CS274

Machine Learning

Aug 28, 2021

Last class:-

Analytical solutions for $J(\theta)$

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

$h_\theta(x^{(i)}) \Downarrow \theta^T x^{(i)}$

$$\Rightarrow \frac{1}{2m}\left[(Y-X\theta)^T(Y-X\theta)\right] = J(\theta)$$

$$\nabla_\theta J(\theta) = 2X^TX\theta - 2X^TY$$

Equating $\nabla_\theta J(\theta)=0$

$$\Rightarrow 2X^TX\theta = 2X^TY$$

what if not invertible?

Analytical solution for least squared (linear) regression

$$\Rightarrow \boxed{\theta = \left[(X^TX)^{-1}X^T\right]Y}$$

expensive

pseudo inverse.

$$\left[(X^TX)^{-1}X^T\right]X = I$$

$O(\theta n$

$O(n^3)$

$\simeq$ cubic in $(n \times m)$

Why not use analytical solution?

$$(X^TX + \alpha I) \rightarrow$$

$\alpha \leq 1$

identity

$\Rightarrow$

Normalization of Data :-

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

For $j = 1$ to $n$ {

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)},$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2}{m}$$

$\forall i :-$ $\left[ x_j^{(i)} \right] \leftarrow \left[ \frac{x_j^{(i)} - \mu_j}{\sigma_j} \right]$

}

↳ normalize $x^{(i)}$'s to zero mean & const variance in each dimension.

$\Rightarrow$ The mean of resulting features is zero & standard deviation $= 1$ ✓

$$\mu_j' = \frac{1}{m} \sum_{i=1}^{m} \left[ \frac{x_j^{(i)} - \mu_j}{\sigma_j} \right]$$

$$= \frac{1}{\sigma_j m} \left[ m\mu_j - m\mu_j \right] = 0$$

$\cancel{\sigma_j^2 = \left[ x_j^{(i)} - \mu_j \right]^2 - \mu_j}$

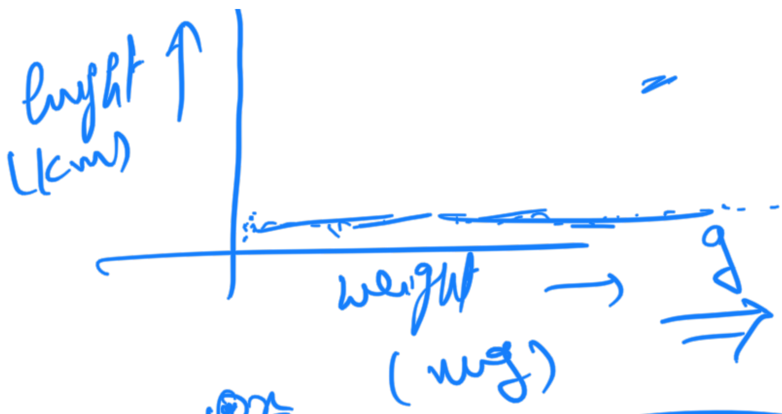
$$\sigma_j'^2 = \frac{1}{m} \sum_{i=1}^{m} \left[ x_j^{(i)'} - \underbrace{\mu_j'}_{0} \right]^2$$

$$\Rightarrow \frac{1}{m} \sum_{i=1}^{m} \left[ \frac{x_j^{(i)} - \mu_j}{\sigma_j} \right]^2$$

$$= \frac{1}{m\sigma_j^2} \cdot \left[ m\sigma_j^2 \right] = 1$$

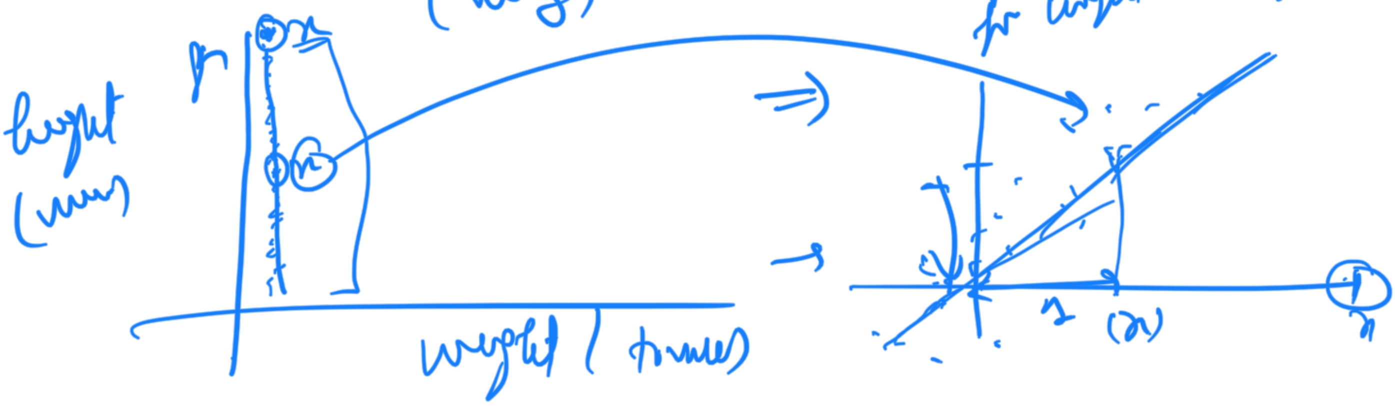
Why do this?

$\{ x^{(i)}, y^{(i)} \}_{i=1}^{m}$ ✓




The distribution become very skewed across dimensions

⟹ Many face challenge for weight tuning.




Imp:- 2 At test time:-

$x$: $\underline{h_\theta(x)}$

↳ $x \leftarrow \left[\underline{\left(\frac{x-\mu}{\sigma}\right)}\right] \Rightarrow h_\theta\left(\cancel{x = \frac{x-\mu}{\sigma}}\right)$

---

H.W: [Exam]

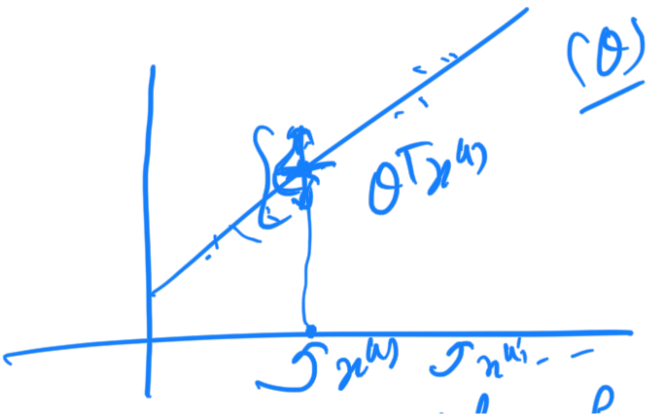
$h_\theta(x)$ Prediction is invariant under the transformation.

---

## Probabilistic Interpretation of Least Squares :-

→ A new class of model :-

ML estimate: Maximum Likelihood Estimation




$(\theta)$

Generative Model

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

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

Model :- how was data might have been generate?

$\Rightarrow$ suppose :- Data :- $\theta$ & linear model

$\forall i = 1$ to $m$ Given :- $[x^{(i)}]$ I

1. compute $\theta^T x^{(i)}$
2. $\epsilon^{(i)} \sim N(0, \sigma^2)$
3. $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$

$$\prod_{i=1}^{m} P(y^{(i)} | x^{(i)}; \theta) = \text{Likelihood function}$$

$$\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} P(y^{(i)} | x^{(i)}; \theta)$$

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

$$P(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y^{(i)} - \theta^T x^{(i)}]^2}{2\sigma}}$$

$$Z \sim N(\mu, \sigma^2) :- \quad P(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

$\Rightarrow$ Likelihood over all examples :-

$$P(y^{(1)} \ldots y^{(m)} | x^{(1)} \ldots x^{(m)}; \theta)$$

$$\overset{\text{i.i.d. assumption}}{=} \prod^{m} P(y^{(i)} | x^{(i)}; \theta) \Rightarrow$$

$i=1$

Likelihood

$$\rightarrow \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

log $\prod$

Find the parameter which maximize the likelihood

$$\arg\max_{\theta} \prod_{i=1}^{m} \left[ p(y^{(i)} \mid x^{(i)}; \theta) \right]$$

$$\equiv \arg\max_{\theta} \log \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$\equiv \arg\max_{\theta} LL(\theta)$$ :- likelihood seen as a function of $\theta$ parameter

$$\Rightarrow \arg\max_{\theta} \log \prod_{i=1}^{m} \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}} \right]$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \left[ \underbrace{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)}_{\text{constant wrt } \theta} + \log e^{\frac{(y^{(i)} - \theta^T x^{(i)})^2}{-2\sigma^2}} \right]$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \cancel{\log}\ e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

$$= \underset{\theta}{\arg\max} \; - \frac{\sum \left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}$$

$$= \underset{\theta}{\arg\max} \; - \sum_{i=1}^{m} \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2}$$

$$= \underset{\theta}{\arg\max} \; \sum_{i=1}^{m} -\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2m}$$

$$= \underset{\theta}{\arg\min} \; \underbrace{\sum_{i=1}^{m} \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2m}}$$

$$= \underset{\theta}{\arg\min} \; J(\theta) \quad :)$$