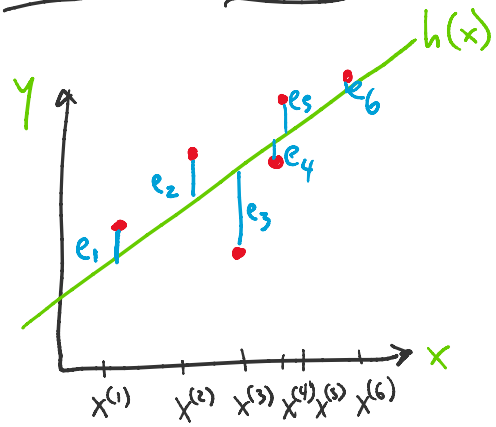
Section B: Linear Regression + Gradient Descent

Wednesday, September 21, 2022 12:49 PM

Residuals

$D = \{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ $\quad \hat{y} = h_{\vec{\theta}}(\vec{x})$

Def: a residual is the vertical distance from observed output $y^{(i)}$ to predicted output $\hat{y} = h(\vec{x}^{(i)})$



$$e_i = |y^{(i)} - h(\vec{x}^{(i)})|$$
$$= |y^{(i)} - (\vec{w}^T \vec{x}^{(i)} + b)|$$

Key Idea for Lin. Reg.

find the linear function h (w/parameters $\vec{w}$ and b) that minimizes the squares of the residuals for a training dataset

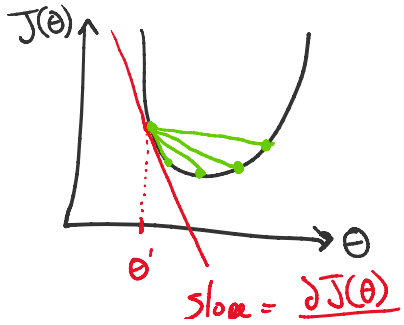
(the squares of: one option among many)

Def: mean squared error (MSE)

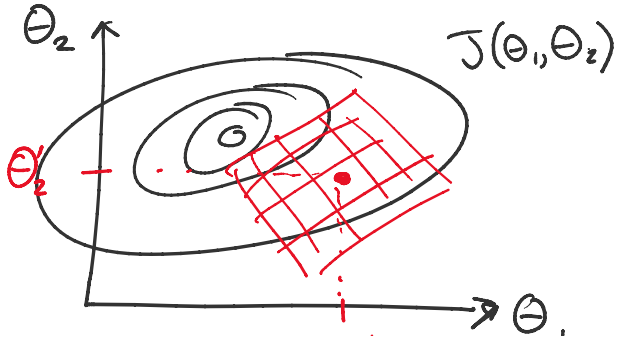
$$J_D(\vec{w}, b) = \frac{1}{N} \sum_{i=1}^{N} (e_i)^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - (\vec{w}^T \vec{x}^{(i)} + b)\right)^2$$

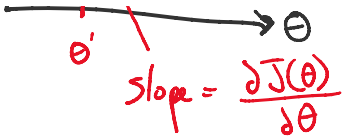
Derivatives

① Deriv. as Slope of Tangent



③ Partial Deriv. as Tangent Plane





② Deriv. as a Limit

$$\frac{\partial J(\theta)}{\partial \theta} = \lim_{e \to 0} \frac{J(\theta+e) - J(\theta)}{e}$$

↑ limit of secants is the tangent

$$\begin{bmatrix} \frac{\partial J(\theta_1,\theta_2)}{\partial \theta_1} \\ \frac{\partial J(\theta_1,\theta_2)}{\partial \theta_2} \end{bmatrix}$$

Gradient

Def: the gradient of $J: \mathbb{R}^M \to \mathbb{R}$

$$\nabla J(\vec{\theta}) = \begin{bmatrix} \partial J(\vec{\theta})/\partial \theta_1 \\ \partial J(\vec{\theta})/\partial \theta_2 \\ \vdots \\ \partial J(\vec{\theta})/\partial \theta_M \end{bmatrix}$$

first order partial derivative

Gradient Descent

Algorithm:

① Choose an initial $\vec{\theta}$

② Repeat $t = 1, 2, 3, \ldots$

a) Compute gradient $\vec{g} = \nabla J(\vec{\theta})$

b) Choose a step size γ_t

c) Update $\vec{\theta} \leftarrow \vec{\theta} - \gamma_t \vec{g}$

③ Return θ when stopping criterion is reached

Initial Point Step Size Stopping Criterion

Initial Point
- randomly
- $\vec{\theta}$ = all zeros

Step Size
- fixed value $\gamma = 0.1$
- schedule $\gamma_t = \frac{\gamma_0}{(t-1)\gamma + 1}$

Stopping Criterion
- when $\vec{g} \approx 0$
- $\|\nabla J(\theta)\|_2 < \epsilon$ for $\epsilon = 10^{-8}$

Gradient for Lin. Reg.

MSE: $J(\vec{\theta}) = \frac{1}{N}\sum_{i=1}^{N} J^{(i)}(\theta)$ where $J^{(i)}(\theta) = \frac{1}{2}\left(y^{(i)} - (\vec{\theta}^T \vec{x}^{(i)})\right)^2$

$\frac{1}{2}$ does not change the argmin

Partial Deriv.s

$$\frac{\partial J^{(i)}(\vec{\theta})}{\partial \theta_j} = \frac{1}{2} 2 \left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right)$$

$$= \left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(y^{(i)} - \sum_{m=1}^{M} \theta_m x_m^{(i)}\right)$$

$$= -\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) x_j^{(i)}$$

Gradient:

$$\nabla J^{(i)}(\vec{\theta}) = \begin{bmatrix} \partial J^{(i)}(\theta)/\partial \theta_1 \\ \vdots \\ \partial J^{(i)}(\theta)/\partial \theta_M \end{bmatrix} = \underbrace{\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right)}_{\text{scalar}} \underbrace{\vec{x}^{(i)}}_{\text{vector}}$$

$$\nabla J(\theta) = \nabla\left(\frac{1}{N}\sum_{i=1}^{N} J^{(i)}(\vec{\theta})\right) = \frac{1}{N}\sum_{i=1}^{N} \nabla J^{(i)}(\theta)$$

$$= \frac{1}{N}\sum_{i=1}^{N} -\left(y^{(i)} - \vec{\theta}^T \vec{x}^{(i)}\right) \vec{x}^{(i)}$$

know b/c dataset

known b/c we passed it in

$$D = \{(\vec{x}^{(1)}, y^{(1)}), (\vec{x}^{(2)}, y^{(2)}), \ldots, (\vec{x}^{(N)}, \vec{y}^{(N)})\}$$

$$\vec{x}^{(2)} = [x_1^{(2)} \quad x_2^{(2)} \quad \cdots \quad x_M^{(2)}]$$

$$h(\vec{x}) = \vec{\theta}^T \vec{x}$$

$$y = w_1 x_1 + b$$

$$y = w_1 x_1 + w_2 x_2 + b$$

$$y = \vec{w}^T \vec{x} + b$$

intercept $= 1$ slope

$$y = \theta_1 x_1 + \theta_2 x_2$$

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

θ_1: intercept; θ_2: slope$_2$; θ_3: slope$_3$

$$y = \vec{\theta}^T \vec{x}$$