COL774
Machine Learning
Sep 1, 2021

Last class:–

- Newton's Method.] using second order information for optimization

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

$y \in \mathbb{R}$ ← $h_\theta(x) = \theta^T x$

Locally weighted linear Regression] :- non parametric methods

First Classification Algorithm:–

$x \in \mathbb{R}^{(n+1)}$

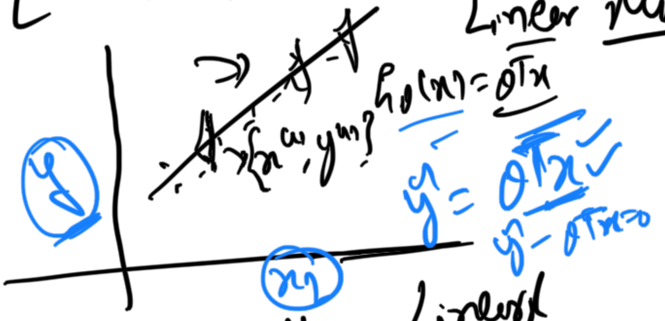
$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

$y \in$ Discrete Set ↗ $\mathbb{R}$ (regression)

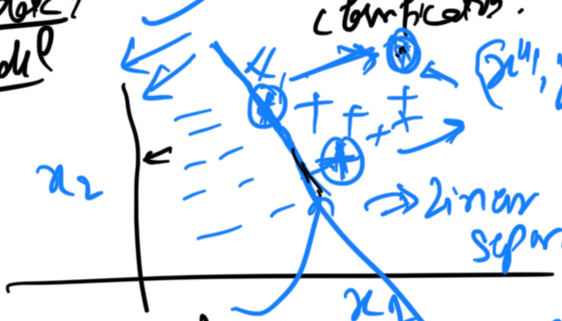
$\{0, 1\}$ ⇐ generalize to multi-class classification.

[What kind of models can be learned over this data?

Linear Model



Regression Linear → Logistic Regression (Classification)



⇒ separator "Decision" Boundary

$\theta^T x = 0$ ⇐ hyperplane $x \in \mathbb{R}^{(n+1)}$

{ $y = mx + c$
{ $y - mx - c = 0$

$\theta_2 x_2 + \theta_1 x_1 + \theta_0 = 0$

$$x_2 = \frac{-(\theta_1 x_1 + \theta_0)}{\theta_2}$$

[$y^{(i)} \mid x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$

⇓

$\max_\theta LL(\theta) \equiv$ Least squares

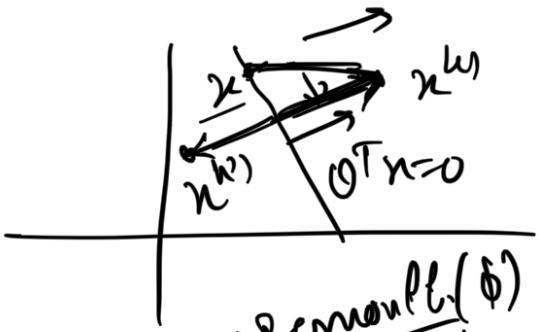
⇒ Classification:-

$y^{(i)} | x^{(i)}; \theta \sim$ Bernoulli $(g(\theta^T x^{(i)}))$ $\in [0,1]$

$\in \{0,1\}$ ~ Be $\downarrow \phi$

⇒ Probability of a point (x to this (signed) being true distance of the point from the line

$\phi \in [0,1]$:- $\theta^T x = 0$



$\theta^T (x^{(i)} - x) =$ unnormalized dist

$\theta^T x^{(i)} - \theta^T x$

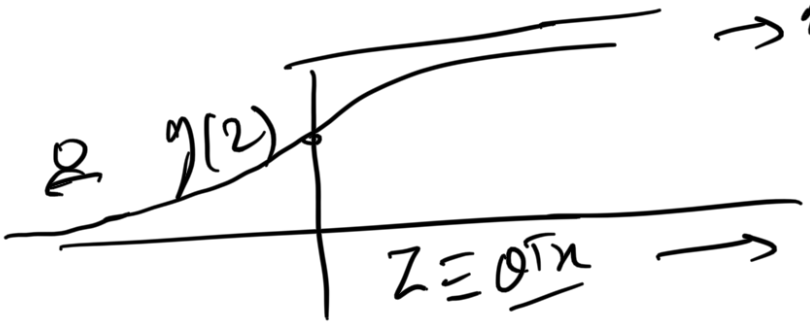
0

$y^{(i)} | x^{(i)}; \theta \sim$ Bernoulli(ϕ)

$\theta^T x^{(i)}$ (signed) $\theta^T x^{(i)} \equiv$ unnormalized distance of $x^{(i)}$ from $\theta^T x = 0$

$\phi = g\left[\theta^T x^{(i)}\right] \in [0,1]$

⇒ Larger the distance from (signed) the hyperplane → 2



greater is ϕ

S-shaped] logistic

$\phi = g(z) = \frac{1}{1+e^{-z}}$ sigmoid / logistic

$z = \theta^T x$

$\phi^{(i)}$

$y^{(i)} | x^{(i)}; \theta \sim \text{Bernoulli}(g(\theta^T x^{(i)}))$

$p(y^{(i)}=1 | x^{(i)}; \theta) = g(\theta^T x^{(i)}) = \begin{cases} \frac{1}{1+e^{-\theta^T x^{(i)}}} \equiv \phi^{(i)} \\ \text{if } \theta^T x = 0 \end{cases}$

$y \sim \text{Bernoulli}(\phi) \quad y \in \{0,1\}$

$p(y=1) = \phi$

$p(y=0) = 1-\phi$

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m} \quad i = 1 \dots m$

$(y^{(i)} | x^{(i)}; \theta) \sim \text{Bernoulli}(\phi^{(i)})$

$\Downarrow$

$g(\theta^T x^{(i)})$

$\Rightarrow$ log-likelihood of the training data under the above modeling assumptions.

$p(y^{(1)} \dots y^{(m)} | x^{(1)} \dots x^{(m)}; \theta)$

$= \prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; \theta) \equiv$ likelihood of the data

$\Rightarrow \log \prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; \theta)$

$\sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta) \quad -(1)$

$= \sum_{i=1}^{m}$ [...]

$\hookrightarrow y^{(i)}=1 \mid \frac{1}{1+e^{-\theta^T x^{(i)}}}$

$\hookrightarrow y^{(i)}=0 \mid 1 - \frac{1}{1+e^{-\theta^T x^{(i)}}} = \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}$

$y \in \{0,1\}$ → indicator fn

$1\{y^{(i)}=1\}$ Boolean expression → 1 if exp is true

$$= \sum_{i=1}^{m} \Big[y^{(i)} \log P(y^{(i)}=1 \mid x^{(i)}; \theta) + (1-y^{(i)}) \log P(y^{(i)}=0 \mid x^{(i)}; \theta) \Big]$$

$1\{y^{(i)}=0\}$

$1\{\langle \text{boolean} \rangle\}$ → 1 if exp is true, 0 if exp is false

$$LL(\theta) = \sum_{i=1}^{m} \Big[y^{(i)} \log \Big[\frac{1}{1+e^{-\theta^T x^{(i)}}} \Big] + (1-y^{(i)}) \log \Big[\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}} \Big] \Big]$$

log-likelihood expression

$$1 - \frac{1}{1+e^{-\theta^T x^{(i)}}} = \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}$$

What is the set of parameters that we are interested in finding?

⇒ maximize the likelihood (log) of generating the observed $y^{(i)} | x^{(i)}$

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\ LL(\theta)$$

$$\Rightarrow \underset{\theta}{\operatorname{argmax}} \frac{1}{2m} \sum_{i=1}^{m} y^{(i)} \log \frac{1}{[1+e^{-\theta^T x^{(i)}}]} + (1-y^{(i)}) \left[\log \frac{e^{-\theta^T x^{(i)}}}{(1+e^{-\theta^T x^{(i)}})} \right]$$

⇒ How to find θ's?

↳ Gradient Descent.

$$\nabla_\theta LL(\theta)$$

$$\frac{1}{2m} \nabla_\theta \left[\sum_{i=1}^{m} y^{(i)} \log \frac{1}{[1+e^{-\theta^T x^{(i)}}]} + [1-y^{(i)}] \log \underbrace{\frac{[e^{-\theta^T x^{(i)}}]}{[1+e^{-\theta^T x^{(i)}}]}}_{B} \right]$$ ← A

$$= \frac{1}{2m} \nabla_\theta \left[\sum_{i=1}^{m} y^{(i)} \log \left[\frac{1}{1+e^{-\theta^T x^{(i)}}} \right] + [1-y^{(i)}] \log \left[\frac{1}{1+e^{-\theta^T x^{(i)}}} \right] + [1-y^{(i)}] \log [e^{-\theta^T x^{(i)}}] \right]$$

$$= \frac{1}{2m} \nabla_\theta \left[\sum_{i=1}^{m} \log\left[\frac{1}{1+e^{-\theta^T x^{(i)}}}\right] + \left[1-y^{(i)}\right]\left(-\theta^T x^{(i)}\right) \right]$$

$$= \frac{1}{2m} \nabla_\theta \left[\sum_{i=1}^{m} -\log\left[1+e^{-\theta^T x^{(i)}}\right] + \left(1-y^{(i)}\right)\left(-\theta^T x^{(i)}\right) \right]$$

$$= \frac{1}{2m} \sum_{i=1}^{m} -\nabla_\theta \log\left[1+e^{-\theta^T x^{(i)}}\right] + \left(1-y^{(i)}\right) \nabla_\theta\left(-\theta^T x^{(i)}\right)$$

$$= \frac{1}{2m} \sum_{i=1}^{m} -\left[\frac{1 \cdot \left[e^{-\theta^T x^{(i)}}\right](-x^{(i)})}{\left[1+e^{-\theta^T x^{(i)}}\right]} + \left(1-y^{(i)}\right)\left[-x^{(i)}\right]\right]$$

$$= \frac{1}{2m} \sum_{i=1}^{m} \left[1-y^{(i)} - \left[\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right]\right]\left[-x^{(i)}\right]$$

$$= \frac{1}{2m} \sum_{i=1}^{m} \left[\cancel{e^{-\theta^T}} \left[1 - \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}\right] + y^{(i)} \right] \left[+x^{(i)}\right]$$

$$\nabla_\theta L(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right] x^{(i)}$$