CS541 a3

Chenxu Wang

10457625

Gradient Calculation:

Suppose $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ are known.

Calculate the gradient of the following functions.

- Sigmoid function: $F(w) = \dfrac{1}{1+e^{-x \cdot w}}$

Since $\left(\dfrac{1}{u}\right)' = \dfrac{-u'}{u^2}$

Using Chain Rule:

$$\frac{\partial F(w)}{\partial w} = \frac{-e^{-xw} \cdot (-x)}{(1+e^{-x \cdot w})^2} = \frac{x \cdot e^{-xw}}{(1+e^{-xw})(1+e^{-xw})}$$

$$= \frac{x}{(1+e^{-xw})} \cdot \frac{1+e^{-xw}-1}{(1+e^{-xw})} = x \cdot \frac{1}{1+e^{-xw}} \cdot \left(1 - \frac{1}{1+e^{-xw}}\right)$$

$$= x \cdot F(w) \cdot (1 - F(w))$$

∘ Logistic loss $F(w) = \log(1 + e^{-yx \cdot w})$

Since $(\log u)' = \frac{1}{u}$

Using Chain Rule :

$$\frac{\partial}{\partial w} \log(1 + e^{-y \cdot x \cdot w})$$

$$= \frac{1}{1 + e^{-yxw}} \cdot e^{-yxw} \cdot (-y \cdot x)$$

$$= \frac{-y \cdot x}{e^{yxw} + 1}$$

Linear Regression:

1. Since we have, for vector $z \in \mathbb{R}^d$, $z^T \cdot z = \sum_{i=1}^{d} z_i^2$

$$\therefore \quad F(w) = \frac{1}{2} \| y - Xw \|_2^2$$

$$= \frac{1}{2} (y - X \cdot w)^T \cdot (y - Xw)$$

$$\nabla F(w) = \nabla \frac{1}{2} (y^T \cdot y - y^T \cdot Xw - w^T \cdot X^T \cdot y + w^T X^T \cdot X \cdot w)$$

$$= \frac{1}{2} (0 - y^T \cdot X - X^T \cdot y + X^T X \cdot w + w^T \cdot X^T X)$$

$$= \frac{1}{2} (-2 X^T y + 2 X^T X \cdot w)$$

$$= X^T X \cdot w - X^T y$$

Let $\nabla F(w) = 0$, we have:

$$X^T X \cdot w = X^T y$$

When $n > d$, $(X^T X)^{-1}$ exists:

$$\Rightarrow \quad w = (X^T X)^{-1} \cdot X^T y$$

For Hessian matrix of $F(w)$:

$$H(F) = \frac{\partial^2 F(w)}{\partial w \cdot \partial w^T} = \frac{\partial}{\partial w^T} \cdot \left(\frac{\partial F(w)}{\partial w} \right) = \frac{\partial}{\partial w^T} \cdot \nabla F(w)$$

$$= \frac{\partial}{\partial w^T} (X^T \cdot X \cdot W - X^T \cdot y) = X^T X$$

$$= \begin{bmatrix} \sum^n X_{i1}^2 & \sum^n X_{i1} \cdot X_{i2} & \cdots & \sum^n X_{i1} \cdot X_{id} \\ \sum^n X_{i2} \cdot X_{i1} & \sum^n X_{i2}^2 & \cdots & \sum^n X_{i2} \cdot X_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \sum^n X_{id} \cdot X_{i1} & \sum^n X_{id} \cdot X_{i2} & \cdots & \sum^n X_{id}^2 \end{bmatrix}_{d \times d}$$

The matrix $X^T \cdot X \geq 0$, is positive semidefinite, which shows that $F(w)$ is a convex program.

2.

When we using the least squares formulation, it is equal to calculate the maximum likelihood estimation for the samples.

we can prove it:

The samples are (x_i, y_i), the prediction is $\hat{y}_i|_w$, then we have $y = \hat{y} + \varepsilon$

we assume $\varepsilon \sim N(0, \sigma^2)$.

$$\therefore \quad y - \hat{y} \sim N(0, \sigma^2)$$

$$\Rightarrow \quad y \sim N(\hat{y}, \sigma^2)$$

The likelihood function is:

$$L(w) = P(y|x; w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma}\right)$$

$$\Rightarrow \log L(w) = n \log \frac{1}{\sqrt{2\pi}} + \sum_{i=0}^{n} -\frac{(y_i - \hat{y}_i)^2}{2\sigma}$$

If we want the maximum of $\log L(w)$, $\sum_{i=0}^{m} (y-\hat{y})^2$ is minimal.

2. The least squares formulation will larger progress for getting to the optimum W. than $\|y-\hat{y}\|_2^{100}$

So the least squares will faster when we iterate $w^t = w^{t-1} - \eta \cdot \nabla(w^{t-1})$ to get W_{opt}.

3.

When the rank of $X^T \cdot X$: $Rank(X^T X) = d$, then $H(F) = X^T X \geq m$ $(m > 0)$. $F(w)$ is strongly-convex.

otherwise, is not strongly-convex.