CSE 517A Homework 1

1. (a) $XX^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix}$

$w^TXX^Tw = (w_1, w_2)\begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$

$= \begin{pmatrix} w_1\sum_{i=1}^n x_{1i}^2 + w_2\sum_{i=1}^n x_{1i}x_{2i} & w_1\sum_{i=1}^n x_{1i}x_{2i} + w_2\sum_{i=1}^n x_{2i}^2 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$

$= w_1^2\sum_{i=1}^n x_{1i}^2 + w_1w_2\sum_{i=1}^n x_{1i}x_{2i} + w_1w_2\sum_{i=1}^n x_{1i}x_{2i} + w_2^2\sum_{i=1}^n x_{2i}^2$

$= w_1^2\left(\sum_{i=1}^n x_{1i}^2\right) + 2w_1w_2\left(\sum_{i=1}^n x_{1i}x_{2i}\right) + w_2^2\left(\sum_{i=1}^n x_{2i}^2\right)$

(b) $(X^Tw)^T = \left(\begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right)^T = \begin{pmatrix} w_1x_{11} + w_2x_{21} \\ \vdots \\ w_1x_{1n} + w_2x_{2n} \end{pmatrix}^T = \begin{pmatrix} w_1x_{11} + w_2x_{21} & \cdots & w_1x_{1n} + w_2x_{2n} \end{pmatrix}$

$w^TX = (w_1, w_2)\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix} = \begin{pmatrix} w_1x_{11} + w_2x_{21} & \cdots & w_1x_{1n} + w_2x_{2n} \end{pmatrix}$

Thus, $(X^Tw)^T = w^TX$

(c) $\sum_{i=1}^n x_iy_i = \sum_{i=1}^n \begin{pmatrix} x_{1i}y_i \\ x_{2i}y_i \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \end{pmatrix}$

$Xy = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \end{pmatrix}$

Thus, $\sum_{i=1}^n x_iy_i = Xy$

(d) $\sum_{i=1}^n x_ix_i^Tw = \sum_{i=1}^n \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}\begin{pmatrix} x_{1i} & x_{2i} \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \sum_{i=1}^n \begin{pmatrix} x_{1i}^2 & x_{1i}x_{2i} \\ x_{1i}x_{2i} & x_{2i}^2 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$

$= \sum_{i=1}^n \begin{pmatrix} w_1x_{1i}^2 + w_2x_{1i}x_{2i} \\ w_1x_{1i}x_{2i} + w_2x_{2i}^2 \end{pmatrix} = \begin{pmatrix} w_1\sum_{i=1}^n x_{1i}^2 + w_2\sum_{i=1}^n x_{1i}x_{2i} \\ w_1\sum_{i=1}^n x_{1i}x_{2i} + w_2\sum_{i=1}^n x_{2i}^2 \end{pmatrix}$

$XX^Tw = \begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} w_1\sum_{i=1}^n x_{1i}^2 + w_2\sum_{i=1}^n x_{1i}x_{2i} \\ w_1\sum_{i=1}^n x_{1i}x_{2i} + w_2\sum_{i=1}^n x_{2i}^2 \end{pmatrix}$

Thus, $\sum_{i=1}^n x_ix_i^Tw = XX^Tw$

(e) $L(w) = \sum_{i=1}^n (w^Tx_i - y_i)^2 = \sum_{i=1}^n \left((w^Tx_i)^2 - 2w^Tx_iy_i + y_i^2\right)$

$= \sum_{i=1}^n w^Tx_ix_i^Tw - \sum_{i=1}^n 2w^Tx_iy_i + \sum_{i=1}^n y_i^2$

$= w^T\left(\sum_{i=1}^n x_ix_i^Tw\right) - 2w^T\left(\sum_{i=1}^n x_iy_i\right) + \sum_{i=1}^n y_i^2$

$= w^TXX^Tw - 2w^TXy + y^Ty$

$= (X^Tw - y)^T(X^Tw - y)$

2. (a) $L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$

$= (X^T w - y)^T (X^T w - y) + \lambda w^T w$

$\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w}\left(w^T X X^T w - 2w^T X y + y^T y + \lambda w^T w\right)$

$= 2XX^T w - 2Xy + 2\lambda w = 2X(X^T w - y) + 2\lambda w$

$w_{t+1} = w_t - c\frac{\partial L(w_t)}{\partial w_t} = w_t - c\left(2X(X^T w_t - y) + 2\lambda w_t\right)$

(b) $L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1$

$= w^T X X^T w - 2w^T X y + y^T y + \lambda \|w\|_1$

$\frac{\partial \|w\|_1}{\partial w} = sign(w) = \begin{cases} +1, & w > 0 \\ -1, & w < 0 \\ \text{any value in } [-1, +1], & w = 0 \end{cases}$

$\frac{\partial L(w)}{\partial w} = 2XX^T w - 2Xy + sign(w) = 2X(X^T w - y) + sign(w)$

$w_{t+1} = w_t - c\left(2X(X^T w_t - y) + sign(w_t)\right)$

(c) $L(w) = \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T x_i))$

$\frac{\partial}{\partial w} \log(1 + \exp(-y_i w^T x_i)) = \frac{1}{1 + \exp(-y_i w^T x_i)} \cdot \exp(-y_i w^T x_i) \cdot (-y_i x_i) = \frac{-y_i x_i}{\exp(y_i w^T x_i) + 1}$

$\frac{\partial L(w)}{\partial w} = \sum_{i=1}^{n} \frac{\partial}{\partial w} \log\left(1 + \exp(-y_i w^T x_i)\right)$

$= \sum_{i=1}^{n} \frac{-y_i x_i}{\exp(y_i w^T x_i) + 1}$

$w_{t+1} = w_t - c \sum_{i=1}^{n} \frac{-y_i x_i}{\exp(y_i w_t^T x_i) + 1}$

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

1) when $1 - y_i w^T x_i > 0$, $\max\{1 - y_i w^T x_i, 0\} = 1 - y_i w^T x_i$, $\frac{\partial}{\partial w}(1 - y_i w^T x_i) = -y_i x_i$

2) when $1 - y_i w^T x_i < 0$, $\max\{1 - y_i w^T x_i, 0\} = 0$, $\frac{\partial}{\partial w}(0) = 0$

Thus, $\frac{\partial}{\partial w} \sum_{i=1}^{n} \max\{1 - y_i w^T x_i, 0\} = \sum_{i=1}^{n} 1(1 - y_i w^T x_i > 0) \cdot (-y_i x_i)$

$\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w}\left(C \sum_{i=1}^{n} \max\{1 - y_i w^T x_i, 0\} + \|w\|_2^2\right)$

$= C \sum_{i=1}^{n} \frac{\partial}{\partial w} \max\{1 - y_i w^T x_i, 0\} + 2w$

$= C \sum_{i=1}^{n} 1(1 - y_i w^T x_i > 0)(-y_i x_i) + 2w$

$w_{t+1} = w_t - c\left[C \sum_{i=1}^{n} 1(1 - y_i w_t^T x_i > 0)(-y_i x_i) + 2w_t\right]$

3. (a) In logistic regression with $y_i \in \{0, 1\}$, we model

$$P(y_i = 1 \mid x_i) = sigm(w^T x_i) \qquad P(y_i = 0 \mid x_i) = 1 - sigm(w^T x_i)$$

Then, we have $P(y_i \mid x_i) = (sigm(w^T x_i))^{y_i} (1 - sigm(w^T x_i))^{1-y_i}$

Given a dataset $\{(x_i, y_i)\}_{i=1}^n$, then we have

$$\log P(\{y_i\} \mid \{x_i\}) = \sum_{i=1}^n \log \left[(sigm(w^T x_i))^{y_i} (1 - sigm(w^T x_i))^{1-y_i} \right]$$

$$= \sum_{i=1}^n \left(y_i \log(sigm(w^T x_i)) + (1 - y_i) \log(1 - sigm(w^T x_i)) \right)$$

Thus, $L(w) = -\sum_{i=1}^n \left(y_i \log(sigm(w^T x_i)) + (1 - y_i) \log(1 - sigm(w^T x_i)) \right)$

(b) $$\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w} \left[-\sum_{i=1}^n \left(y_i \log(sigm(w^T x_i)) + (1 - y_i) \log(1 - sigm(w^T x_i)) \right) \right]$$

$$= -\sum_{i=1}^n \frac{\partial}{\partial w} \left(y_i \log(sigm(w^T x_i)) + \frac{\partial}{\partial w} (1 - y_i) \log(1 - sigm(w^T x_i)) \right)$$

$$\frac{\partial}{\partial w} (y_i \log(sigm(w^T x_i))) = y_i \frac{1}{sigm(w^T x_i)} \cdot \frac{\partial\, sigm(w^T x_i)}{\partial w}$$

$$= y_i \cdot \frac{1}{sigm(w^T x_i)} \; sigm(w^T x_i)(1 - sigm(w^T x_i)) \cdot x_i$$

$$= y_i (1 - sigm(w^T x_i)) \cdot x_i$$

$$\frac{\partial}{\partial w} (1 - y_i) \log(1 - sigm(w^T x_i)) = -(1 - y_i) \frac{1}{1 - sigm(w^T x_i)} \cdot \frac{\partial\, sigm(w^T x_i)}{\partial w}$$

$$= -(1 - y_i) \frac{1}{1 - sigm(w^T x_i)} \; sigm(w^T x_i)(1 - sigm(w^T x_i)) \cdot x_i$$

$$= -(1 - y_i)\, sigm(w^T x_i) \cdot x_i$$

Thus, $$\frac{\partial}{\partial w} L(w) = -\sum_{i=1}^n y_i (1 - sigm(w^T x_i)) \cdot x_i - (1 - y_i) \cdot sigm(w^T x_i) \cdot x_i$$

$$= -\sum_{i=1}^n y_i x_i - y_i\, sigm(w^T x_i)\, x_i + y_i\, sigm(w^T x_i)\, x_i - sigm(w^T x_i) \cdot x_i$$

$$= -\sum_{i=1}^n (y_i - sigm(w^T x_i))\, x_i$$

(c) $$H = \frac{\partial^2}{\partial w^2} L(w) = \frac{\partial}{\partial w} \sum_{i=1}^n (sign(w^T x_i) - y_i) x_i$$

$$= \sum_{i=1}^n x_i\, sigm(w^T x_i)(1 - sigm(w^T x_i))\, x_i^T$$

$$= \sum_{i=1}^n x_i\, W_{ii}\, x_i^T = X W X^T$$

Since $W_{ii} = sign(w^T x_i)(1 - sign(w^T x_i))$, when $sign(w^T x_i) = \frac{1}{2}$, i.e. $w^T x_i = 0$, W_{ii} achieves the largest value $\frac{1}{4}$

when $sign(w^T x_i) = 0$ or 1 i.e. $w^T x_i \to -\infty$ or $+\infty$, W_{ii} achieves the smallest value 0

(d) $\forall v \in R^d$, $v^T H v = v^T X^T W X v = (Xv)^T W (Xv)$

since W is diagonal matrix and all diagonal entries $W_{ii} = sigm(w^T x_i)(1 - sigm(w^T x_i))$

which are non-negative, it follows that $(Xv)^T W (Xv) \geq 0$

Thus, $v^T H v \geq 0$ which means H is positive semi-definite

(e) Newton's method $w_{new} = w - H^{-1} \frac{\partial L}{\partial w}$

$$= w + (X^T W X)^{-1} \sum_{i=1}^{n} (y_i - sigm(w^T x_i)) x_i$$

since $z_i = x_i^T w + \frac{1}{w_{ii}} (y_i - sign(w^T x_i))$, then we have $(y_i - sign(w^T x_i)) = w_{ii}(z_i - x_i^T w)$

Thus, $w_{new} = w + (X^T W X)^{-1} \sum_{i=1}^{n} w_{ii} (z_i - x_i^T w) x_i$

$$= w + (X^T W X)^{-1} \left[\sum_{i=1}^{n} x_i^T w_{ii} z_i - \sum_{i=1}^{n} x_i^T w_{ii} x_i w \right]$$

$$= w + (X^T W X)^{-1} (X W z - X^T W X w)$$

$$= w + (X^T W X)^{-1} X W z - w = (X^T W X)^{-1} X W z$$

4 (a) $L(w) = \sum_{i=1}^{n} p_i (w^T x_i - y_i)^2 + \lambda w^T w$

$= (X^T w - y)^T P (X^T w - y) + \lambda w^T w$

(b) $\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w}\left[(X^T w - y)^T P (X^T w - y) + \lambda w^T w\right]$

$= \frac{\partial}{\partial w}\left((w^T X - y^T) P (X^T w - y)\right) + 2\lambda w$

$= \frac{\partial}{\partial w}\left[w^T X P X^T w - w^T X P y - y^T P X^T w + y^T P y\right] + 2\lambda w$

$= \frac{\partial}{\partial w}\left[w^T X P X^T w - 2 w^T X P y + y^T P y\right] + 2\lambda w$

$= X P X^T w + (X P X^T)^T w - 2 X P y + 2\lambda w$

$= 2 X P X^T w - 2 X P y + 2\lambda w$

$= (2 X P X^T + 2\lambda I) w - 2 X P y = 0$

$\Rightarrow \quad w^* = (X P X^T + \lambda I)^{-1} X P y$

(c) Let $\lambda = 0$, $w^* = (X P X^T)^{-1} X P y$

In problem 3(e), $w_{new} = (X W X^T)^{-1} X W z$

Newton's Method can be shown to solve each iteration by fitting a new weighted least squares problem, where the weights get updated at every iteration. Because one is reweighting the least squares objective at each step, that procedure is known as iteratively reweighted least squares.

5. (a) Using $y \in \{-1, +1\}$ is a good choice for implementation project 1 because it simplifies the mathematics, particularly in gradient calculations. And this label choice directly reflects the directionality of classification, enabling the model to more clearly learn the boundary

(b) For regression tasks, the 'true' regression loss is often considered to be the Mean Squared Error ($MSE = \frac{1}{n}\sum_{n=1}^{n}(y_i - \hat{y}_i)^2$) This is because it directly measures the average squared difference between the predicted and true values. MSE is widely used due to its mathematical properties, interpretability and its strong connection to statistical models. Other loss functions, like Absolute-loss and Huber-loss may serve as surrogates to handle specific scenarios.