X \in R^{Nxm}$$
$$w \in R^{mx1}$$
$$y \in R^{Nx1}$$

$$NLL(w) = \frac{1}{2} w^T (X^T X) w - w^T (X^T y)$$

$$\frac{\partial NLL(w)}{\partial w} = X^T X w - X^T y$$

$$X^T X w - X^T y = 0$$

$$w = (X^t X)^{-1} X^T y$$

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

Regresión

- **Regresión Lineal**

$$X^T X w = X^T y$$ (Ecuaciones normales)

- Y la solución conocida como **Ordinary Least Squares (OLS)** se obtiene resolviendo

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

Regresión

- **Interpretación Geométrica**

- Asumiendo $N > m$ (más ejemplos que features). Sea x_j la columna j que es un vector de tamaño N. Además y también es un vector de tamaño N. Por ejemplo $N = 3$ y $D = 2$

$$X = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix} \qquad y = \begin{pmatrix} 8.98 \\ 0.61 \\ 1.77 \end{pmatrix}$$

- Entonces buscamos el vector $\hat{y}$ que cumpla:

Regresión

- **Interpretación Geométrica**

$$X = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix} \qquad y = \begin{pmatrix} 8.98 \\ 0.61 \\ 1.77 \end{pmatrix}$$

- Entonces buscamos el vector $\hat{y}$ que cumpla:

$$argmin_{\hat{y} \in span(\{x_1, \ldots, x_D\})} || y - \hat{y} ||_2$$

- Dado que $\hat{y} \in span(X)$ existe w tal que $\hat{y} = w_1 x_1 + \ldots w_m x_m = Xw$

Regresión

- **Interpretación Geométrica**

$$X = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix} \qquad y = \begin{pmatrix} 8.98 \\ 0.61 \\ 1.77 \end{pmatrix}$$

- Entonces buscamos el vector $\hat{y}$ que cumpla:

$$argmin_{\hat{y} \in span(\{x_1,\ldots,x_D\})} ||y - \hat{y}||_2$$

- Esto se logra cuando el vector residual $y - \hat{y}$ es ortogonal a cada columna de X

$$X^T(y - Xw) = 0 \implies w = (X^TX)^{-1}X^Ty$$

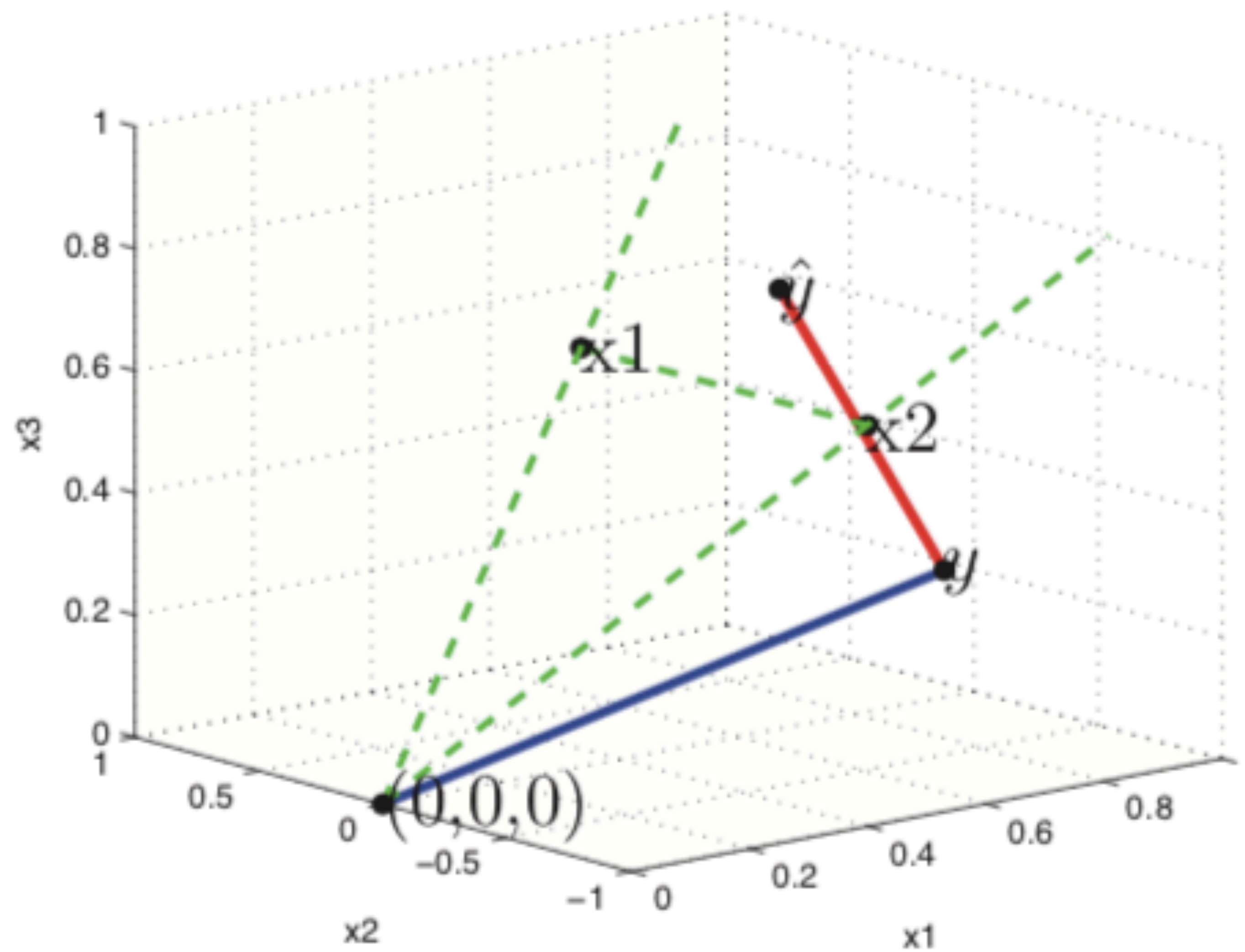
Regresión

- **Interpretación Geométrica**

$$X = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix} \quad y = \begin{pmatrix} 8.98 \\ 0.61 \\ 1.77 \end{pmatrix}$$

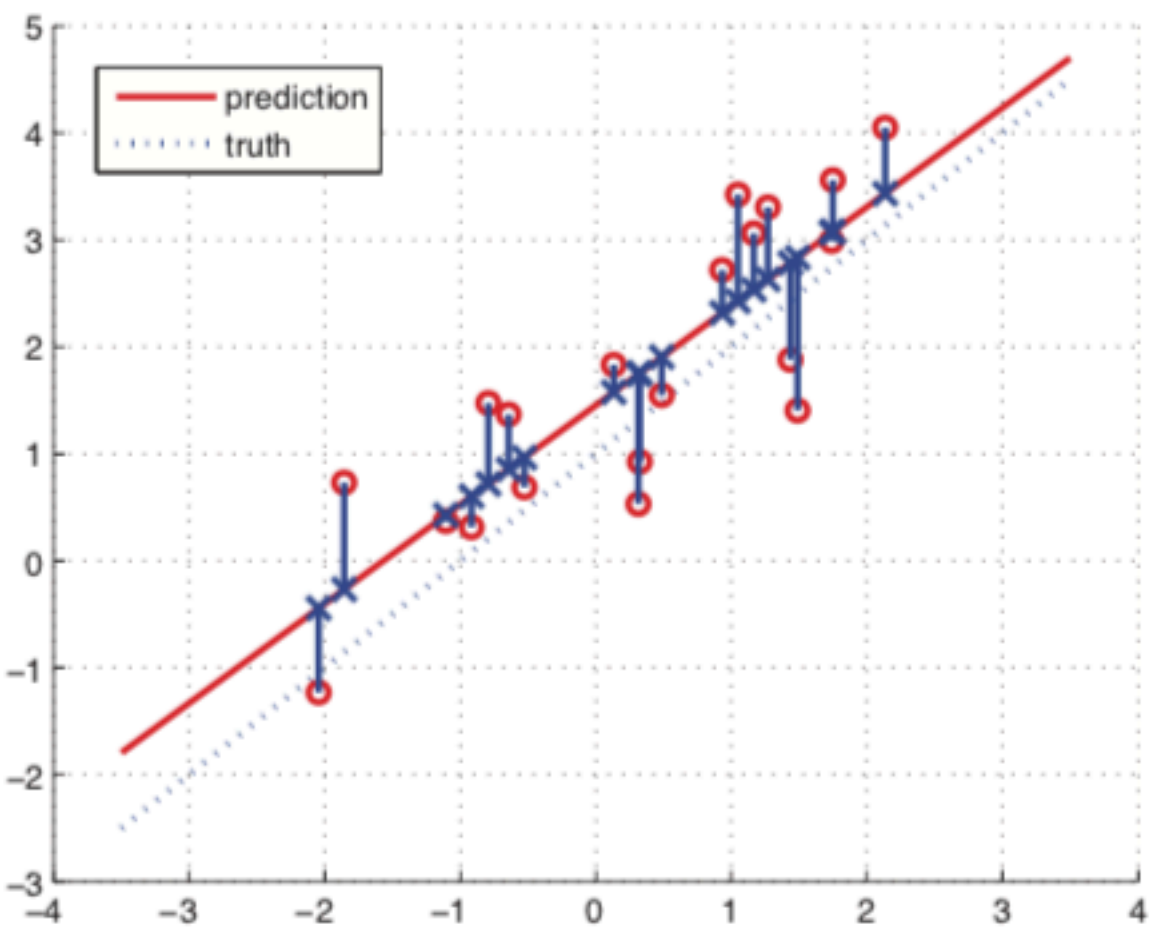
- $\hat{y} = Xw = X(X^T X)^{-1} X^T y$

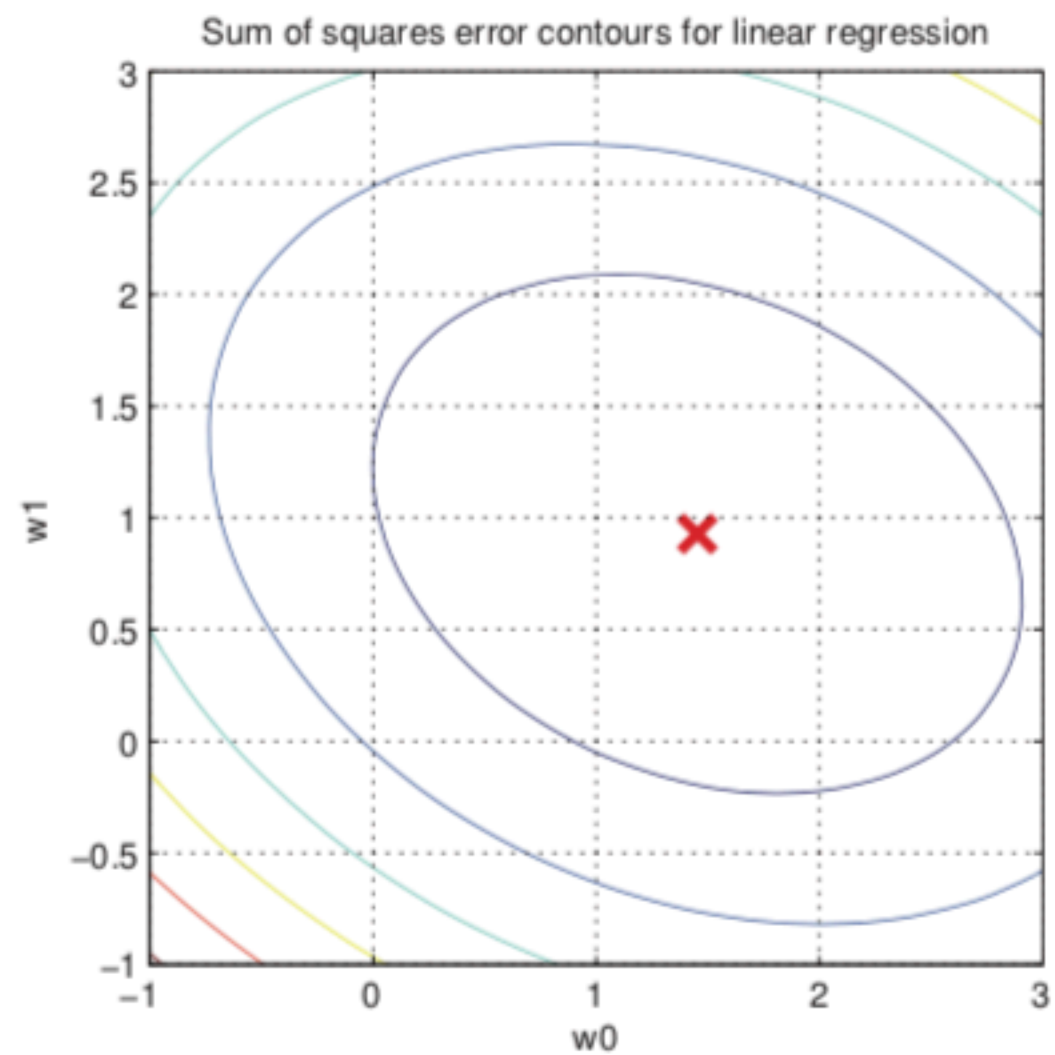
Matriz de proyección $\hat{P}$: proyecta y en el espacio de columnas de X



Regresión

- **Regresión Lineal**





https://seeing-theory.brown.edu/regression-analysis/index.html#section1

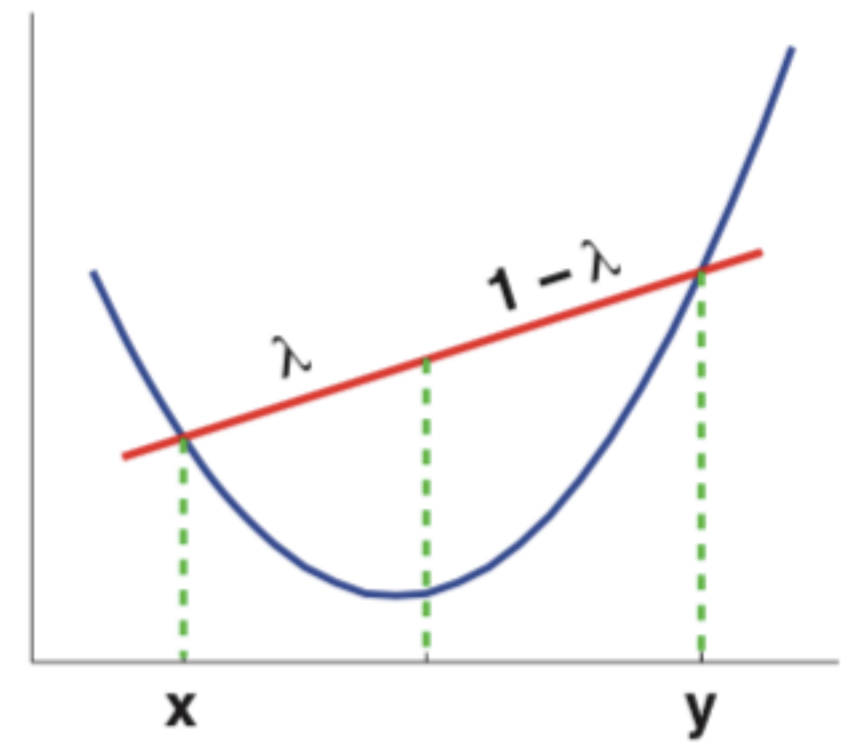
Regresión

- **Optimización convexa.**
  Una importante característica de la regresión lineal es que la superficie del NLL es convexa

  Una función es convexa si para

$$0 \leq \lambda \leq 1$$

$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda) f(\theta')$$



Sí la función es convexa tiene un **mínimo global** θ^*

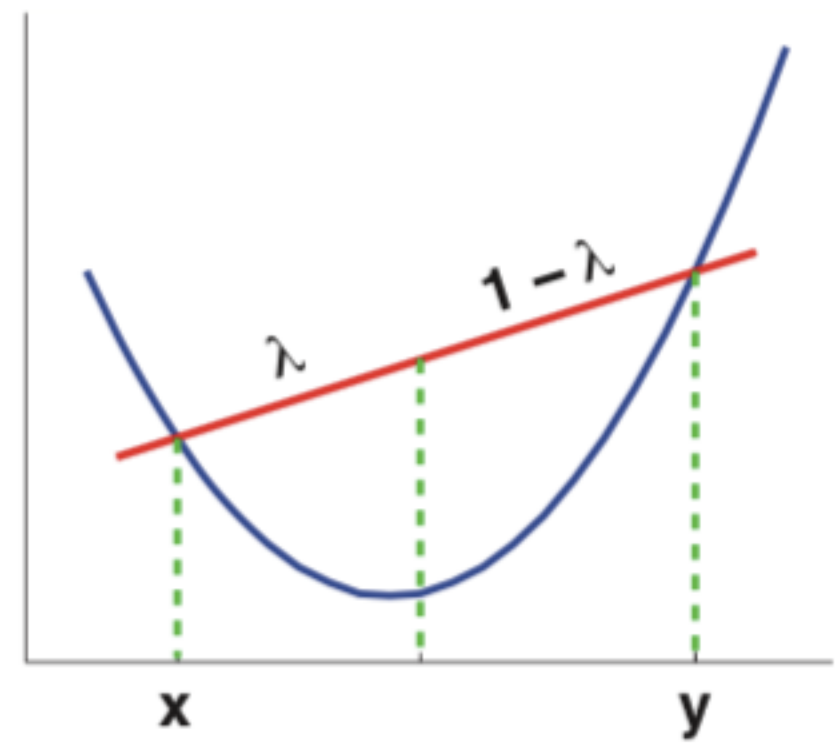
Regresión

- **Optimización convexa.**
  Una importante característica de la regresión lineal es que la superficie del NLL es convexa

  Una función es convexa si para

$$0 \leq \lambda \leq 1$$

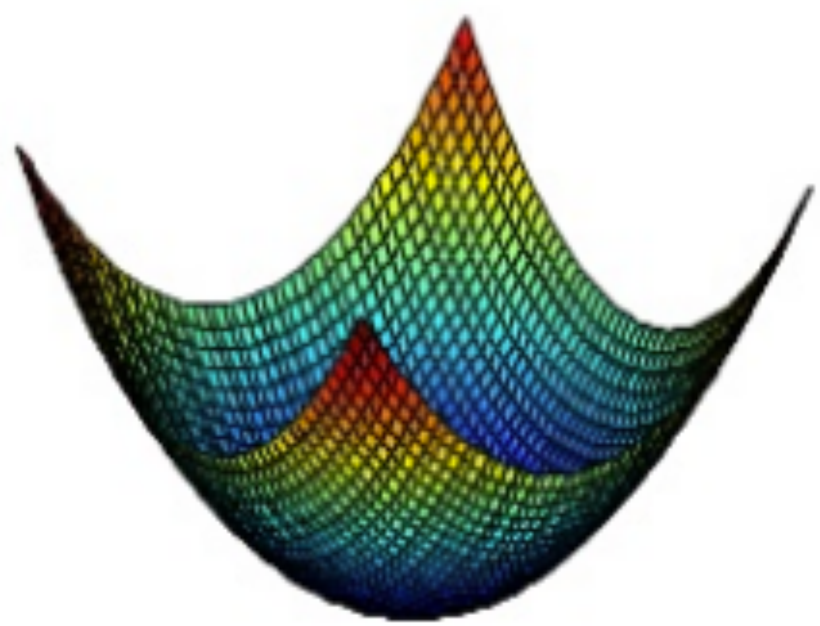
$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda) f(\theta')$$

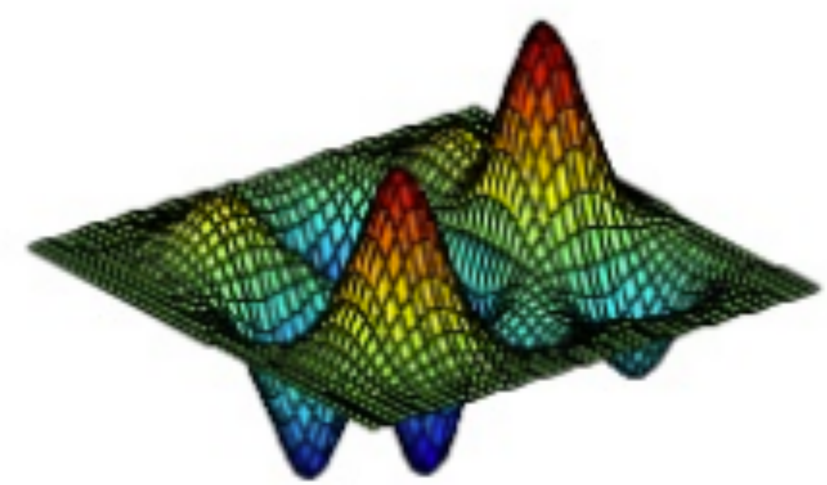


Sí la función es convexa tiene un **mínimo global** θ^*

Regression

- **Optimización convexa**





Convexa vs No Convexa

Regression

- **Regresión Lineal**
  También puede modelar relaciones no lineales utilizando **funciones basales**.

$$p(y \mid x, w) = N(w^T \phi(x), \sigma^2(x))$$

- Por ejemplo, es muy común utilizar polinomios

$$\phi(x) = [1, x, x^2, \ldots, x^d]$$

- **La optimización es aún convexa**.

Regression

- También puede modelar relaciones no lineales utilizando **funciones basales**.

$$p(y\,|\,x, w) = N(w^T\phi(x), \sigma^2(x))$$

- Por ejemplo, es muy común utilizar polinomios

$$\phi(x) = [1, x, x^2, \ldots, x^d]$$

- **La optimización es aún convexa**.

Regression Polinomial

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M + \epsilon.$$

- Para entrenar este modelo, se crea un nuevo dataset, con variables polinomiales, luego la optimización es igual que el caso normal

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}$$

Regression Polinomial

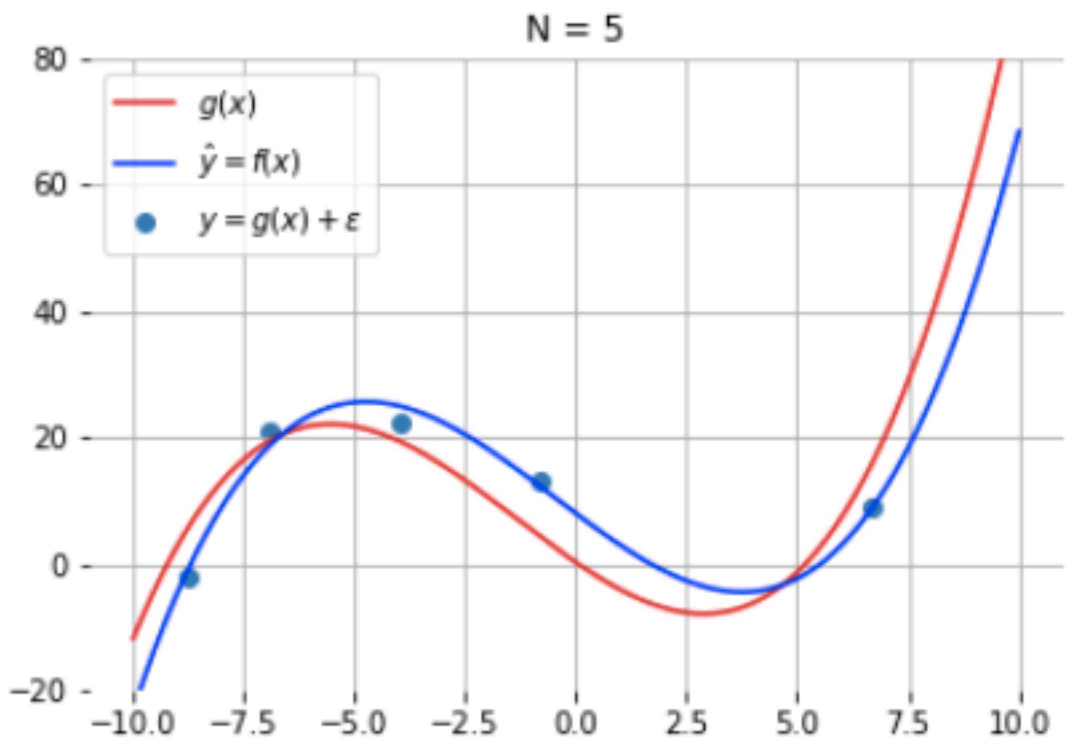
$$
\begin{aligned}
y = &\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_1^M \\
&+ \beta_{M+1} x_2 + \ldots + \beta_{2M} x_2^M \\
&+ \ldots \\
&+ \beta_{M(J-1)+1} x_J + \ldots + \beta_{MJ} x_J^M
\end{aligned}
$$

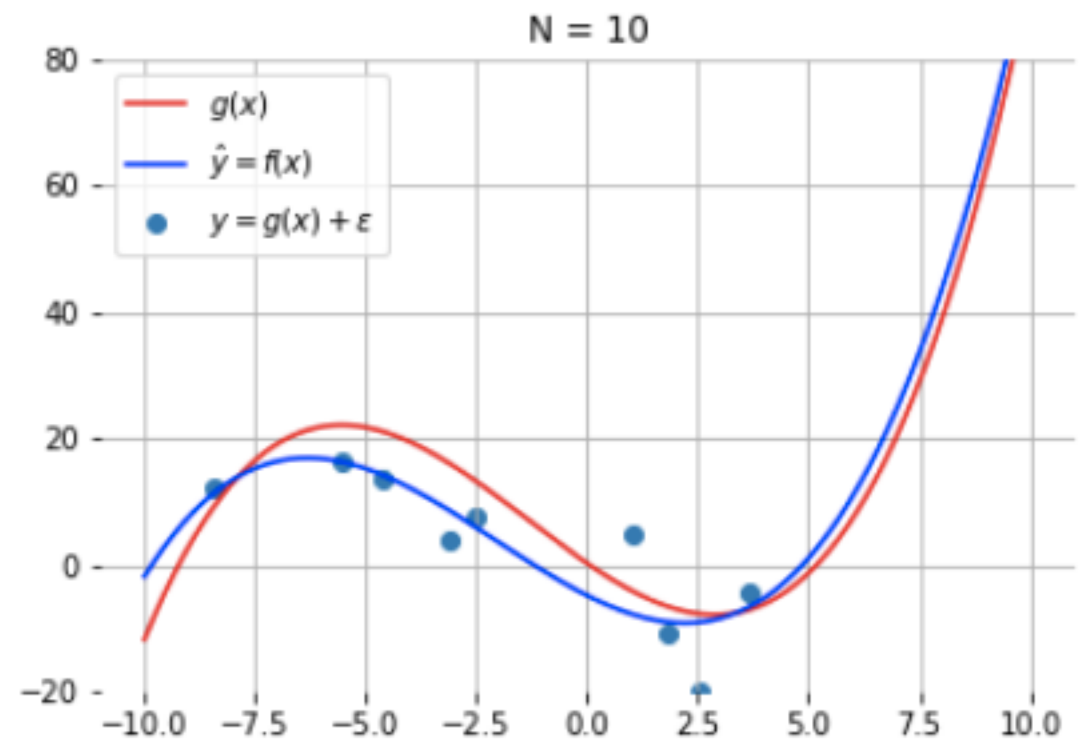
- También se pueden agregar términos de interacción

$$
\begin{aligned}
y = &\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_1^M \\
&+ \beta_{1+M} x_2 + \ldots + \beta_{2M} x_2^M \\
&+ \beta_{1+2M} (x_1 x_2) + \ldots + \beta_{3M} (x_1 x_2)^M
\end{aligned}
$$

Regression

- **Regresión Lineal** (efecto de más datos)

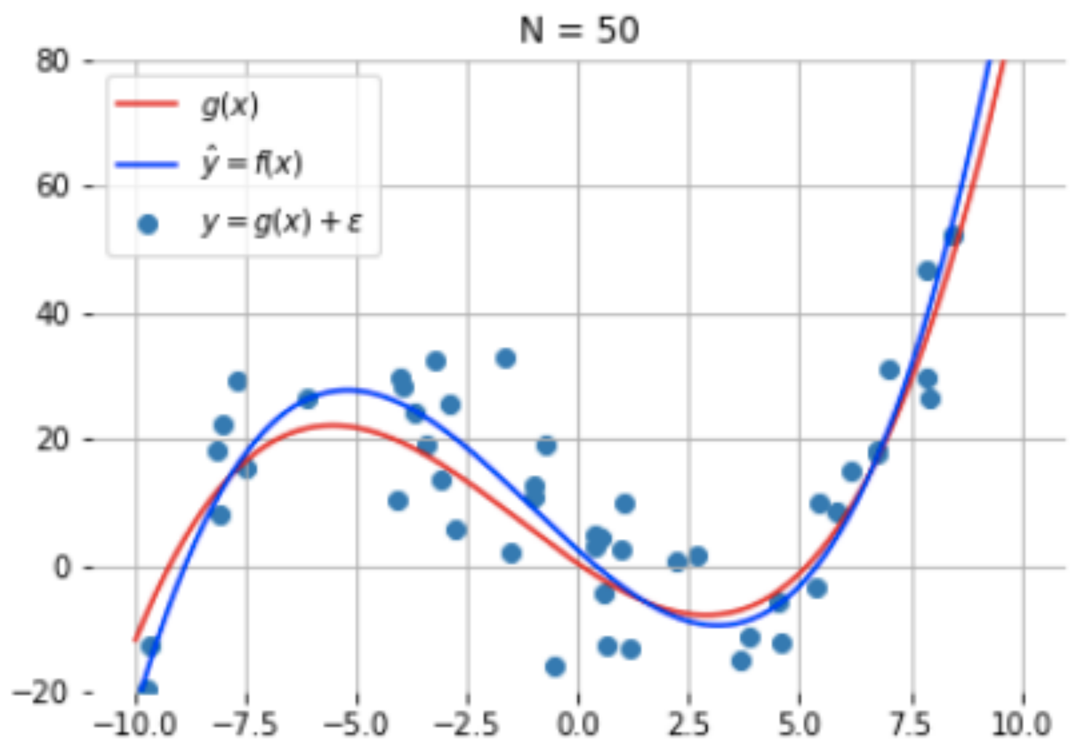


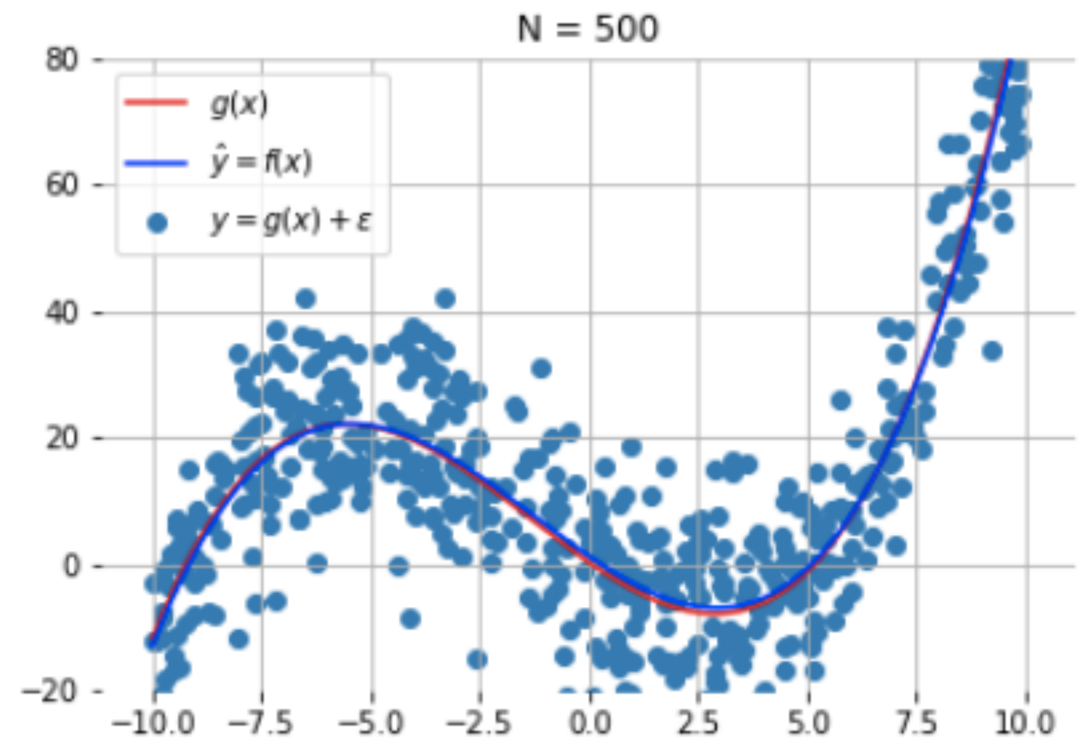


Credit: Deep Learning Class. Louppe G.

Regression

- **Regresión Lineal** (efecto de más datos)

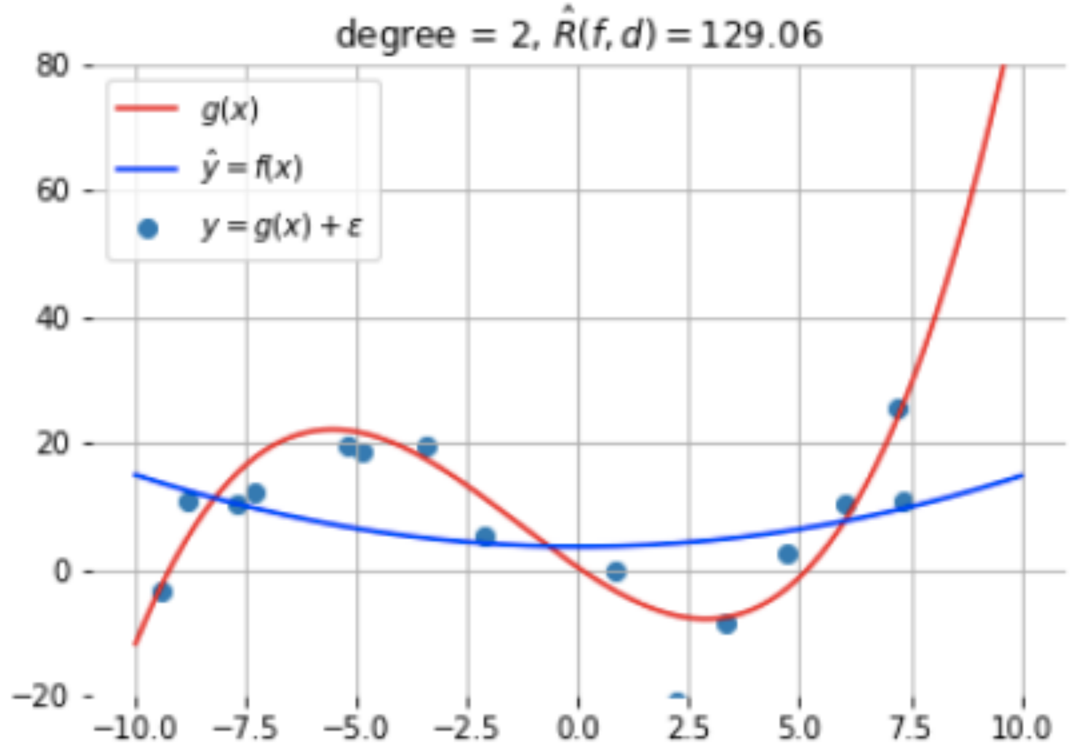


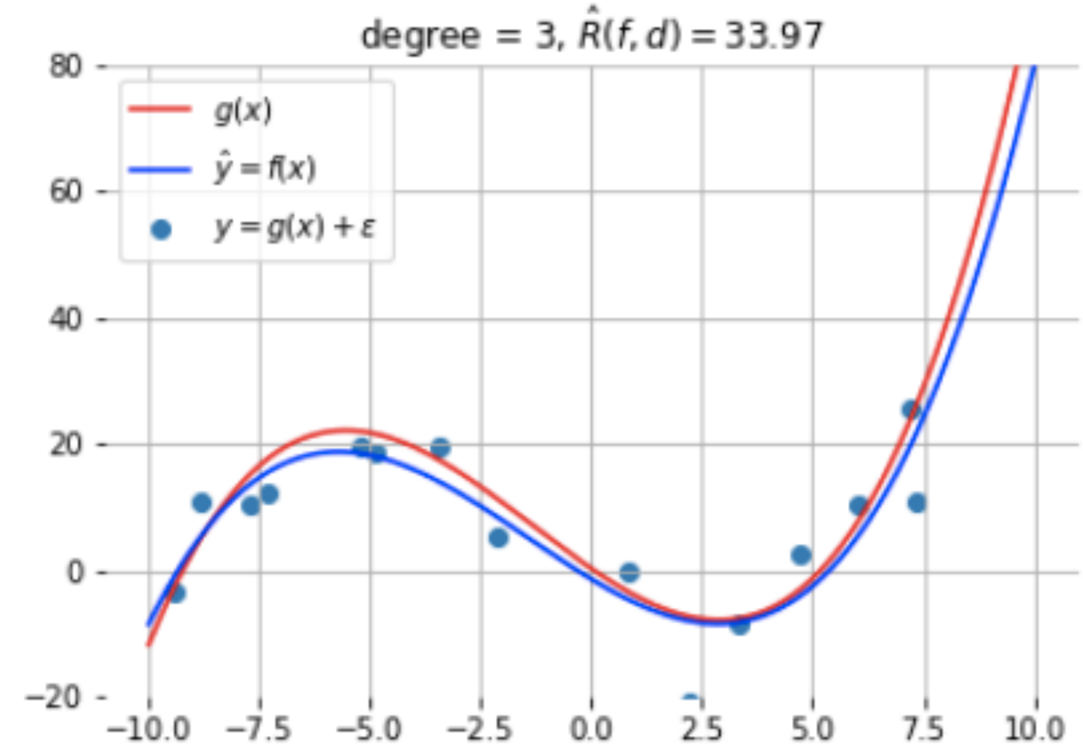


Credit: Deep Learning Class. Louppe G.

Regression

- **Regresión Lineal** (efecto de complejidad del modelo)

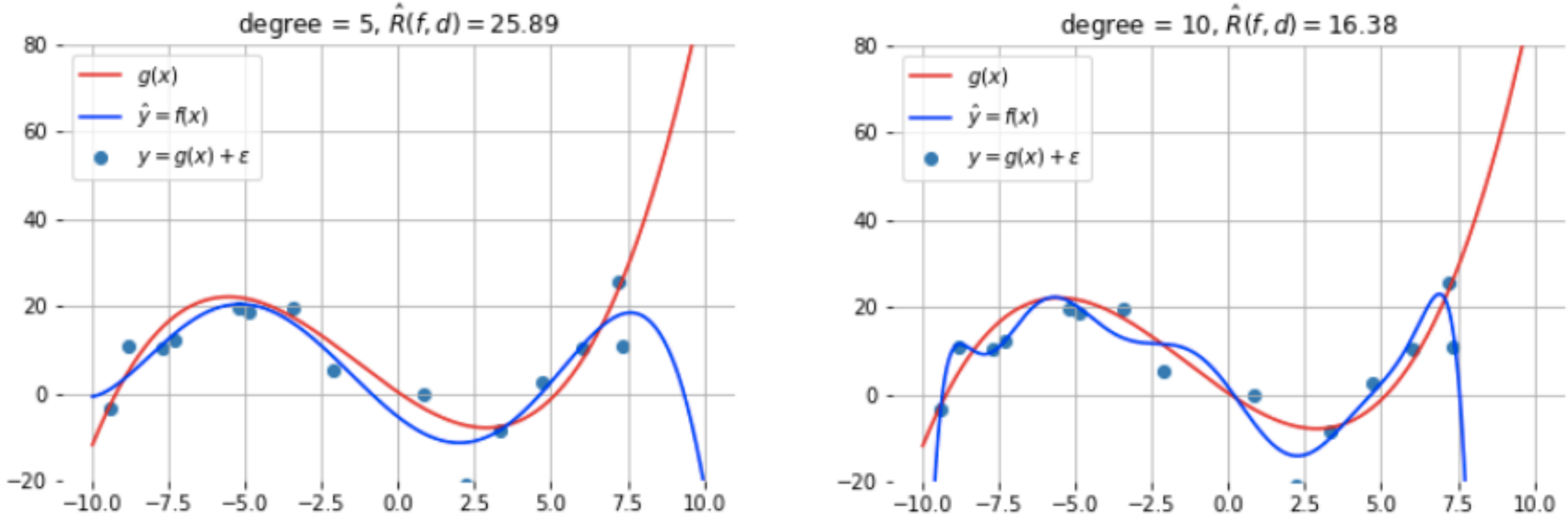




Credit: Deep Learning Class. Louppe G.

Regression

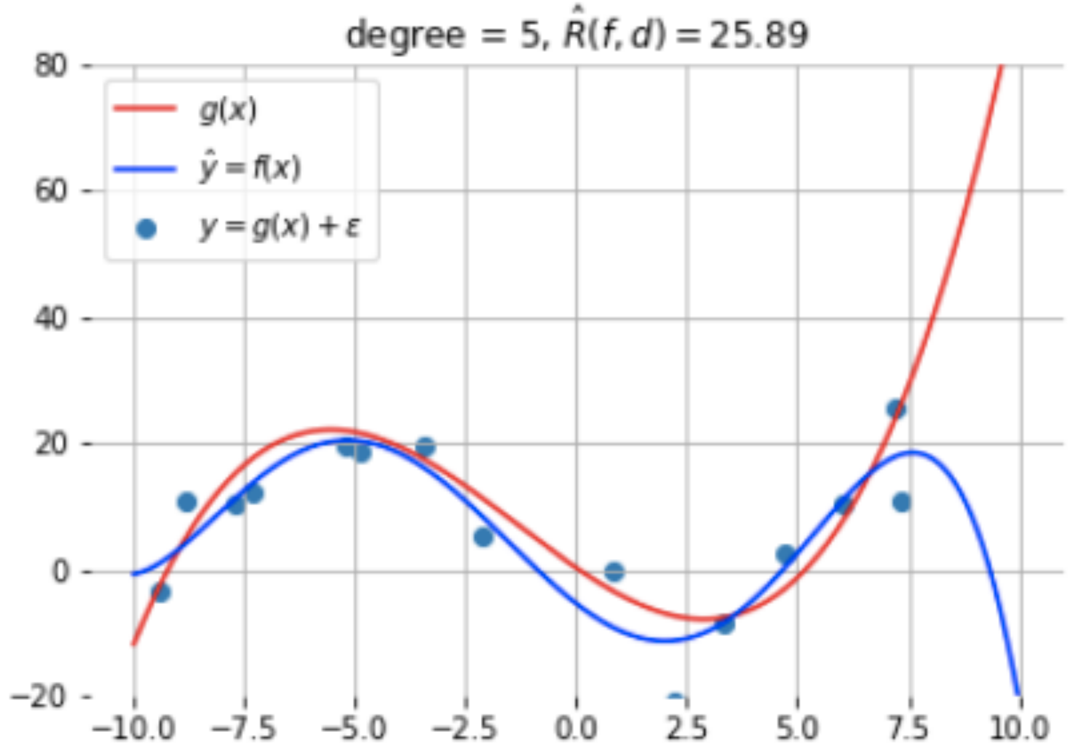
- **Regresión Lineal** (efecto de complejidad del modelo)

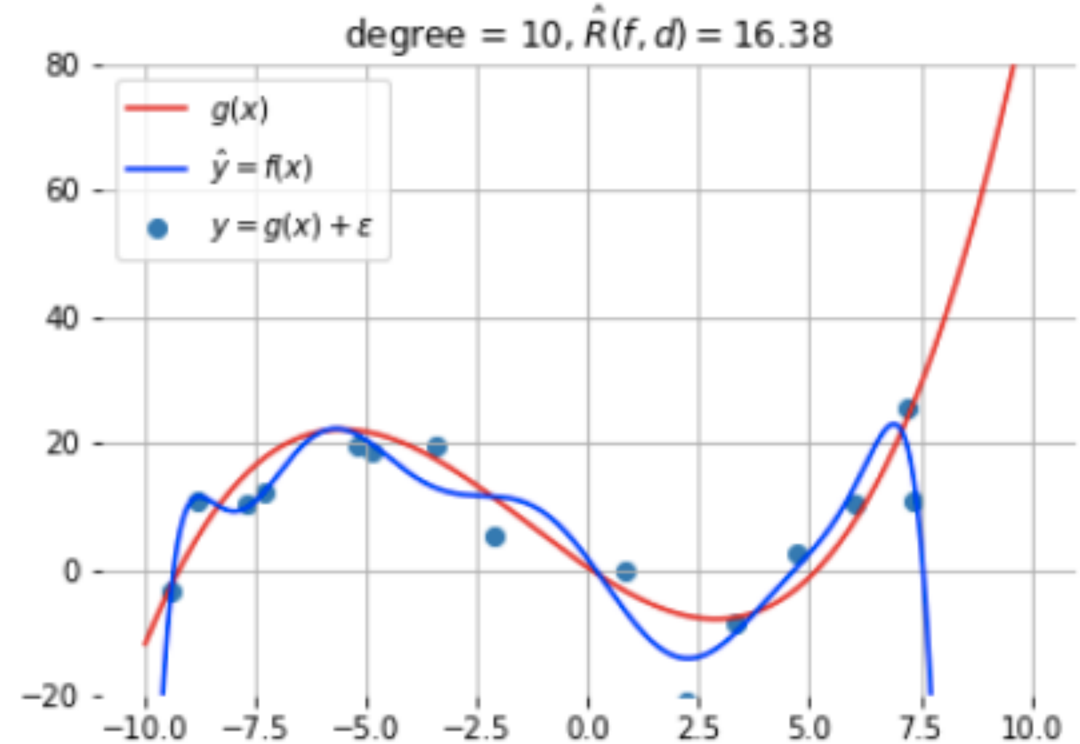


Credit: Deep Learning Class. Louppe G.

Regression

- **Regresión Lineal** (efecto de complejidad del modelo)





Este efecto se llama **overfitting**.
Lo estudiaremos más en las próximas clases

Credit: Deep Learning Class. Louppe G.

Regression

- **Regresión Ridge**
  La regresión lineal **sobreajusta** porque puede tomar cualquier valor para modelar los datos

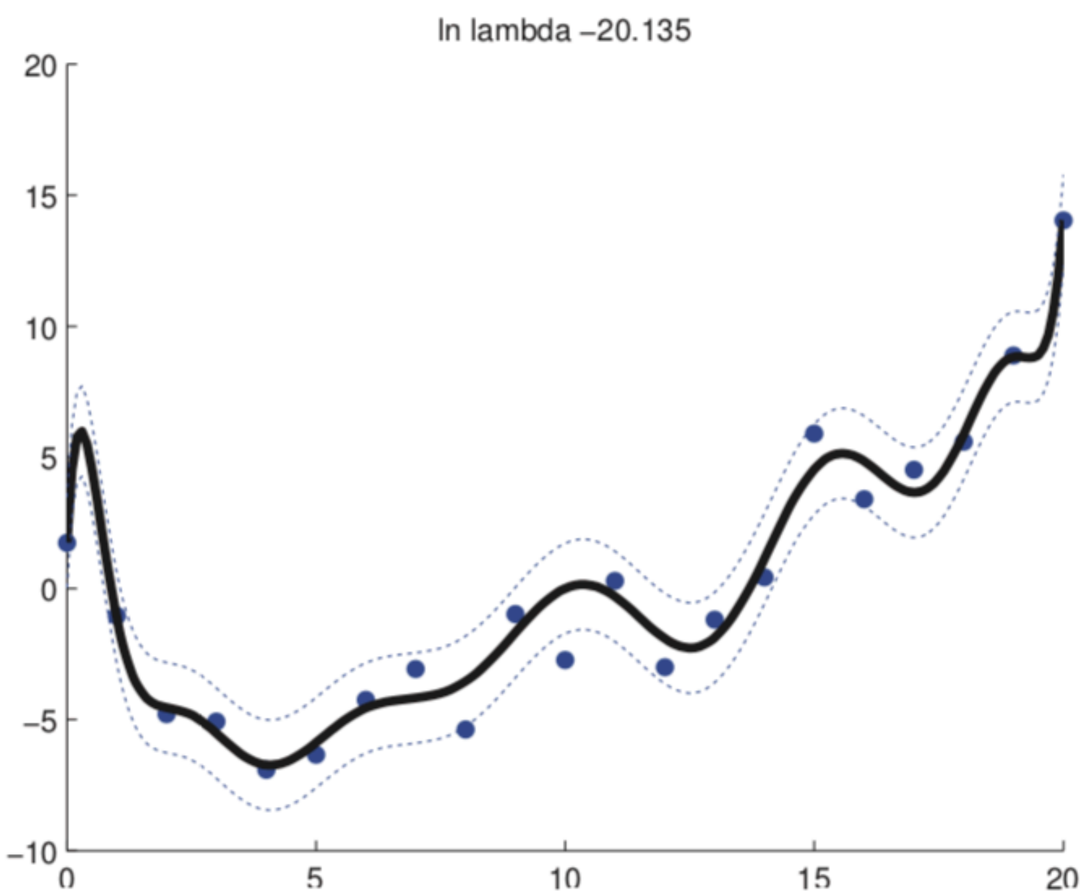
  Si la data tiene mucho **ruido** y la **capacidad** del modelo es suficiente, los parámetros modelaran funciones muy complejas.

  Esto produce **inestabilidad**: Un pequeño cambio en los datos puede cambiar totalmente los parámetros.

Regression

- **Regresión Ridge**

6.560, -36.934, -109.255, 543.452, 1022.561, -3046.224, -3768.013, 8524.540, 6607.897, -12640.058, -5530.188, 9479.730, 1774.639, -2821.526



Regression

- **Regresión Ridge**
  Podemos motivar parámetros pequeños usando un a priori en los pesos:
  $$p(w) = \prod_{j=1}^{D} N(w_j \,|\, 0, \tau^2)$$
  τ^2 pequeños forzarán parámetros alrededor de 0.

- El a posteriori es

- $$p(w \,|\, X) = \prod_{i=1}^{N} N(y_i \,|\, w_0 + w^T x_i) \prod_{j=1}^{D} \log N(w_j \,|\, 0, \tau^2)$$

Regression

- **Regresión Ridge**
  Podemos motivar parámetros pequeños usando un a priori en los pesos:

$$p(w) = \prod_{j=1}^{D} N(w_j \mid 0, \tau^2)$$

  τ^2 pequeños forzarán parámetros alrededor de 0.

- La estimación MAP es:

$$argmax_w \sum_{i=1}^{N} \log N(y_i \mid w_0 + w^T x_i, \sigma^2) + \sum_{i=1}^{D} \log N(w_j \mid 0, \tau^2)$$

Regression

- **Regresión Ridge**

$$argmax_w \sum_{i=1}^{N} \log N(y_i \,|\, w_0 + w^T x_i, \sigma^2) + \sum_{i=1}^{D} \log N(w_j \,|\, 0, \tau^2)$$

- Que es equivalente a minimizar

$$J(w) = \frac{1}{2\sigma^2} RSS(w) + \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\tau} \sum_{j=1}^{D} w_j^2$$

$$J(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T x_i))^2 + \lambda \,||\, w \,||_2^2$$

- Con $\lambda = \sigma^2/\tau^2$ y $||\, w \,||_2^2 = \sum_j w_j^2 = w^T w$

Regression

- **Ridge Regression**

$$J(w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w^T x_i))^w + \lambda || w ||_2^2$$

Añadir el término ||w|| es una técnica común conocida como regularización ℓ_2 o **weight decay** (puede ser derivada sin una interpretación probabilística)

- El mínimo obtenido es:

$$\hat{w}_{ridge} = (\lambda I_D + X^T X)^{-1} X^T y$$

Regression

- **Regresión Ridge**

$$\hat{w}_{ridge} = (\lambda I_D + X^T X)^{-1} X^T y$$

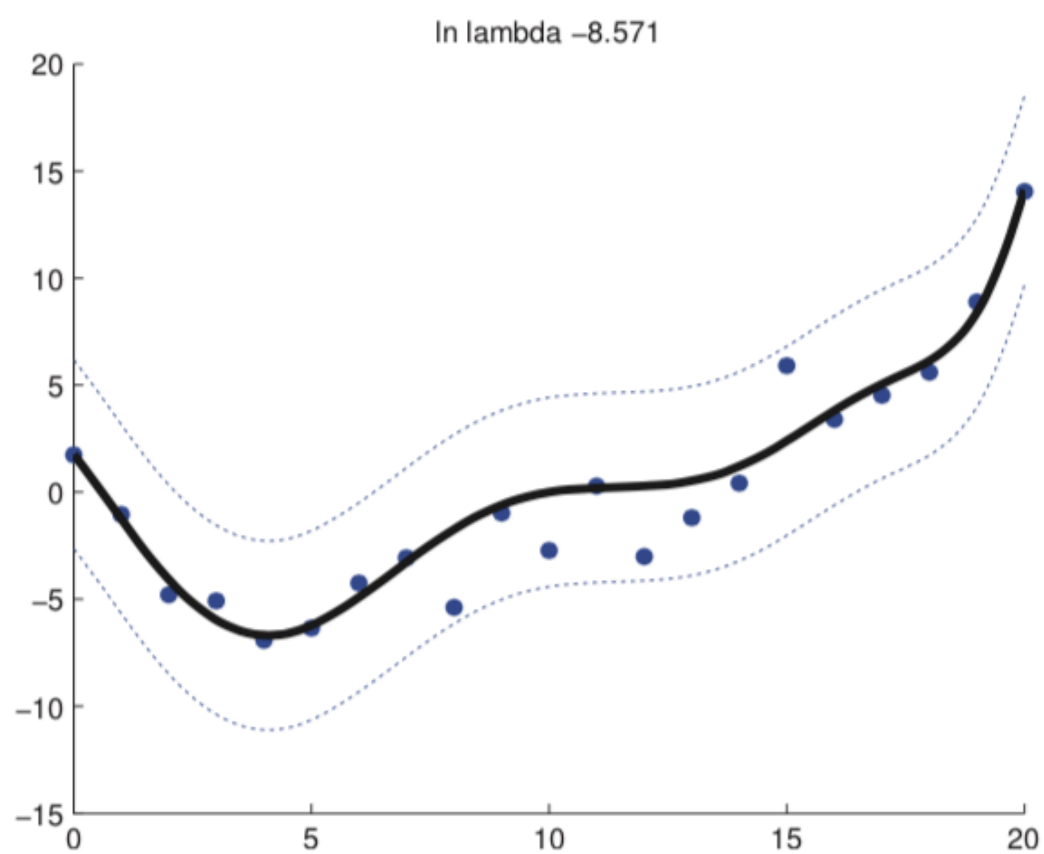
- Además de tener mejores propiedades estadísticas, la solución es más estable.

- $(\lambda I_D + X^T X)^{-1}$ está mejor condicionada que $X^T X^{-1}$

Regression

- **Regresión Ridge**

```
2.128, 0.807, 16.457, 3.704, -24.948, -10.472, -2.625, 4.360, 13.711,
10.063, 8.716, 3.966, -9.349, -9.232
```



$$\lambda = 10^{-3}$$