# Assignment 1

Question 2

1. Show that ridge regression can be rewritten as a non regularized linear regression

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2n} \|Xw + b\mathbf{1} - y\|_2^2 + \lambda \|w\|_2^2$$

$$= \frac{1}{2n} \left( \|Xw + b\mathbf{1} - y\|^2 + (\sqrt{2n\lambda})^2 \|w\|_2^2 \right)$$

$$= \frac{1}{2n} \left( \underbrace{(Xw + b\mathbf{1} - y)^T (Xw + b\mathbf{1} - y)}_{\text{dot product}} + \underbrace{\sqrt{2n\lambda}\, w^T (\sqrt{2n\lambda}\, w)}_{\text{dot product}} \right)$$

This is similar to a dot product

$$= \frac{1}{2n} \left( \begin{bmatrix} Xw + b\mathbf{1} - y \\ \sqrt{2n\lambda}\, I_d w \end{bmatrix}^T \begin{bmatrix} Xw + b\mathbf{1} - y \\ \sqrt{2n\lambda}\, I_d w \end{bmatrix} \right)$$

$$= \frac{1}{2n} \left\| \begin{bmatrix} Xw + b\mathbf{1} - y \\ \sqrt{2n\lambda}\, \cancel{I}\, w + 0_d \end{bmatrix} \right\|_2^2 \quad \text{since } w \text{ is of size } d$$

$$= \frac{1}{2n} \left\| \begin{bmatrix} Xw + b\mathbf{1} \\ \sqrt{2n\lambda}\, w \end{bmatrix} - \begin{bmatrix} y \\ 0_d \end{bmatrix} \right\|_2^2$$

$Iw = w$ since $I$ is identity matrix of size $d$

$$= \frac{1}{2n} \left\| \begin{bmatrix} Xw + b\mathbf{1}_n \\ \sqrt{2n\lambda}\, I_d w \end{bmatrix} - \begin{bmatrix} y \\ 0_d \end{bmatrix} \right\|_2^2$$

$\mathbf{1} \to \mathbf{1}_n$ because $y \in \mathbb{R}^n$

$$= \frac{1}{2n} \left\| \begin{bmatrix} X & \mathbf{1}_n \\ \sqrt{2n\lambda}\, I_d & 0_d \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} - \begin{bmatrix} y \\ 0_d \end{bmatrix} \right\|_2^2$$

$\therefore$ Proved.

2. $$\frac{\partial}{\partial w} \frac{1}{2n} \|Xw + b1 - y\|_2^2 + \lambda \|w\|_2^2$$

1st term $$= \frac{1}{2n}\left((Xw + b1 - y)^T (Xw + b1 - y)\right)$$

$$= \frac{1}{2n} (w^T X^T + 1^T b - y^T)(Xw + b1 - y)$$

$$= \frac{1}{2n} (w^T X^T X w + 1^T b X w - y^T X w + w^T X^T b1 + b1^T b$$
$$- y^T b1 - w^T X^T y - 1^T b y + y^T y$$

Now derive wrt $w$

$$= \frac{1}{2n}\left(\frac{\partial}{\partial w}\left(w^T X^T X w + b^T X w - y^T X w + w^T X^T b - w^T X^T y\right)\right)$$

$$(b^T X w)^T = w^T X^T b \qquad (y^T X w)^T = w^T X^T y$$

$$= \frac{1}{2n}\left(\frac{\partial}{\partial w}\left(w^T X^T X w + 2 w^T X^T b - 2 w^T X^T y\right)\right)$$

$$= \frac{1}{2n}\left(\frac{\partial}{\partial w}\left(w^T X^T X w + 2(X^T b)^T w - 2(X^T y)^T w\right)\right)$$

$$= \frac{1}{2n}\left(\frac{\partial}{\partial w}(w^T X^T X w) + 2X^T b - 2X^T y\right)$$

$\cancel{\frac{\partial}{\partial w}(w^T X^T X w) = w^T ((X^T X)^T + X^T X)}$

Use product rule of matrix partial derivative

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

$$= \frac{1}{n} X^T (Xw + b - y)$$

2nd term $\Rightarrow \frac{\partial}{\partial w} \lambda \|w\|_2^2 = 2\lambda w$

Combining terms

$$= \frac{1}{n} X^T (Xw + b\mathbf{1} - y) + 2\lambda w$$

$\therefore$ Proved

$$\frac{\partial}{\partial b} \frac{1}{2n} \|Xw + b\mathbf{1} - y\|_2^2 + \lambda \|w\|_2^2$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} \|Xw + b\mathbf{1} - y\|_2^2$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} (Xw + b\mathbf{1} - y)^T (Xw + b\mathbf{1} - y)$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} (w^T X^T + b^T - y^T)(Xw + b - y)$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} (w^T X^T Xw + b^T Xw - y^T Xw + w^T X^T b + b^T b - y^T b - w^T X^T y - b^T y + y^T y)$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} (2b^T Xw + b^T b - 2y^T b)$$

$$\cancel{= \frac{1}{2n} \frac{\partial}{\partial b} (2(Xw)^T b + 1^T b^2 - 2y b^T)}$$

$$= \frac{1}{2n} \frac{\partial}{\partial b} (2b\mathbf{1}^T Xw + \mathbf{1}^T b^2 - 2b\mathbf{1}^T y)$$

$$= \frac{1}{n} \mathbf{1}^T (Xw + b - y)$$

$\therefore$ Proved

Question 2

3. Algorithm submitted
4. Algorithm submitted

5. Comparing run time of 2 approaches

For $\lambda = 10$

Runtime of closed form approach = 0.0003

Runtime of gradient descent approach = 0.442

$\lambda = 0$

Closed form Training error = 0.062297

~~Tr~~ Training loss = 0.062297

Test error = 0.529986

$\lambda = 10$

Closed form Training error = 0.41589

Training loss = 0.45481

Test error = 0.44393

$\lambda = 0$

Gradient Descent Training error = 0.062595

Training loss = 0.062595

Test error = 0.50952

$\lambda = 10$

Gradient Descent Training error = 0.41642

~~Tr~~ Training loss = 0.45481

Testing error = 0.44419

I believe $\lambda = 10$, closed form is a better approach because it not only has a faster runtime but also a smaller testing error

$\therefore$ For this dataset, $\lambda = 10$ closed form approa[ch] is better.