HW 2

Gawun Kim

① $X^TXw = X^Ty$ where $X \in R^{N\times M}$, $y \in R$

Prove : X^TX is non-singular if and only if X has linearly independent columns.

ⓐ * With $N \gg M$, there will likely be M linearly indep vector X_n

* Gram Matrix X^TX is likely to be invertible.

ⓑ * X has linearly independent columns if $AX=0$ only for $x=0$

ⓒ * $Ax=0$ has only trivial Solution $(x=0)$ if and only if A is non-singular

If A is singular, there exists $z\neq 0$ sh that $Az=0$

✓ Let $K = X^TX$,

if Column vectors of X are linearly independent, then there is a vector to satisfy ($Xu=0$ when $u\neq 0$)

So that $X^TXu = 0 = Ku$. Since $u\neq 0$, K is not invertible.

✓ K is not invertible and there exists $(V\neq 0)$ vector (V) so that $KV=0$.

→ * $V^TKV = \underbrace{V^TX^T}_{= V^TX^T = (XV)^T}XV = (XV)^TXV = \|XV\|^2 = 0$

→ so that $XV=0$

since $XV=0$ and $V\neq 0$, the column of vectors of A are linearly indep

② Hat Matrix $H = X(X^TX)^{-1}X^T$, where $X\in R^{N\times M}$ and X^TX is invertible

ⓐ show H is Symmetric $(H=H^T)$

$$(H)^T = (X(X^TX)^{-1}X^T)^T$$
$$= (X^T)^T((X^TX)^{-1})^TX^T = (X)\,((X^TX)^T)^{-1}(X^T)$$

* $(X^TX)^T = (X)^T(X^T)^T = X^TX$

$$= X(X^TX)^{-1}X^T$$

∴ $H = H^T$ proved.

ⓑ Show $H^K = H$

$H = X(X^TX)^{-1}X^T$

when $K=2$ $H^2 = HH = [X(X^TX)^{-1}X^T][X(X^TX)^{-1}X^T]$

$= X\cancel{(X^TX)^{-1}}\cancel{(X^TX)}(X^TX)^{-1}X^T$

$= X(X^TX)^{-1}X^T = H$

When $K=\alpha$ (α is some integer, which is bigger than 2)

$H^\alpha = [\cancel{X(X^TX)^{-1}X^T}][\cancel{\quad}]\cdots[\cancel{\quad}][\cancel{X}(X^TX)^{-1}X^T]$

$= X(X^TX)^{-1}X^T = H$

∴ $H^K = H$

ⓒ Show $(I-H)^K = (I-H)$

$I-H = I - X(X^TX)^{-1}X^T$

$(I-H)^K$

When K is 2 $(I-H)^2 = I^2 - 2\cdot H + \overbrace{(H)^2}^{\text{by Question ⓑ}} = I - 2H + H = \underline{I-H}$

$(I-H)^K$ when K is some big positive integer.

$(I-H)^K = (I-H)(I-H)(I-H)(I-H)\cdots\cdots(I-H)$

$= \underbrace{(I-H)(I-H)}_{\text{it is } (I-H) \text{ from above.}}(I-H)\cdots(I-H)$

$= (I-H)(I-H)\cdots(I-H)$

$=$ $(I-H)$

∴ $(I-H)^K = I-H$

ⓓ show Trace$(H) = M$

According to question ⓐ, H is symmetric and invertible.

$HH = H$ and $HH^T = H^TH$ by hint and question ⓑ

So that H should be identify matrix.

So that, in H, each of diagonal elements is 1

and the summation of all diagonal element is

the number of column (because $N \gg M$ from question 1)

③ $J(w_0, w_1) = \sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n)^2$ → Calculate the gradient by computing partial derivative to each of parameters (w_0, w_1)

$\alpha_n > 0$

derivative w.r.t w_0

$$\frac{d}{dw_0}J(w_0,w_1) = \sum_{n=1}^{N}(\alpha_n)(2)(w_0 + (w_1x_n) - y_n)$$
$$= \boxed{\sum_{n=1}^{N} 2\alpha_n(w_0 + w_1x_{n1} - y_n)}$$

derivative w.r.t w_1

$$\frac{\partial}{\partial w_1}J(w_0,w_1) = \sum_{n=1}^{N}(\alpha_n)(2)(x_{n,1})(w_0 + w_1x_{n,1} - y_n)$$
$$= \boxed{\sum_{n=1}^{N} 2\alpha_n x_{n,1}(w_0 + w_1x_{n,1} - y_n)}$$

$$\nabla J(w_0, w_1) = \left[\sum_{n=1}^{N}2\alpha_n(w_0 + w_1x_{n,1} - y_n) \quad \sum_{n=1}^{N}2\alpha_n x_{n,1}(w_0 + w_1x_{n,1} - y_n)\right]^T$$

When α_n is 0, the gradient of $J(w_0,w_1)$ ends up being 0

it means that the rate of change in the function is zero

If α_j is much bigger than other $\alpha_{j'}$, $j\neq j'$ in gradient descent.

Training data does not throw out any data.

and more computationally efficient.

④ $J(w) = -\sum_{n=1}^{N}[y_n \log h_w(x_n) + (1-y_n)\log(1-h_w(x_n))] + \frac{1}{2}\sum_i w_i^2$

$h_w(x_n) = \sigma(w^Tx_n) = \dfrac{1}{1+\exp(-w^Tx_n)} = f(w)$

$\dfrac{\partial}{\partial w}[\log(\sigma(w^Tx))] = \dfrac{1}{\sigma(w^Tx)}\cdot\dfrac{\partial(\sigma(w^Tx))}{\partial w}$

* $f'(w) = \dfrac{\partial}{\partial w}(w^Tx) = \dfrac{x\,e^{-w^Tx}}{(1+e^{-w^Tx})^2}$

✓ $\dfrac{d(\log f(w))}{dw} = \dfrac{1}{\sigma(w^Tx)}\cdot\dfrac{\partial}{\partial w}(\sigma(w^Tx))$

$= \dfrac{1}{\sigma(w^Tx)}\quad\dfrac{e^{-w^Tx}}{(1+e^{-w^Tx})^2}x$

$= \dfrac{1}{\sigma(w^Tx)}\cdot\dfrac{1}{(1+e^{-w^Tx})}\cdot\dfrac{e^{-w^Tx}}{(1+e^{-w^Tx})}x$

$= \cancel{\dfrac{1}{\sigma(w^Tx)}}\ \cancel{\sigma(w^Tx)}\cdot\dfrac{1+e^{-w^Tx}-1}{(1+e^{-w^Tx})}x$

$= [1-\sigma(w^Tx)]\,x$

✓ $\dfrac{\partial(\log(1-f(w)))}{\partial w} = -\sigma(w^Tx)\,x$

$\dfrac{\partial J(w)}{\partial w_j} = -\sum_{n=1}^{N}[y_n(1-\sigma(w_j^Tx))x + (1-y)(-\sigma(w_j^Tx))x]$

$+ \dfrac{d(\sum_i w_i^2)}{\partial w}$

$= -\sum_{n=1}^{N} x(y_n - \sigma(w_i^Tx)) + (\cancel{\quad})$

$$\boxed{\frac{\partial(J(w))}{\partial w_j} = -\sum_{n=1}^{N} x(y_n - \sigma(w_i^Tx)) + w_j}$$

→ since derivative by w_j, others will be consider as constant

In each element, there is unique (one) corresponded weight and get updated

so it derives the gradient update rules for weight

⑤ $\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0, I)$

For $z \sim \mathcal{N}(0, I_m)$

$$f(w) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\sum_{i=1}^{m} \left(\frac{w_i^2}{2}\right)\right)$$

$$f_z(z) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\sum_{i=1}^{m} \frac{(z_i)^2}{2}\right)$$

$$w^* = \arg\max_w \left[\prod^{n} P(y_i | x_i, w)\, f(w)\right] \quad ※\ P(y_i | x_i\, w) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma_i^2}\right)$$

$$= \arg\max_w \left(\text{const} - \frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{\sigma_i^2}\right) f(w)$$

$f(w) \geq 0$

So that

$$= \arg\min_w \left[\left(\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{\sigma_i^2}\right) f(w)\right]$$

$$= \arg\min \left[\cancel{\frac{1}{2}}\sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{\sigma_i^2}\right)\left(\cancel{\frac{1}{(2\pi)^{m/2}} \exp\left(-\sum_{i=1}^{m}\left(\frac{w_i^2}{2}\right)\right)}\right]$$

It is alway bigger than 0

$$= \arg\left(\sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{\sigma_i^2}\right)$$

So This is the identical objective.

a $J(w)$ for weighted least squares with $\alpha_n = 1/\sigma_n^2$