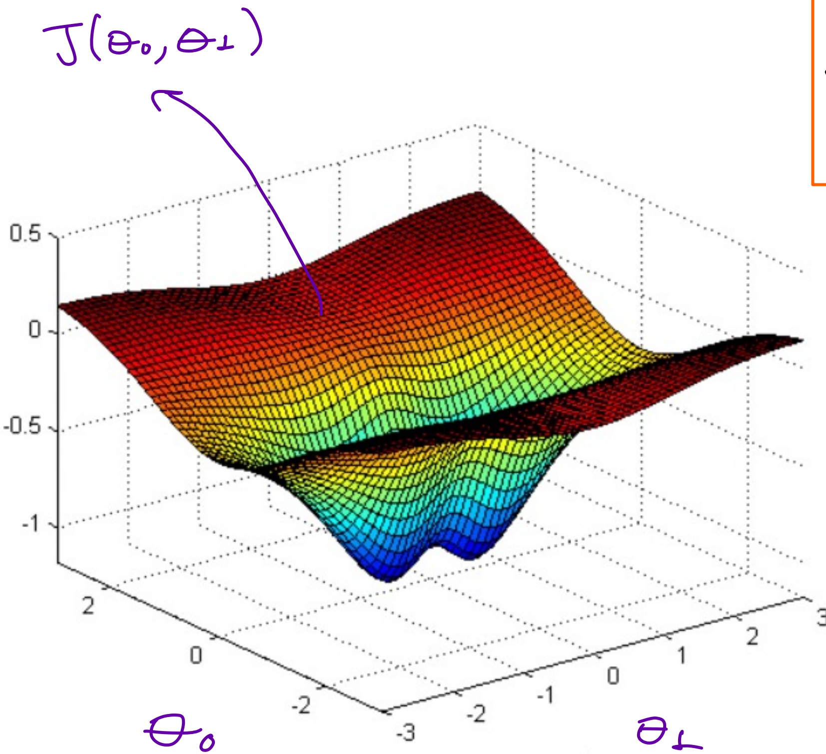
Hoy

- ☐ Ecuaciones descenso de gradiente
- ☐ sobreajuste
- ☐ Técnicas de regularización :
  - ☐ Regresión Ridge (L_2)
  - ☐ Regresión Lasso (L_1)
- ☐ Diferencia entre L_1 y L_2
- ☐ Una intuición geométrica

Complementando lo hablado anteriormente sobre descenso de gradiente



Gradiente de la función J

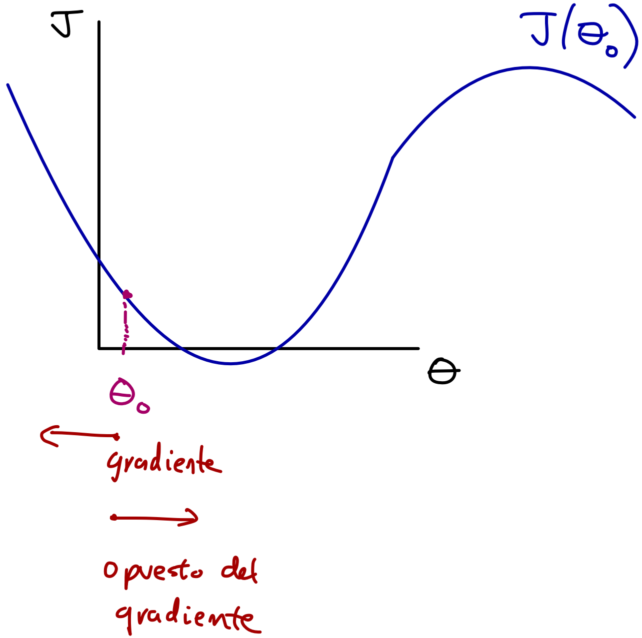
$$\nabla J(\theta_0,\theta_1) = \left(\frac{\partial J}{\partial \theta_0}, \frac{\partial J}{\partial \theta_1}\right)$$

Partimos de cualquier punto (θ_0,θ_1) en la superficie.

Queremos un nuevo punto (θ_0',θ_1') que resulte de dar un "paso" en dirección opuesta al gradiente

$$(\theta_0',\theta_1') := (\theta_0,\theta_1) - \alpha\nabla J(\theta_0,\theta_1)$$

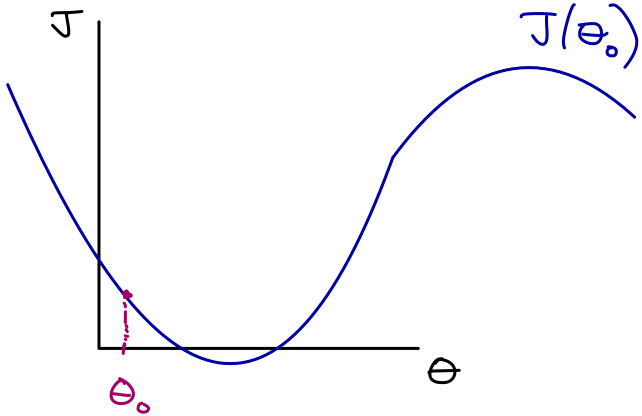
Por ejemplo, si J sólo tiene un parámetro



$$\theta_0' = \theta_0 - \alpha \frac{d}{d\theta_0} J(\theta_0)$$

Determina el tamaño del paso

Por ejemplo, si J sólo tiene un parámetro



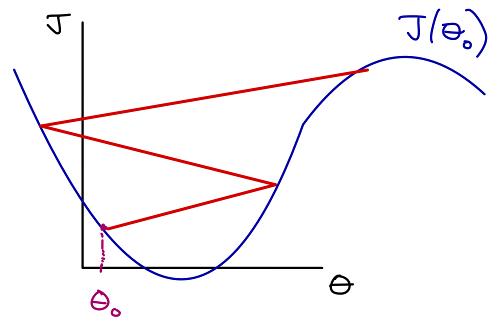
gradiente

opuesto del gradiente

$$\theta_0' = \theta_0 - \alpha \frac{d}{d\theta_0} J(\theta_0)$$

Determina el tamaño del paso

si el paso es muy grande..



si el paso es muy pequeño? Puede tardar mucho el proceso !

Descenso de gradiente para regresión lineal (Ejemplo 2 parámetros θ_0, θ_1)

$$J(?) = \frac{1}{2N} \sum_{i=1}^{N} \left(h(x_i) - y_i \right)^2$$

Antes no estaba, pero se puede agregar
se puede tomar promedio de las diferencias

se pone estratégicamente para simplificar
la fórmula al derivar

Descenso de gradiente para regresión lineal (Ejemplo 2 parámetros Θ_0, Θ_1)

$$J(?) = \frac{1}{2N} \sum_{i=1}^{N} \left(h(x_i) - y_i \right)^2$$

Antes no estaba, pero se puede agregar

se puede tomar promedio de las diferencias

se pone estratégicamente para simplificar la fórmula al derivar

$$h(x) = \sum_{j=0}^{N} x_j \theta_j$$

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

se considera $\Theta_0 = \beta_0$

y $x_0 = 1$ para el termino independiente

Descenso de gradiente para regresión lineal (Ejemplo 2 parámetros θ_0, θ_1)

$$J(?) = \frac{1}{2N} \sum_{i=1}^{N} \left(h(x_i) - y_i \right)^2$$

Antes no estaba, pero se puede agregar

se puede tomar promedio de las diferencias

se pone estratégicamente para simplificar la fórmula al derivar

$$h(x) = \sum_{j=0}^{N} x_j \theta_j$$

$$f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

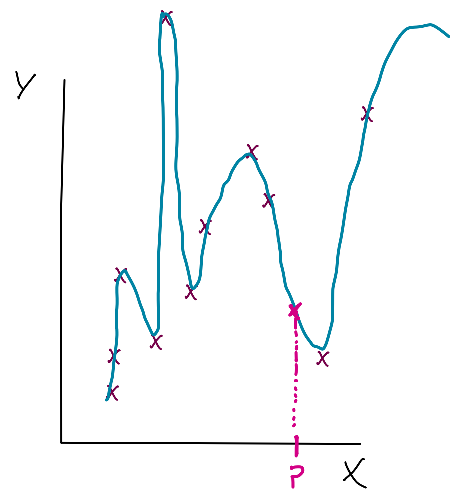
se considera $\theta_0 = \beta_0$ y $x_0 = 1$ para el termino independiente

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} \frac{1}{2N} \sum_{i=1}^{N} \left(\theta_0 + \theta_1 x_i - y_i \right)^2$$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \sum_{i=1}^{N} \frac{1}{N} \left(\theta_0 + \theta_1 x_i - y_i \right)$$

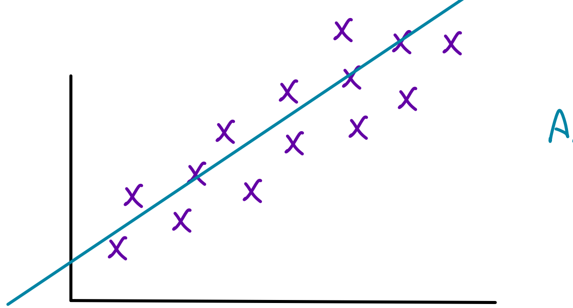
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \sum_{i=1}^{N} \frac{1}{N} \left(\theta_0 + \theta_1 x_i - y_i \right) x_i$$

Sobreajuste : Se presenta cuando el modelo aprende "en exceso". Esto es, aprende detalle y ruido del conjunto de entrenamiento, al punto de impactar negativamente el desempeño en nuevos datos



Regularización Son técnicas para evitar el sobreajuste. Hablaremos de dos comunes: Regresión Ridge y Regresión Lasso.

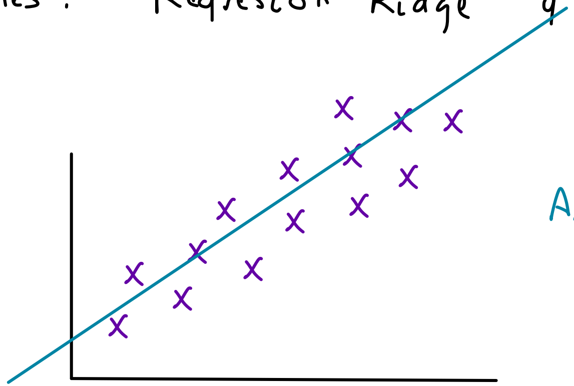
Intuición:



Antes RSS

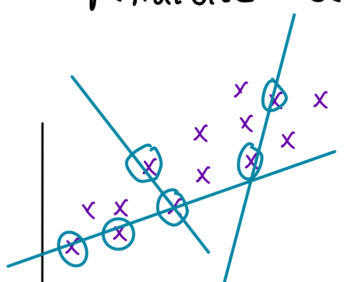
Regularización Son técnicas para evitar el sobreajuste. Hablaremos de dos comunes: Regresión Ridge y Regresión Lasso.

Intuición:



Antes RSS

Si tengo muestras pequeñas de los datos (para ejemplificar 2 datos), en ocasiones la pendiente se eleva de una forma no natural



En la práctica, (realidad), los datos con comportamientos lineales no suelen tener pendientes muy altas

La Regresión de Ridge evita esta situación penalizando pendientes altas.

método de "encogimiento" (shrinkage method)

¿Cómo penaliza las pendientes altas?

En la práctica, (realidad), los datos con comportamientos lineales no suelen tener pendientes muy altas

La Regresión de Ridge evita esta situación penalizando pendientes altas.

método de "encogimiento" (shrinkage method)

¿Cómo penaliza las pendientes altas?

Con una nueva función de costo:

λ: Parámetro a definir en cada entrenamiento

$$\underbrace{\hat{\beta}^{ridge}}_{\text{"}L_2\text{"}} = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij} \beta_j \right)^2}_{RSS} + \underbrace{\lambda \overbrace{\sum_{j=1}^{p} \beta_j^2}^{= \|\beta\|^2}}_{Nuevo} \right\}$$

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Observación 1 Si $\lambda = 0$ no es más que la fórmula para regresión lineal original

Observación 2 Los resultados de la regresión de Ridge pueden cambiar significativamente dependiendo de la escala de los datos, por lo que se aconseja "estandarizar" los datos

Beneficios

- Mejores métricas cuando la muestra es pequeña
- Evita el sobreajuste y disminuye la varianza → veremos después
- Útil cuando hay varias columnas con alta correlación

Caso una variable

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i \underbrace{- \beta_0 - x_i \beta_1}_{\text{recta}} \right)^2 + \lambda \underbrace{\beta_1}_{\text{pendiente}}^2 \right\}$$

Regresión de Lasso (LASSO) ("L_2")

Least Absolute Shrinkage and Selection Operator

Es otro método de encogimiento similar a Ridge

$$\hat{\beta}^{Lasso} = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \underbrace{|\beta_j|} \right\}$$

"L_2"

valor absoluto de las pendientes, en lugar de cuadrados

¿Diferencia entre las dos?

Tomemos $p=1$ un único atributo $N=1$ y asumamos un modelo sin intercepto $\beta_0=0$ (se puede centralizar para evitar el intercepto)

Ridge $L_2 = \sum (y - x\beta)^2 + \lambda \beta$

Minimizar: Derivamos

$$\sum (2y - 2x\beta)(-x) + 2\lambda\beta = 0$$

$$-2yx + 2x^2\beta + 2\lambda\beta = 0$$

$$\boxed{\beta = \frac{\sum yx}{\lambda + \sum x^2}}$$

β podría hacerse pequeño pero no se hará nulo

$\lambda \to \infty$

$\beta \to 0$

Lasso $L_1 = (y - x\beta)^2 + \lambda|\beta$

si $\beta > 0$

$$= (y - x\beta)^2 + \lambda\beta$$

Derivamos

$$2(y - x\beta)(-x) + \lambda = 0$$

$$-2yx + 2x^2\beta + \lambda = 0$$

$$\boxed{\beta = \frac{2yx - \lambda}{2x^2}}$$

β puede dar cero!

En conclusión, con el método de Ridge las pendientes (β) se "encogen" pero no se anularán.

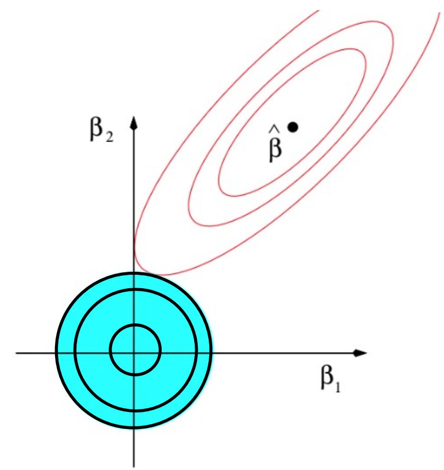
En cambio al usar Lasso, algunas variables se pueden anular y esto hace que algunas variables desaparezcan de la ecuación del modelo.

Intuición Geométrica

Caso $p=2$ $N=1$ sin intercepto

$$L_2: \underbrace{(y-\beta_1 x_1-\beta_2 x_2)^2}_{\text{Elipses centradas en el mínimo de RSS}} + \underbrace{\beta_1^2+\beta_2^2}_{\text{Círculos}}$$

Si las elipses no están centradas en ningún eje, no interceptarán a un círculo en un eje



Intuición Geométrica

Caso $p=2$ $\quad N=1$ $\quad$ sin intercepto

$$L_2 : \underbrace{(y - \beta_1 x_1 - \beta_2 x_2)^2}_{\text{Elipses centradas en el mínimo de RSS Mas no centradas en los ejes}} + \underbrace{\beta_1^2 + \beta_2^2}_{\text{Círculos}}$$

$$L_1 : (y - \beta_1 x_1 - \beta_2 x_2)^2 + \underbrace{|\beta_1| + |\beta_2|}$$

Aunque la elipse no esté centrada en un eje, puede interceptar al rombo en un eje.