Introducción a Machine Learning

Regresion

Ronald Cárdenas Acosta

Agosto, 2016

Regresión Lineal

Hipótesis

$$h_w(x) = w_0.1 + w_1.x_1 + ... + w_M.x_M$$

- $y^i \in \mathbb{R}$, ej: precio de una casa, riesgo crediticio, energia producida, etc.
- Objetivo: minimizar función de costo o función objetivo (errores cuadráticos)

$$L(w) = \frac{1}{2 * M} \sum_{i=1}^{N} (h_w(x^i) - y^i)^2 \quad (1)$$

Regresión Lineal: Ecuación Normal

- En forma vectorial: $Y = X * W$
- Despejando: $W = (X^T X)^{-1} X^T Y$
  Donde $(X^T X)^{-1} X^T$ es la inversa generalizada de X
- Cuando $X^T.X$ no es invertible:
  - $X^T.X$ es matriz singular o degenerada
  - Algunos features son linealmente dependientes entre si
  - Se utiliza la pseudo-inversa de $X^T.X$
- Costo computacional: $O(M^3)$
- No es escalable: no aplicable para $M > 100$

Optimización basada en Gradientes

- Usa un grupo pequeño de muestras (batch) o una muestra (online) para actualizar $\hat{w}$
- Procedimiento por defecto cuando M o N tienen ordenes mayores a 10^3
- Algorimos mas usados
  - Gradient Descent/Ascent
  - Stochastic Gradient Descent/Ascent
  - AdaGrad, ADAM (populares en Deep Learning)
  - Alg. basados en momentos, decaimiento, entre otros.

Gradient Descent

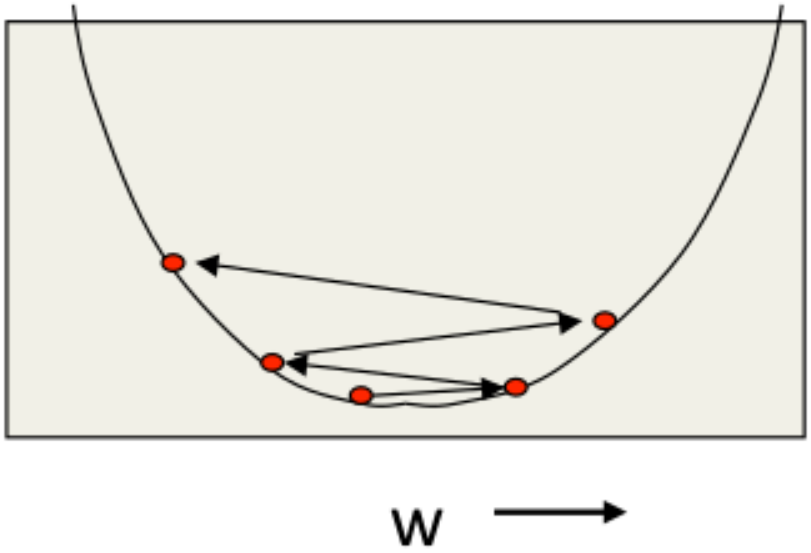
Para cada muestra:

- $w = w - \alpha * \frac{dL}{dw}$

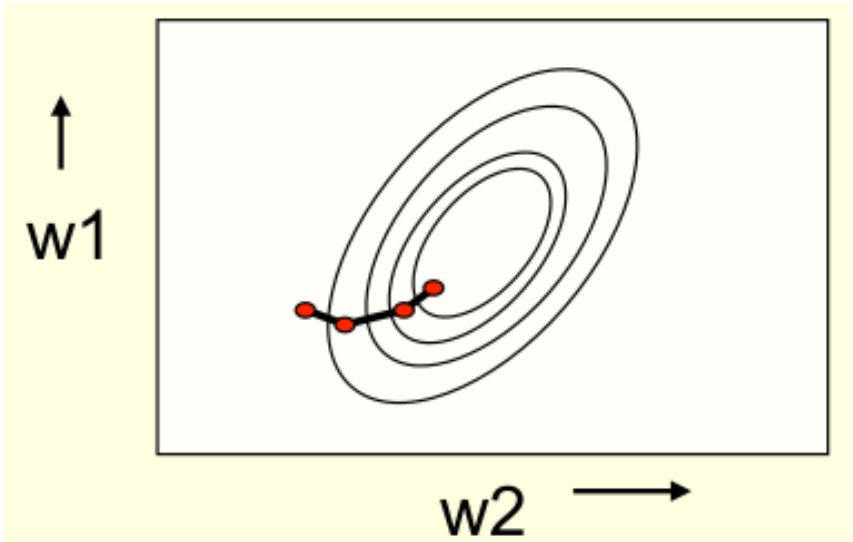
Donde α es la velocidad de aprendizaje:

- α muy bajo: entrenamiento lento
- *alpha* muy alto: algoritmo puede no converger

Gradient Descent



(a) Para $h_w(x) = 1 + w_1 * x_1$



(b) Para
$h_w(w) = 1 + w_1(x_1) + w_2 * x_2$