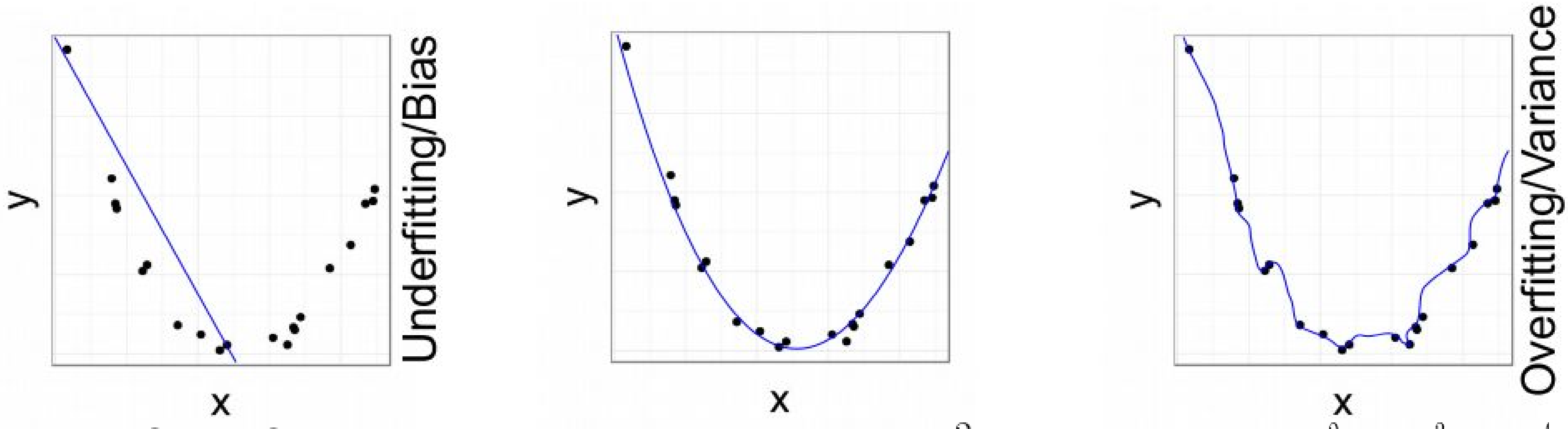
Regularización

Overfitting y Underfitting: Ejemplo en Regresión



$\Theta_0 + \Theta_1 x$ $\qquad$ $\Theta_0 + \Theta_1 x + \Theta_2 x^2$ $\qquad$ $\Theta_0 + \Theta_1 x + \Theta_2 x^2 + \Theta_3 x^3 + \Theta_4 x^4$

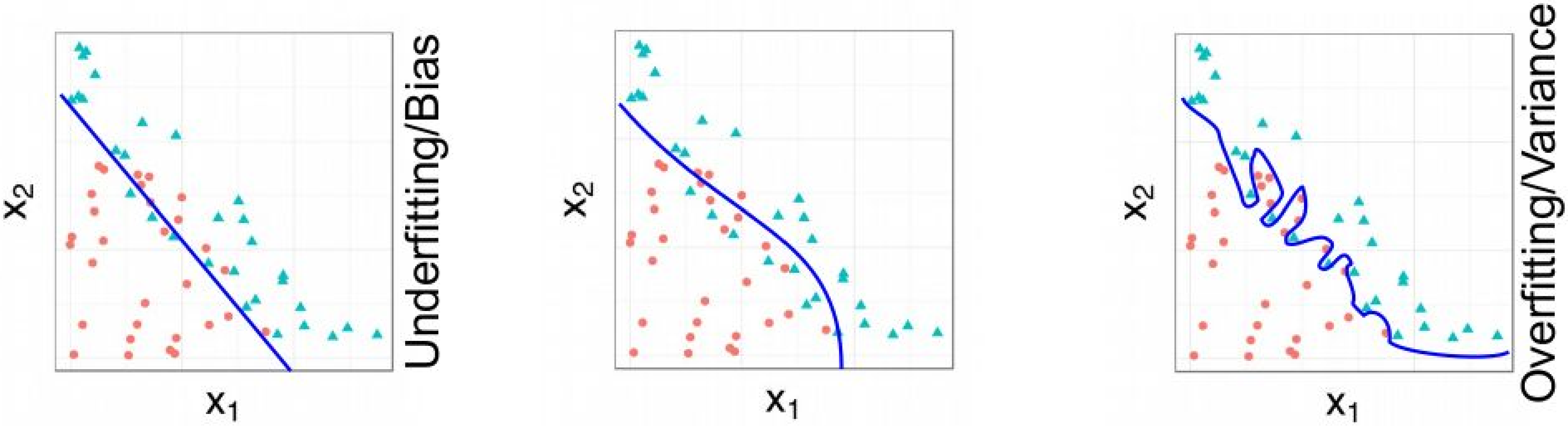
Underfitting/Bias

- Error en conjunto de entrenamiento es alto.
- Hipótesis ***Simple*** falla en generalizar nuevos ejemplos.

Overfitting/Variance

- Error en conjunto de entrenamiento es bajo.
- Hipótesis ***Compleja*** falla en generalizar nuevos ejemplos.

Overfitting y Underfitting: Ejemplo en Clasificación



$$h_\Theta(x) = g(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2)$$

$$g(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \Theta_3 x_1^2 + \Theta_4 x_2^2)$$

$$g(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_1^2 + \Theta_3 x_1^2 x_2 + \Theta_4 x_1^2 x_2^2 + \ldots)$$

Underfitting/Bias

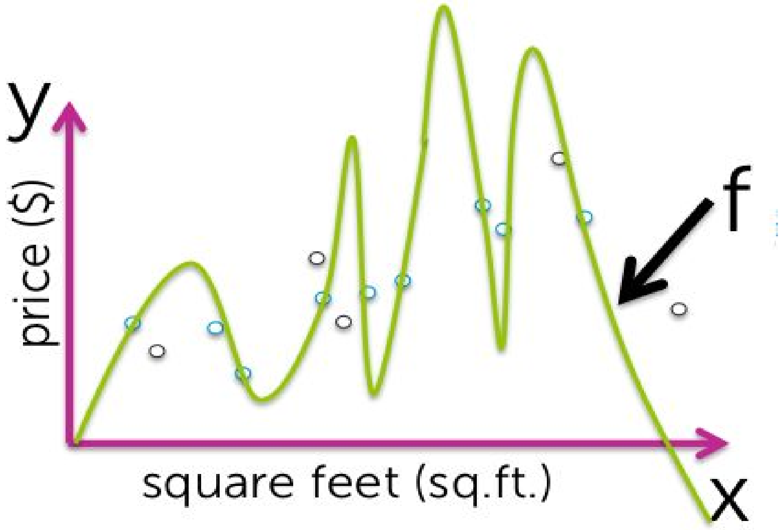
- Error en conjunto de entrenamiento es alto.
- Hipótesis ***Simple*** falla en generalizar nuevos ejemplos.

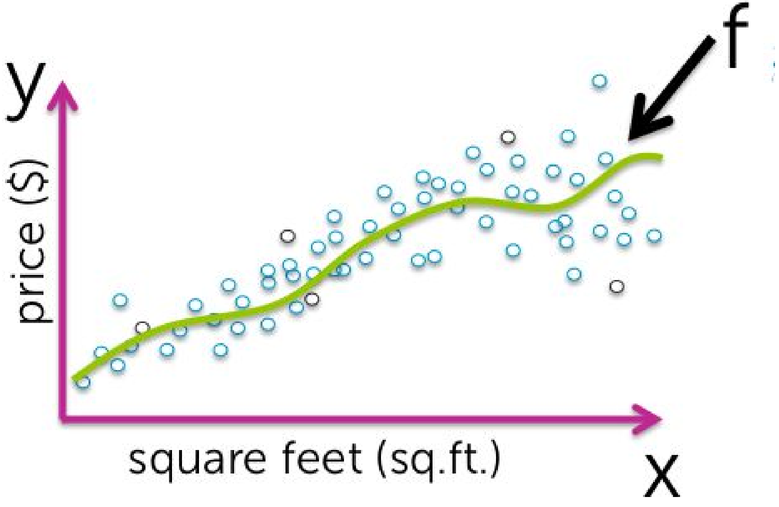
Overfitting/Variance

- Error en conjunto de entrenamiento es bajo.
- Hipótesis ***Compleja*** falla en generalizar nuevos ejemplos.

Como el número de observaciones influye en Overfitting?

- Pocas observaciones (n pequeño)
  - Rápidamente tenemos overfitting debido a la complejidad del modelo.
- Muchas observaciones (n muy grande)
  - Dificil de tener overfitting.





Tratando con Underfitting y Overfitting

Previniendo Underfitting

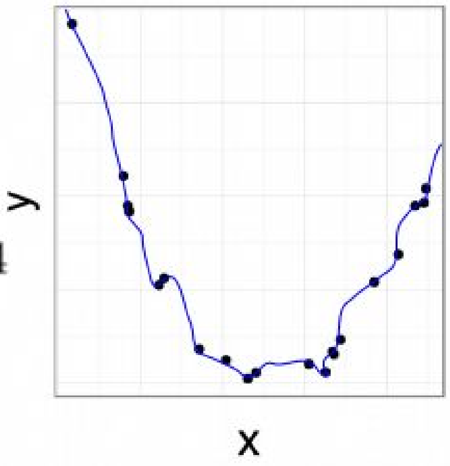
- Adicionar más características polinomiales - Incrementar la complejidad del modelo.

Previniendo Overfitting

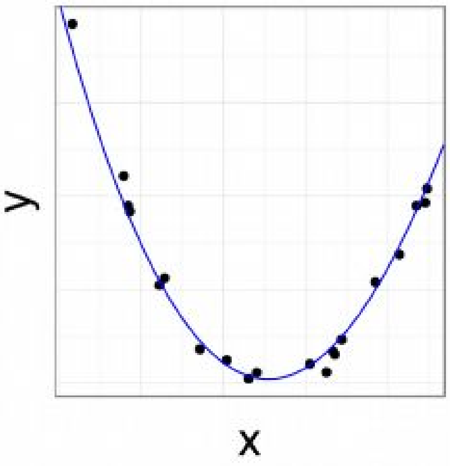
- Aumentar más data de entrenamiento
- Reducir el número de características.
  - Manualmente.
  - Uso de algoritmo de selección.
- Regularización
  - Mantener todas las características pero reducir el valor/importancia de los parametros θj.
  - Trabaja bien para varias características que contribuyan un poco en la predicción de y.

Regularización (penalización de parámetros)

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



$$\theta_1, \theta_3, \theta_4 \text{ small:}$$
$$\approx \theta_0 + \theta_2 x^2$$



$$\min_{\Theta} \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2$$

$$\min_{\Theta} \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2 + 1000\Theta_1^2 + 1000\Theta_3^2 + 1000\Theta_4^2$$

Regularización

Pequeños valores para parámetros $\Theta_0, \Theta_1, \ldots, \Theta_n$

- Hipótesis “simple” (función más suave)
- Menos propenso a overfitting

El problema está cuando:

- Tenemos muchas variables $x_0, x_1, x_2, \ldots, x_{150}$
- Por lo tanto tendremos muchos parámetros: $\theta_0, \theta_1, \theta_2, \ldots, \theta_{150}$
- Qué parámetros penalizar?

Regularización L2

- La solución está en penalizar todos los parámetros al mismo tiempo

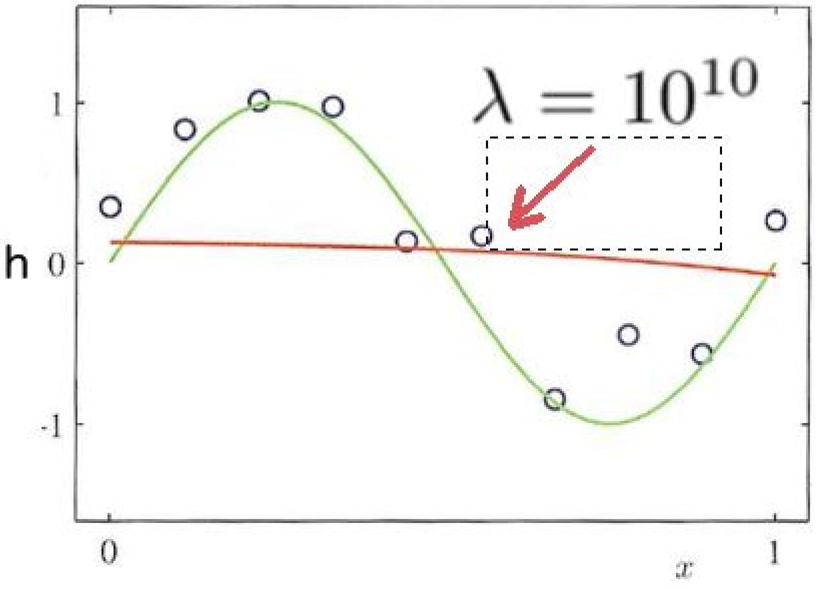
$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

- Parámetro de regularización λ controla el nivel de simplicidad que tendrá la función hipótesis.
- Por convención no se penaliza θ_0

Efecto del parámetro λ

- Un λ muy pequeño no ayuda en nada.
- Un λ demasiado grande causaria underfitting.

$$\theta_1 \approx 0$$
$$\theta_0 \approx 0$$
$$\dots$$
$$\theta_n \approx 0$$
$$\rightarrow h(\Theta) \approx \theta_0$$



Gradiente Descendente

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left(\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right)$$

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{1}{m} \left(\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} \right)$$

Ecuación normal

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

Supongamos que m ≤ n , entonces

$$\theta = \boxed{(X^T X)^{-1}} X^T y$$

- No es invertible.
- Usar función ***pinv*** de R, python, Matlab/octave.

Si λ > 0:

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \right)^{-1} X^T y$$