CSE517A

Machine Learning (SP18)

Written Homework 1.

Chunyuan Li, Jianxi Xing

Problem 1.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Problem 2.

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

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

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

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

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

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

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

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

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

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

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

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

(c) By definition,

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

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

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

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

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

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

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

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

$$= H$$

Problem 2 (continue)

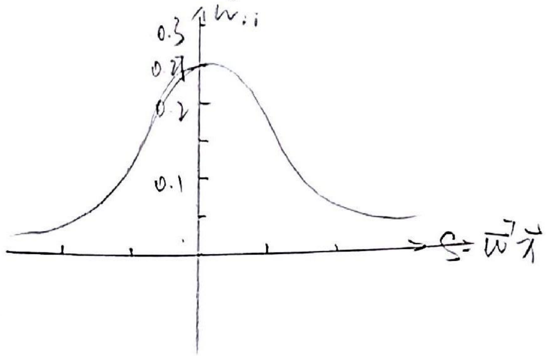
(c) (Cont.)

Note $sigm(\cdot)$ 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 $P(y_i=1 \mid x_i) = \dfrac{e^{\vec{w}^T\vec{x}_i}}{1+e^{\vec{w}^T\vec{x}_i}}$

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

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

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) $\because \forall s \in R$, $sigm(s) \in (0,1)$, $1-sigm(s) \in (0,1)$

$\forall \vec{v} \in R^d$, $\vec{v}^T\vec{v} = \sum_{i=1}^{d} v_i^2 \geq 0$

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

$$\geq \sum_{i=1}^{n} \cdot 0 = 0$$

i.e. H is a nonnegative scalar

$\therefore \forall \vec{\lambda} \in R^d$, $\vec{\lambda}^T H \vec{\lambda} = H \cdot \sum_{i=1}^{d} \lambda_i^2 \geq 0$

$\therefore H$ is positive semi-definite

(e) By definition of Newton's method:

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

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

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

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

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

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

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

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

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

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.