Section B

Wednesday, February 15, 2023 12:12 PM

Optimization for Linear Models

MSE for Linear Reg:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \vec{\theta}^T\vec{x}^{(i)}\right)^2 \longrightarrow \nabla J(\theta) = \cdots \longrightarrow GD$$

MSE for Perceptron:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \underbrace{\operatorname{sign}(\vec{\theta}^T\vec{x}^{(i)})}_{\hat{y}^{(i)}}\right)^2 \longrightarrow \nabla J(\theta) = \boxed{FAIL}$$

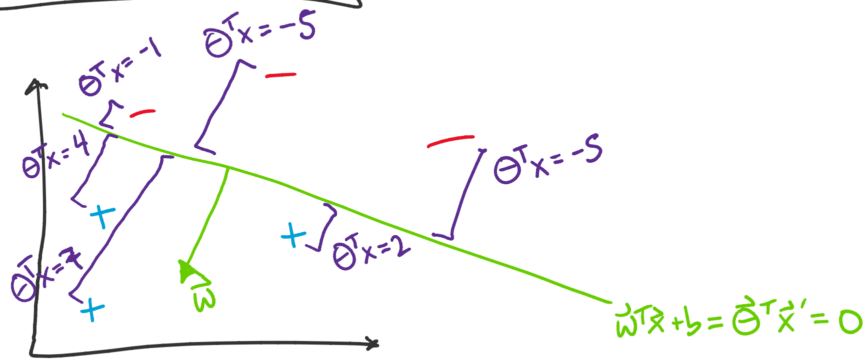
$\hat{y}^{(i)}$ not differentiable

MSE for Logistic Regression

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \sigma\left(\vec{\theta}^T\vec{x}^{(i)}\right)\right) \longrightarrow \nabla J(\vec{\theta}) = \cdots \longrightarrow GD$$

chosen to be differentiable

What is $\vec{\theta}^T\vec{x}$?



Binary Logistic Regression

① Model:

$y \sim \text{Bernoulli}(\phi)$

$\phi = \sigma(\vec{\theta}^T\vec{x})$ where $\sigma(u) = \dfrac{1}{1+\exp(-u)}$

$$p(y \mid \vec{x}, \theta) = \begin{cases} \sigma(\vec{\theta}^T\vec{x}) & \text{if } y = 1 \\ \underbrace{1 - \sigma(\vec{\theta}^T\vec{x})}_{1-\phi} & \text{if } y = 0 \end{cases}$$

$\log(a \cdot b) = \log a + \log b$

② Objective:

$$\ell(\vec{\theta}) = \log p(D \mid \vec{\theta}) = \log \prod_{i=1}^{N} p(y^{(i)} \mid \vec{x}^{(i)}, \vec{\theta})$$

$$= \sum^{N} \log p(y^{(i)} \mid \vec{x}^{(i)}, \theta)$$

★ negative average conditional log-likelihood, $J(\theta)$ for Log. Reg. is convex!

$$= \sum_{i=1}^{N} \log p(y^{(i)} | \vec{x}^{(i)}, \theta)$$

$$J(\vec{\theta}) = -\frac{1}{N}\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{-\log p(y^{(i)} | x^{(i)}, \theta)}_{J^{(i)}(\theta)}$$

likelihood, $J(\theta)$ for Log. Reg. is convex!

$$\theta_{MLE} = \operatorname{argmax} \ell(\theta)$$
$$= \operatorname{argmin} -\ell(\theta)$$
$$= \operatorname{argmin} -\frac{1}{N}\ell(\theta)$$

③ Derivatives

$$\frac{\partial J^{(i)}(\vec{\theta})}{\partial \theta_m} = \frac{\partial}{\partial \theta_m} -\log\left(p(y^{(i)} | \vec{x}^{(i)}, \vec{\theta})\right)$$

$$= \begin{cases} \partial/\partial\theta_m -\log(\sigma(\vec{\theta}^T \vec{x}^{(i)})) & \text{if } y^{(i)} = 1 \\ \partial/\partial\theta_m -\log(1 - \sigma(\vec{\theta}^T \vec{x}^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$$

$= \dots$

$= \dots$

$= \dots$

recitation or HW

$$= -\big(\underbrace{y^{(i)}}_{\text{true value}} - \underbrace{\sigma(\vec{\theta}^T \vec{x}^{(i)})}_{\text{prob. of } y^{(i)}=1}\big)\underbrace{x_m^{(i)}}_{\text{mth feature}}$$

$$\nabla J^{(i)}(\theta) = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} = -(y^{(i)} - \sigma(\theta^T x^{(i)}))\vec{x}^{(i)}$$

$$\nabla J(\vec{\theta}) = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N} \nabla J^{(i)}(\theta)$$

④ Find $\vec{\theta}$ by gradient descent or SGD

⑤ Predict the most probable class, for new $\vec{x}$

$$\hat{y} = \operatorname*{argmax}_{y \in \{0,1\}} p(y | \vec{x}, \vec{\theta})$$

$$= \begin{cases} y = 1 & \text{if } p(y=1 | \vec{x}, \vec{\theta}) \geq 0.5 \\ y = 0 & \text{otherwise} \end{cases}$$

$$= \text{"sign"}(\vec{\theta}^T \vec{x})$$

Exercise: Prove that this is true

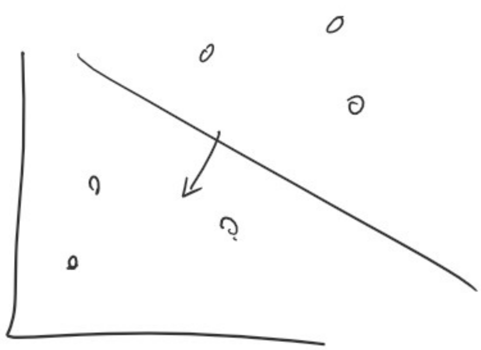
OH

$$w^T x + b = 0$$

$$x_2 = \underline{\left(-\frac{w_1}{w_2}\right)} x_1 + \underline{\left(-\frac{b}{w_2}\right)}$$

repeat

1. pick hyp. α, β
2. train w/ α and β
3. report val error



$$\left(y^{(i)} - \left(\theta_1 x_1 + \theta_2 x_2\right)\right)^2$$

$$\left(7 - \left(\theta_1 3 + \theta_2 \cdot 2\right)\right)^2$$