CSE517A

Machine Learning (SP18)

Written Homework I.

Chunyuan Li, Jianxi Xing

Problem I.

$$X = [\vec{x}_1, \vec{x}_2 \dots \vec{x}_n]^T = \begin{bmatrix} — \vec{x}_1^T — \\ \vdots \\ — \vec{x}_n^T — \end{bmatrix}$$

(a) $\because L(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda \|\vec{w}\|_2^2$

$$= \sum_{i=1}^{n} [(\vec{w}^T \vec{x}_i)^2 - 2y_i \vec{w}^T \vec{x}_i + y_i^2] + \lambda \vec{w}^T \vec{w}$$

$$= \sum_{i=1}^{n} [\vec{w}^T \vec{x}_i \vec{x}_i^T \vec{w} - 2y_i \vec{w}^T \vec{x}_i + y_i^2] + \lambda \vec{w}^T \vec{w}$$

$$\therefore \frac{\partial}{\partial \vec{w}} L(\vec{w}) = \sum_{i=1}^{n} (2\vec{x}_i \vec{x}_i^T \vec{w} - 2y_i \vec{x}_i) + 2\lambda \vec{w}$$

$$= \sum_{i=1}^{n} 2\vec{x}_i (\vec{x}_i^T \vec{w} - y_i) + 2\lambda \vec{w}$$

The matrix form is:

$$L(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda \|\vec{w}\|_2^2$$

$$= (X\vec{w} - Y)^T (X\vec{w} - Y) + \lambda \vec{w}^T \vec{w}$$

$$= \vec{w}^T X^T X \vec{w} - 2\vec{w}^T X^T Y + Y^T Y + \lambda \vec{w}^T \vec{w}$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = 2X^T X \vec{w} - 2X^T Y + 2\lambda \vec{w}$$

(b) Note $\nabla \vec{w}$:

$$\nabla \vec{w} = \frac{\partial}{\partial \vec{w}} |\vec{w}|$$

$$\nabla w_\alpha = \frac{\partial}{\partial \vec{w}} |w_\alpha| = \frac{\partial}{\partial w_\alpha} |w_\alpha| = \begin{cases} 1, & w_\alpha > 0 \\ -1, & w_\alpha < 0 \\ \text{not defined}, & w_\alpha = 0 \end{cases}, \quad \alpha = 1, \dots, d$$

Since in practice the chance that $w_\alpha = 0$ is very rare, we can define $\nabla \vec{w}$:

$$\nabla w_\alpha = \begin{cases} 1, & w_\alpha \geq 0 \\ -1, & w_\alpha \leq 0 \end{cases}, \quad \alpha = 1, 2, \dots d$$

Therefore,

$$\frac{\partial}{\partial \vec{w}} L(\vec{w}) = \sum_{i=1}^{n} 2\vec{x}_i (\vec{x}_i^T \vec{w} - y_i) + \lambda \nabla \vec{w},$$

where $\nabla \vec{w}$: $\nabla w_\alpha = \begin{cases} 1, & w_\alpha \geq 0 \\ -1, & w_\alpha < 0 \end{cases}, \quad \alpha = 1, \dots, d$

The matrix form is:

$$L(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda |\vec{w}|$$

$$= \vec{w}^T X^T X \vec{w} - 2\vec{w}^T X^T Y + Y^T Y + \lambda |\vec{w}|$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = 2X^T X \vec{w} - 2X^T Y + \lambda\, \text{sign}(\vec{w})$$

CSE517A

Machine Learning (SP18)

Written Homework 1.

Chunyuan Li, Jianxi Xing

Problem 1.

(c) $$\frac{\partial}{\partial \vec{w}} L(\vec{w}) = \sum_{i=1}^{n} \frac{1}{1+e^{-y_i \vec{w}^T \vec{x}_i}} \cdot e^{-y_i \vec{w}^T \vec{x}_i} \cdot (-y_i \vec{x}_i)$$

$$= -\sum_{i=1}^{n} \frac{e^{-y_i \vec{w}^T x_i}}{1+e^{-y_i \vec{w}^T \vec{x}_i}} y_i \vec{x}_i$$

$$= -\sum_{i=1}^{n} \frac{y_i \vec{x}_i}{1+e^{y_i \vec{w}^T \vec{x}_i}}$$

The matrix form is:

$$\frac{\partial}{\partial \vec{w}} L(\vec{w}) = \frac{Y^T}{[1+\exp(diag(Y)\cdot X \cdot w)]^T} \cdot X$$

where X is a $n \times d$ matrix $\begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \end{bmatrix}$

and $\div$ is elementwise division.

d) $$\because L(\vec{w}) = C \sum_{i=1}^{n} \max\{1-y_i \vec{w}^T \vec{x}_i, 0\} + \|\vec{w}\|_2^2$$

$$= C\sum_{i=1}^{n} \begin{cases} 1-y_i \vec{w}^T \vec{x}_i, & y_i \vec{w}^T \vec{x}_i < 1 \\ 0, & y_i \vec{w}^T \vec{x}_i \geq 1 \end{cases} + \vec{w}^T \vec{w}$$

$$\therefore \frac{\partial}{\partial \vec{w}} L(\vec{w}) = C\sum_{i=1}^{n} \begin{cases} -y_i \vec{x}_i, & y_i \vec{w}^T \vec{x}_i < 1 \\ 0, & y_i \vec{w}^T \vec{x}_i \geq 1 \end{cases} + 2\vec{w}$$

The matrix form is:

$$L(\vec{w}) = C \sum_{i=1}^{n} \max\{1-y_i \vec{w}^T \vec{x}_i, 0\} + \|\vec{w}\|_2^2$$

$$= C \cdot \frac{sign(1-\vec{w}^T X^T Y)+1}{2} \cdot (1-\vec{w}^T X^T Y) + \vec{w}^T \vec{w}$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = C\, Y\, diag\left(\frac{sign(1-\vec{w}^T X^T Y)+1}{2}\right) X + 2\vec{w}$$

Problem 2.

(a) $\because Pr(y=1|x) = \dfrac{e^{\vec{w}^T\vec{x}}}{1+e^{\vec{w}^T\vec{x}}} = Sigm(\vec{w}^T\vec{x})$

$Pr(y=0|x) = \dfrac{1}{1+e^{\vec{w}^T\vec{x}}} = 1 - Sigm(\vec{w}^T\vec{x})$

$\therefore Pr(y|x) = [Sigm(\vec{w}^T\vec{x})]^{y} \cdot [1-Sigm(\vec{w}^T\vec{x})]^{1-y}$

$\therefore L(\vec{w}) = -log[Pr(Y|X)]$

$= -log[\prod_{i=1}^{n} Pr(y_i|x_i)]$

$= -\sum_{i=1}^{n} log[Pr(y_i|x_i)]$

$= -\sum_{i=1}^{n} [log[sigm(\vec{w}^T\vec{x}_i)]^{y_i} + log[1-sigm(\vec{w}^T\vec{x}_i)]^{1-y_i}]$

$= -\sum_{i=1}^{n} [y_i log[sigm(\vec{w}^T\vec{x}_i)] + (1-y_i) log[1-sigm(\vec{w}^T\vec{x}_i)]]$

(b) $\dfrac{\partial}{\partial\vec{w}} L = -\sum_{i=1}^{n} [y_i \cdot \dfrac{1}{sigm(\vec{w}^T\vec{x}_i)} \cdot sigm(\vec{w}^T\vec{x}_i)[1-sigm(\vec{w}^T\vec{x}_i)] \cdot \vec{x}_i$

$- (1-y_i) \dfrac{1}{1-sigm(\vec{w}^T\vec{x}_i)} sigm(\vec{w}^T\vec{x}_i)[1-sigm(\vec{w}^T\vec{x}_i)] \cdot \vec{x}_i]$

$= -\sum_{i=1}^{n} [y_i \vec{x}_i [1-sigm(\vec{w}^T\vec{x}_i)] - (1-y_i)\vec{x}_i \, sigm(\vec{w}^T\vec{x}_i)]$

$= -\sum_{i=1}^{n} [y_i \vec{x}_i - \vec{x}_i \, sigm(\vec{w}^T\vec{x}_i)]$

$= -\sum_{i=1}^{n} \vec{x}_i [y_i - sigm(\vec{w}^T\vec{x}_i)]$

(c) By definition,

$H = \dfrac{\partial^2}{\partial\vec{w}^2} L(\vec{w})$

$= \dfrac{\partial}{\partial\vec{w}}\left(\dfrac{\partial}{\partial\vec{w}} L(\vec{w})\right)$

$= \dfrac{\partial}{\partial\vec{w}} (-1) \sum_{i=1}^{n} \vec{x}_i [y_i - sigm(\vec{w}^T\vec{x}_i)]$

$= \sum_{i=1}^{n} \vec{x}_i \dfrac{\partial}{\partial\vec{w}} sigm(\vec{w}^T\vec{x}_i)$

$= \sum_{i=1}^{n} \vec{x}_i \cdot sigm(\vec{w}^T\vec{x}_i)[1-sigm(\vec{w}^T\vec{x}_i)] \vec{x}_i^T$

$= \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T sigm(\vec{w}^T\vec{x}_i)[1-sigm(\vec{w}^T\vec{x}_i)]$

$X^T W X = \sum_{i=1}^{n} \vec{x}_i W_{ii} \vec{x}_i^T$

$= \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T sigm(\vec{w}^T\vec{x}_i)[1-sigm(\vec{w}^T\vec{x}_i)]$

$= H$

Problem 2 (continue)

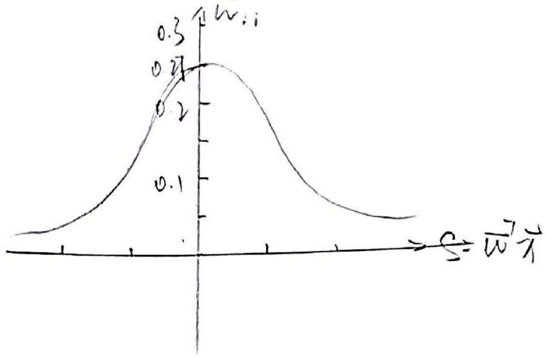
(c) (Cont.)

Note sigm(·) as $\theta(\cdot)$, $\vec{w}^T\vec{x}_i$ as s,

then: $W_{ii} = \theta(s)[1-\theta(s)]$

$$= \frac{e^s}{1+e^s}\cdot\frac{1}{1+e^s}$$

$$= \frac{e^s}{(1+e^s)^2}$$



From the plot we can see that W_{ii} gets larger when $|\vec{w}^T x_i| \to 0$, and gets smaller when $|\vec{w}^T\vec{x}_i| \to \infty$

Since $P(y_i=1 \mid x_i) = \frac{e^{\vec{w}^T\vec{x}_i}}{1+e^{\vec{w}^T\vec{x}_i}}$

$$P(y_i=0 \mid \vec{x}_i) = \frac{1}{1+e^{\vec{w}^T\vec{x}_i}},$$

when $\vec{w}^T\vec{x}_i = 0$, $P(y_i=1 \mid \vec{x}_i) = P(y_i=0 \mid \vec{x}_i) = \frac{1}{2}$

which means $\vec{x}_i$ is "ambigous" or "hard to classify". In other words, W_{ii} is larger for examples harder to be classified and smaller for examples easier to be classified.

(d) $\forall \vec{\lambda} \in \mathbb{R}^d$, $\vec{\lambda}^T H \vec{\lambda}$

$$\vec{\lambda}^T H \vec{\lambda} = \vec{\lambda}^T \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \,\mathrm{sigm}(\vec{w}^T\vec{x}_i)[1-\mathrm{sigm}(\vec{w}^T\vec{x}_i)]\vec{\lambda}$$

$$= \sum_{i=1}^{n} \vec{\lambda}^T\vec{x}_i\vec{x}_i^T\vec{\lambda}\cdot \mathrm{sigm}(\vec{w}^T\vec{x}_i)[1-\mathrm{sigm}(\vec{w}^T\vec{x}_i)]$$

$$= \sum_{i=1}^{n} \|\vec{\lambda}^T\vec{x}_i\|_2^2 \cdot \mathrm{sigm}(\vec{w}^T\vec{x}_i)[1-\mathrm{sigm}(\vec{w}^T\vec{x}_i)]$$

$\because \forall s \in \mathbb{R}$, $\mathrm{sigm}(s) \in (0,1)$, $[1-\mathrm{sigm}(s)] \in (0,1)$

$\therefore \vec{\lambda}^T H \vec{\lambda} \geq \sum_{i=1}^{n} 0 = 0$

$\therefore$ H is positive semi-definite

(e) By definition of Newton's method:

$$\vec{w}_{new} = \vec{w} - \frac{\frac{\partial}{\partial\vec{w}}L(\vec{w})}{H}$$

$$= \vec{w} + \frac{\sum_{i=1}^{n}\vec{x}_i[y_i - \mathrm{sigm}(\vec{w}^T\vec{x}_i)]}{H}$$

$$= \frac{\sum_{i=1}^{n}\vec{x}_i[y_i - \mathrm{sigm}(\vec{w}^T\vec{x}_i)] + X^T W X\cdot\vec{w}}{X^T W X}$$

$$= \frac{\sum_{i=1}^{n}\vec{x}_i[y_i - \mathrm{sigm}(\vec{w}^T\vec{x}_i)] + \sum_{i=1}^{n}\vec{x}_i\cdot\vec{x}^T W_{ii}\cdot\vec{w}}{X^T W X}$$

$$= \frac{\sum_{i=1}^{n}\vec{x}_i[y_i - \mathrm{sigm}(\vec{w}^T\vec{x}_i) + W_{ii}\vec{x}_i^T\vec{w}]}{X^T W X}$$

$$= \frac{\sum_{i=1}^{n}\vec{x}_i W_{ii}[\frac{1}{W_{ii}}(y_i - \mathrm{sigm}(\vec{w}^T\vec{x}_i)) + \vec{x}_i^T\vec{w}]}{X^T W X}$$

$$= \frac{\sum_{i=1}^{n}\vec{x}_i W_{ii}\vec{z}}{X^T W X}$$

$$= \frac{X W\vec{z}}{X^T W X}$$

$$= (X^T W X)^{-1} X W\vec{z}$$

Problem 3

(a) $L(\vec{w}) = (\vec{w}^T X - Y) P (\vec{w}^T X - Y)^T + \lambda \vec{w}^T \vec{w}$

(b) Note $L_1(\vec{w})$:

$$L_1(\vec{w}) = (\vec{w}^T X - Y) P (\vec{w}^T X - Y)^T$$

$$= (\vec{w}^T X P - Y P)(X^T \vec{w} - Y^T)$$

$$= \vec{w}^T X X^T \vec{w} - \vec{w}^T X P Y^T - Y P X^T \vec{w} + Y P Y^T$$

$$= \vec{w}^T X X^T \vec{w} - 2\vec{w}^T X P Y^T + Y P Y^T$$

$$\frac{\partial}{\partial \vec{w}} L_1(\vec{w}) = 2 X P X^T \vec{w} - 2 X P Y^T$$

$$\therefore \frac{\partial}{\partial \vec{w}} L(\vec{w}) = \frac{\partial}{\partial \vec{w}} L_1(\vec{w}) + 2\lambda \vec{w}$$

$$= 2 X P X^T \vec{w} - 2 X P Y^T + 2\lambda \vec{w}$$

$$= 2(X P X^T + \lambda I)\vec{w} - 2 X P Y^T = 0$$

$$\Rightarrow \vec{w} = (X P X^T + \lambda I)^{-1} X P Y^T$$

(c) Comparing the equation above and that in Problem 2(e), we can see that using Newton's method to solve Logistic ~~R~~ regression is equal to solve a weighted ridge regression at each iteration, where the weight p_i is given by W_{ii} and examples that are harder to be classified will get larger weights.