HW5

1. AdaBoost

a)
$$\sum_{i=1}^{n} \exp\left[-y_i\left(G(x_i) + \alpha F(x_i)\right)\right]$$

$$= \sum_{i=1}^{n}\left[\exp\left[-y_i G(x_i)\right] \times \exp\left[-y_i \alpha F(x_i)\right]\right]$$

$$= \sum_{i=1}^{n} w_i \exp\left(-\alpha y_i F(x_i)\right)$$

since $y_i, F(x_i) \in \{-1, +1\}$ , $\alpha > 0$

$$\sum_{i=1}^{n} w_i \exp\left(-\alpha y_i F(x_i)\right)$$

$$= \sum_{y_i = F(x_i)} \left(w_i e^{-\alpha}\right) + \sum_{y_i \neq F(x_i)} \left(w_i e^{\alpha}\right)$$

$$= e^{-\alpha} \sum_{y_i = F(x_i)} w_i + e^{\alpha} \sum_{y_i \neq F(x_i)} w_i$$

$$= \left(e^{\alpha} - e^{-\alpha}\right) \sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq F(x_i)\right) + e^{-\alpha} \sum_{i=1}^{n} w_i$$

since $e^{\alpha} - e^{-\alpha} > 0$ for $\alpha > 0$ and $e^{-\alpha} \sum_{i=1}^{n} w_i$ is independent of $F$

thus $$\alpha_m, G_m = \arg\min_{\alpha, F \in \mathcal{F}} \left[\left(e^{\alpha} - e^{-\alpha}\right) \sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq F(x_i)\right) + e^{-\alpha} \sum_{i=1}^{n} w_i\right] \quad ①$$

reduced to $$G_m = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq F(x_i)\right)$$

---

b) Plug in optimal $G_m$ to ①, we have

$$\alpha_m = \arg\min_{\alpha} \left[\left(e^{\alpha} - e^{-\alpha}\right) \sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq G_m(x_i)\right) + e^{-\alpha} \sum_{i=1}^{n} w_i\right]$$

$$= \arg\min_{\alpha} L(\alpha)$$

set derivative of Loss fuction with respect to $\alpha$ to be 0

$$\frac{\partial L(\alpha)}{\partial \alpha} = \left(e^{\alpha} - e^{-\alpha}(-1)\right) \sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq G_m(x_i)\right) - e^{-\alpha} \sum_{i=1}^{n} w_i = 0$$

$$\Rightarrow \quad e^{2\alpha} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i \mathbb{1}\left(y_i \neq G_m(x_i)\right)} - 1 = \frac{1}{err_m} - 1$$

$$\Rightarrow \quad \alpha_m = \frac{1}{2} \log\left(\frac{1 - err_m}{err_m}\right)$$

c)

$$w_i = w_i \cdot \exp(-\alpha_m y_i G_m(x_i)) \qquad (d)$$

Since $y_i, G_m(x_i) \in \{-1, +1\}$

$$w_i = \exp(-\alpha_m)\, w_i \times \exp\left[2\alpha_m \mathbb{1}[y_i \neq G_m(x_i)]\right]$$

we can ignore the scalar $\exp(-\alpha_m)$

then $w_i = w_i \times \exp\left[2\alpha_m \mathbb{1}[y_i \neq G_m(x_i)]\right]$

plug in $\alpha_m = \frac{1}{2}\log\left(\frac{1-err_m}{err_m}\right) \qquad (c)$

then $w_i = w_i \times \exp\left[\log\left(\frac{1-err_m}{err_m}\right) \mathbb{1}(y_i \neq G_m(x_i))\right]$

Now this form is the same as P339

where $\alpha_m = \log\left(\frac{1-err_m}{err_m}\right)$

$$w_i = w_i \times \exp\left[\alpha_m \mathbb{1}(y_i \neq G_m(x_i))\right]$$

---

2. Regularization

Part a)

take gradient of Ridge Regression

$$\frac{\partial\left(\frac{1}{2n}\|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2\right)}{\partial\theta} \qquad \theta \in R^P$$

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

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

$$= \frac{1}{2n}\frac{\partial(y-X\theta)^T(y-X\theta)}{\partial(y-X\theta)}\frac{\partial(y-X\theta)}{\partial\theta} + \lambda\theta$$

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

set gradient to be 0

$$\frac{1}{2n}X^T(X\theta - y) + \lambda\theta = 0$$

$$(X^TX + 2n\lambda)\theta = X^Ty$$

$$\hat{\theta} = (X^TX + 2n\lambda I_p)^{-1}X^Ty$$

where $I_p$ is the pxp identity matrix

Part b) $\hat{\theta}^{(\lambda)} = \frac{\hat{\theta}^{(ols)}}{1+2n\lambda} = \frac{y}{1+2n\lambda}$ where $\hat{\theta}_j^{(\lambda)} = \frac{\hat{\theta}_j^{(OLS)}}{1+2n\lambda} = \frac{y_j}{1+2n\lambda}$

Proof:

If $X = I_p$ be the $p \times p$ identity matrix

Then $X^TX = I_p \quad X^Ty = y$

plug in $\hat{\theta}$ computed from part c)

$$\hat{\theta}^{(\lambda)} = (X^TX + 2n\lambda I_p)^{-1} X^Ty = (I_p + 2n\lambda I_p)^{-1} y = \frac{1}{1+2n\lambda} I_p^{-1} y = \frac{1}{1+2n\lambda} y$$

$\hat{\theta}_j^{(\lambda)} = \frac{\hat{\theta}_j^{(OLS)}}{1+2n\lambda}$ where $\hat{\theta}_j^{(OLS)} = y_j$

---

Part (c) Gradient Descent:

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta \left(\frac{1}{2n} \|X\theta - y\|^2 + \lambda \|\theta\|_1\right) \quad ①$$

Lasso Regression:

$$\frac{1}{2n} \|X\theta - y\|^2 + \lambda \|\theta\|_1$$

$$= \frac{1}{2n} \sum_{i=1}^{p} (x_i^T\theta - y_i)^2 + \lambda \sum_{i=1}^{p} |\theta_{(i)}|$$

The j th coordinate of gradient with respect to $\theta$ is

$$\frac{\partial \frac{1}{2n} \sum_{i=1}^{n} (x_i^T\theta - y_i)^2 + \lambda \sum_{i=1}^{p} |\theta_{(i)}|}{\partial \theta_{(j)}} = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} (x_i^T\theta - y_i) x_i^T(j) + \text{sign}(\theta_{(j)})\lambda & \text{if } \theta_{(j)} \neq 0 \\ \frac{1}{n} \sum_{i=1}^{n} (x_i^T\theta - y_i) x_i^T(j) & \text{if } \theta_{(j)} = 0 \end{cases}$$

Thus, the gradient is

$$\nabla_\theta = \frac{1}{n} X^T(X\theta - y) + \lambda\gamma \quad \text{where } \gamma_i \in \begin{cases} \{\text{sign}(\theta_i)\} & \text{if } \theta_i \neq 0 \\ 0 & \text{if } \theta_i = 0 \end{cases} \quad (\text{for } i = 1, \cdots, p)$$

plug the gradient in ①

then $\theta_{k+1} = \theta_k - \eta_k \left(\frac{1}{n} X^T(X\theta_k - y) + \lambda\gamma\right)$

$$\boxed{\begin{array}{l} \theta_{k+1} = \left(1 - \frac{\eta_k}{n} X^TX\right)\theta_k + \eta_k\left(\frac{1}{n} X^Ty - \lambda\gamma\right) \\ \text{where } \gamma_i \in \begin{cases} \{\text{sign}(\theta_{k(i)}) & \text{if } \theta_{k(i)} \neq 0 \\ 0 & \text{if } \theta_{k(i)} = 0 \end{cases} \quad (\text{for } i = 1, \cdots, p) \end{array}}$$

Part d) $\hat{\theta}_j^{L} = \text{sign}(\hat{\theta}_j^{OLS})(|\hat{\theta}_j^{OLS}| - n\lambda)_+$

$$= \begin{cases} \hat{\theta}_j^{OLS} - n\lambda & \text{if } \hat{\theta}_j^{OLS} > n\lambda \\ 0 & \text{if } |\hat{\theta}_j^{OLS}| \leq n\lambda \\ \hat{\theta}_j^{OLS} + n\lambda & \text{if } \hat{\theta}_j^{OLS} < -n\lambda \end{cases}$$ where $\hat{\theta}_j^{OLS} = y_j$

---

Proof

In Part c) we compute the j the coordinate of gradient with respect to $\theta$

$$\nabla_{\theta_j} = \begin{cases} \frac{1}{n}[X^T(X\theta - y)]_{(j)} + \text{sign}(\theta_j)\lambda & \text{if } \theta_j \neq 0 \\ \frac{1}{n}[X^T(X\theta - y)]_{(j)} & \text{if } \theta_j = 0 \end{cases}$$

If X is pxp identity matrix $I_p$

Then $X^TX = I_p$ $X^Ty = y$

Thus $$\nabla_{\theta_j} = \begin{cases} \frac{1}{n}(\theta_j - y_j) + \text{sign}(\theta_j)\lambda & \text{if } \theta_j \neq 0 \\ -\frac{1}{n}y_j & \text{if } \theta_j = 0 \end{cases}$$

set $\nabla_{\theta_j}$ to be 0 to compute $\hat{\theta}_j^{Lasso}$

$$\hat{\theta}_j^{L} = \begin{cases} \hat{\theta}_j^{OLS} - n\,\text{sign}(\hat{\theta}_j^{L})\lambda & \text{if } \hat{\theta}_j^{L} \neq 0 \\ 0 & \text{if } \hat{\theta}_j^{L} = 0 \end{cases}$$ where $\hat{\theta}_j^{OLS} = y_j$

① if $\hat{\theta}_j^{L} \neq 0 \Rightarrow |\hat{\theta}_j^{OLS}| > n\lambda$

i) if $\text{sign}(\hat{\theta}_j^{L}) = +1$

then $\hat{\theta}_j^{L} = \hat{\theta}_j^{OLS} - n\lambda > 0 \Rightarrow \hat{\theta}_j^{OLS} > n\lambda$

ii) if $\text{sign}(\hat{\theta}_j^{L}) = -1$

then $\hat{\theta}_j^{L} = \hat{\theta}_j^{OLS} + n\lambda < 0 \Rightarrow \hat{\theta}_j^{OLS} < -n\lambda$

② if $\hat{\theta}_j^{L} = 0 \Rightarrow |\hat{\theta}_j|^{OLS} \leq n\lambda$