CS480 Assignment 1

Nikhil Nanda
20727218

Exercise 1:

1. python3 alg1.py
2. If there exists w^* and b^* such that for all i,

$$\begin{cases} \langle x_i, w^* \rangle + b \geq 0, & \text{if } y_i = 1 \\ \langle x_i, w^* \rangle + b < 0, & \text{if } y_i = -1 \end{cases}$$

This means that the data points (x_i, y_i) are linearly separable and one of the separating hyperplane is

$$x_1 w_1 + x_2 w_2 + \cdots + x_d w_d + b = 0.$$

The data points with $\langle x_i, w^* \rangle + b = 0$ lie on this separating hyperplane.

Since the data is linearly separable, there exists infinitely many hyperplane that separate data points with $y_i = 1$ and $y_i = -1$.

This means that we can get a different hyperplane by adding a very small constant c.

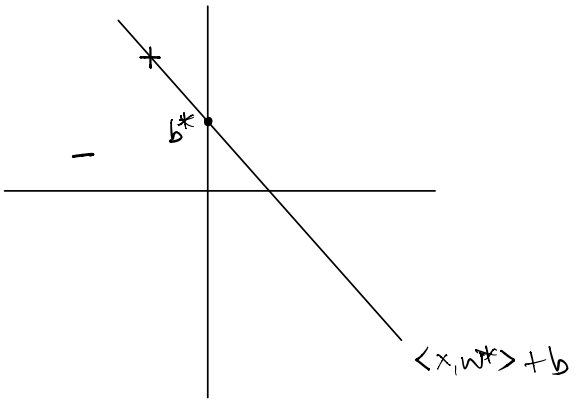
$$w_{new} = w^*$$
$$b_{new} = b + c \text{ , where } c \text{ is a very small constant}$$

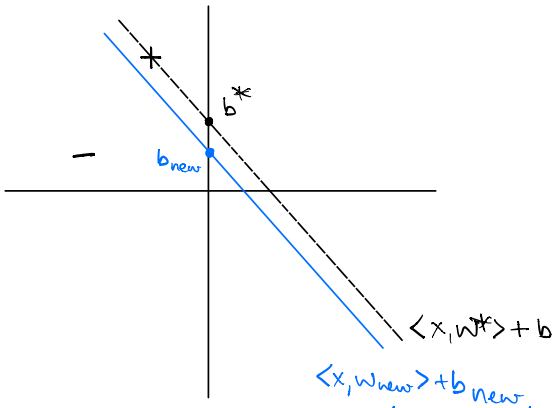
The only difference is that now, after adding c to the offset b, there exists no data point on the new separating hyperplane.

Now, for all i

$$\begin{cases} \langle x_i, w_{new} \rangle + b_{new} > 0 \text{ , if } y_i = 1 \\ \langle x_i, w_{new} \rangle + b_{new} < 0 \text{ , if } y_i = -1 \end{cases}$$

Note: The sign of constant c is determined in such a way that the new hyperplane is a separating hyperplane with no data points lying on the hyperplane





$\langle x, w_{new} \rangle + b_{new}$, constant c is positive in this case

Notation $\begin{cases} +, \text{if } y_i = 1 \\ -, \text{if } y_i = -1 \end{cases}$

Exercise 2:

1. $\text{sign}(w) \cdot \max\{0, |w| - \lambda\} =$

$$\underset{z \in \mathbb{R}}{\text{argmin}} \ \frac{1}{2}(z-w)^2 + \lambda|z|$$

The optimality condition is

$$0 \in \nabla\left(\frac{1}{2}(z-w)^2 + \lambda|z|\right)$$

$$\Rightarrow \quad 0 \in (z-w) + \lambda \nabla|z|$$

When $z \neq 0$:

$$\nabla|z| = \text{sign}(z)$$

For the optimum z,

$$0 = z - w + \lambda \,\text{sign}(z)$$

$$\Rightarrow \quad z = w - \lambda \,\text{sign}(z)$$

Thus, if $z < 0 \Rightarrow w < -\lambda$

and if $z > 0 \Rightarrow w > \lambda$

$\therefore \quad \text{sign}(z) = \text{sign}(w)$

Substituting this in previous equation, we get

$$z = w - \lambda \, \text{sign}(w)$$

When $z = 0$,

Then $\Delta |z|$ is in the interval $[-1, 1]$

and then the optimality condition is,

$$0 \in -w + \lambda [-1, 1]$$

$$\Rightarrow w \in [-\lambda, \lambda]$$

$$\Rightarrow |w| \leq \lambda$$

Combining the above cases for value of z_1

$$z = \begin{cases} 0 & , \text{ if } |w| \leq \lambda \\ w - \lambda \, \text{sign}(w) & , \text{ if } |w| > \lambda \end{cases}$$

$\Updownarrow$ Same as

$$z = \text{sign}(w) \cdot \max\{0, |w| - \lambda\}$$

→ Graph is in Figure_2.1.png
code to plot it → python3 alg2.1.py

2.

$$w_j \leftarrow \underset{z \in \mathbb{R}}{\operatorname{argmin}} \; \frac{1}{2} \left\| (X_{j:})^T z + \sum_{k \neq j} (X_{k:})^T w_k - y \right\|_2^2 + \lambda |z|$$

On simplifying and manipulating the right hand side by using the property that: x is a global minimizer of f iff x is a global minimizer of $\lambda f + c$ for any $\lambda > 0$ and $c \in \mathbb{R}$.

$$w_j \leftarrow \underset{z \in \mathbb{R}}{\operatorname{argmin}} \; \frac{1}{2} \left(z - \frac{X_{j:}(y - X_k^T w_k)}{X_{j:} X_{j:}^T} \right)^2 + \frac{\lambda}{X_{j:} X_{j:}^T} |z| ,$$

where X_k = matrix X excluding the the j-th row $(X_{j:})$

w_k = matrix w excluding the j-th entry (w_j)

From the equation proved in part 1, we can conclude that

$$w_j = z = \text{sign}\left(\frac{X_{j:}(y - X_k^T w_k)}{X_{j:} X_{j:}^T}\right) \cdot \max\left\{0, \left|\frac{X_{j:}(y - X_k^T w_k)}{X_{j:} X_{j:}^T}\right| - \frac{\lambda}{X_{j:} X_{j:}^T}\right\}$$

Thus, step 3 of algorithm 2 can be performed in $O(n)$ time and space.

$X_{j:} \in \mathbb{R}^{1 \times n}$

$X_k \in \mathbb{R}^{d-1 \times n}$

$w_k \in \mathbb{R}^{d-1}$

$y \in \mathbb{R}^n$

→ python3 alg2.2.py

(Currently lambda is set to 1, it can be changed by changing the variable lam in the code)

3. When $\lambda \to \infty$, according to the above equation w_j will be set to zero.

Thus, by the end of the first for loop, $w = 0$ and it has converged.

As λ increases, the percentage of non-zero weights in w decreases, making the algorithm converge faster.

When $\lambda \to 0$, there is no regularization. The weights are dense (high percentage of non-zero weights) and the algorithms takes time to converge as it is computationally heavy due to the high percentage of non-zero weights.

Exercise 3:

1. Implemented along with part 3 $\Rightarrow$ python3 alg3.py

2. $X \in \mathbb{R}^{d \times n}$, $D \in \mathbb{R}_{+}^{n \times n}$

$D(i,j) \leftarrow \|X_{:i} - X_{:j}\|_2 \Leftrightarrow$ euclidean distance between i-th and j-th points

In order to replace the double-for loop in Algorithm 3 with matrix-matrix or matrix-vector products, we can use the fact that

$$\|a-b\|_2^2 = a^T a - 2a^T b + b^T b \quad -(1)$$

Without Loss of generality,

Let $X = [a \quad b \quad c]$, where a, b, c are 3 column vectors

$x \in \mathbb{R}^{d \times n}$, where $n=3$

$a, b, c \in \mathbb{R}^d$

Let $M = X^T X$

Thus $M \in \mathbb{R}^{n \times n}$,

$$M = \begin{bmatrix} a^T a & a^T b & a^T c \\ b^T a & b^T b & b^T c \\ c^T a & c^T b & c^T c \end{bmatrix}$$

All the terms needed to calculate the euclidean distance are on the diagonal of matrix M.

Thus, we need to extract these terms into an nxn matrix in order to calculate the euclidean distance by using equation (1).

With the help of column matrix of only 1's, we can extract the diagonal terms into an nxn matrix in the following way:

$$D(D) = \text{diag}(M) \cdot 1^T - 2 \cdot M + 1 \cdot (\text{diag}(M))^T,$$

where $\text{diag}(M) \in \mathbb{R}^n$

$1 \in \mathbb{R}^n$

and $(\text{diag}(M) \cdot 1^T)^T = 1 \cdot (\text{diag}(M))^T$

To get D, we can use the fact that

$$D = \sqrt{D \odot D}$$

3. It can be seen from Figure_3.png that the implementation without the double - for loop performs better in terms of running time

Exercise 4:

1. Need to compare $\sum_{i=1}^{n} p_i^*$ with n_1, give w^* and b^* at optimum,

which maximizes the likelihood function which is the same as minimizing the negative log likelihood function:

$$-\sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i),$$

where $p_i = p(x_i) = \frac{1}{1+e^{-(w^T x+b)}}$

$$\Updownarrow$$

$$-\sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-(w^T x+b)}}\right) + (1-y_i)\log\left(\frac{e^{-(w^T x+b)}}{1+e^{-(w^T x+b)}}\right)$$

On differentiating above equation w.r.t b,

$$-\sum_{i=1}^{n} y_i \left(1+e^{-(w^T x+b)}\right)\left(1+e^{-(w^T x+b)}\right)^2 \cdot \left(e^{-(w^T x+b)}\right)$$

$$+ (1-y_i) \cdot \frac{\left(1+e^{-(w^T x+b)}\right)}{e^{-(w^T x+b)}} \cdot \frac{\left[-e^{-(w^T x+b)} \cdot \left[1+e^{-(w^T x+b)}\right] + \left(e^{-(w^T x+b)}\right)^2\right]}{\left(1+e^{-(w^T x+b)}\right)^2}$$

$$= -\sum_{i=1}^{n} y_i \cdot \frac{e^{-(w^T x+b)}}{1+e^{-(w^T x+b)}}$$

$$+ (1-y_i) \cdot \frac{\left(1+e^{-(w^T x+b)}\right)}{e^{-(w^T x+b)}} \cdot \frac{\left[-e^{-(w^T x+b)}\right]}{\left[1+e^{-(w^T x+b)}\right]^2}$$

$$= -\sum_{i=1}^{n} y_i \cdot \frac{e^{-(w^T x+b)}}{1+e^{-(w^T x+b)}}$$

$$-\frac{(1-y_i)}{\left(1+e^{-(w^T x+b)}\right)}$$

$$= -\sum_{i=1}^{n} \frac{y_i\left(1+e^{-(w^T x+b)}\right)-1}{1+e^{-(w^T x+b)}}$$

$$= -\sum_{i=1}^{n} y_i \frac{\left(1+e^{-(w^T x+b)}\right)}{1+e^{-(w^T x+b)}} - \frac{1}{1+e^{-(w^T x+b)}}$$

On setting the derivative to zero, to get the optimal value

$$\sum_{i=1}^{n} \frac{y_i \left(1 + e^{-(w^{*T} x + b^*)}\right)^{-}}{1 + e^{-(w^{*T} x + b^*)}} - \frac{1}{1 + e^{-(w^{*T} x + b^*)}} = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - p_i^*\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i^*$$

$$\Rightarrow \boxed{n_1 = \sum_{i=1}^{n} p_i^*}$$, by definition of n_1

2. Need to prove $\text{sign}(w^*) = \text{sign}(\bar{x}_1 - \bar{x}_0)$

On differentiating negative log likelihood function w.r.t w,

$$-\sum_{i=1}^{n} y_i \cdot \left(1+e^{-(wx_i+b)}\right)\cdot\left(1+e^{-(wx_i+b)}\right)^{-2} \cdot e^{-(wx_i+b)} \cdot x_i$$

$$+(1-y_i)\cdot \frac{1}{1-\frac{1}{1+e^{-(wx_i+b)}}} \cdot \left[\left(1+e^{-(wx_i+b)}\right)^{-2} \cdot e^{-(wx_i+b)} \cdot x\right]$$

$$= -\sum_{i=1}^{n} y_i \cdot \frac{x_i e^{-(wx_i+b)}}{1+e^{-(wx_i+b)}}$$

$$-(1-y_i)\frac{\left(1+e^{-(wx_i+b)}\right)}{e^{-(wx_i+b)}} \cdot \left(1+e^{-(wx_i+b)}\right)^{-2} \cdot x_i \cdot e^{-(wx_i+b)}$$

$$= -\sum_{i=1}^{n} y_i \cdot \frac{x e^{-(wx_i+b)}}{1+e^{-(wx_i+b)}}$$

$$-\frac{(1-y_i)x_i}{1+e^{-(wx_i+b)}}$$

$$= -\sum_{i=1}^{n} \frac{x_i\left[y_i e^{-(wx_i+b)} - 1 + y_i\right]}{1+e^{-(wx_i+b)}}$$

$$= -\sum_{i=1}^{n} x_i \frac{\left[y_i\left(1+e^{-(wx_i+b)}\right) - 1\right]}{1+e^{-(wx_i+b)}}$$

On setting the derivative to zero, to get the optimal value

$$-\sum_{i=1}^{n} x_i \frac{\left[y_i\left(1+e^{-(w^* x_i+b^*)}\right) - 1\right]}{1+e^{-(w^* x_i+b^*)}} = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} x_i\left(y_i - p_i^*\right) = 0}$$

Thus, we can say that

$$\frac{\bar{x}}{n_1}\sum_{i=1}^{n}(y_i - p_i) = 0$$

$$\Rightarrow \frac{1}{n_1}\sum_{i=1}^{n}\left(y_i\bar{x} - p_i\bar{x}\right) = 0$$

$$\frac{1}{n_1}\sum_{i=1}^{n} x_i(y_i - p_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(\frac{1}{n_1}\cdot x_i y_i - \frac{1}{n_1} x_i p_i\right) = 0$$

$$\Rightarrow \bar{x}_1 - \frac{1}{n_1}\sum_{i=1}^{n} x_i p_i = 0$$

Equating above two equations

$$\bar{x}_1 - \frac{1}{n_1}\sum_{i=1}^{n} x_i p_i = \frac{1}{n_1}\sum_{i=1}^{n} (y_i \bar{x} - p_i \bar{x})$$

$$\Rightarrow \bar{x}_1 - \frac{1}{n_1}\sum_{i=1}^{n} y_i \bar{x} = \frac{1}{n_1}\sum_{i=1}^{n} x_i p_i - \frac{1}{n_1}\sum_{i=1}^{n} p_i \bar{x}$$

$$\Rightarrow \bar{x}_1 - \bar{x} = \frac{1}{n_1}\sum_{i=1}^{n} p_i (x_i - \bar{x})$$

$$\Rightarrow n_1 (\bar{x}_1 - \bar{x}) = \sum_{i=1}^{n} p_i (x_i - \bar{x})$$

On looking at two cases where $\text{sign}(w^*) > 0$ and $\text{sign}(w^*) < 0$

Since the sigmoid function is monotonic, it can be seen that

when $\text{sign}(w^*) > 0 \Rightarrow \sum_{i=1}^{n} p_i (x_i - \bar{x}) > 0$

and $\text{sign}(w^*) < 0 \Rightarrow \sum_{i=1}^{n} p_i (x_i - \bar{x}) < 0$

Thus, we can say that,

$$\text{sign}(w^*) = \text{sign}\left(\sum_{i=1}^{n} p_i (x_i - \bar{x})\right)$$

$$= \text{sign}\left(n_1 (\bar{x}_1 - \bar{x})\right)$$

$$= \text{sign}(\bar{x}_1 - \bar{x}) \quad \text{, since } n_1 \text{ is positive}$$

$$= \text{sign}(\bar{x}_1 - \bar{x}_0)$$

Thus, $\boxed{\text{sign}(w^*) = \text{sign}(\bar{x}_1 - \bar{x}_0)}$

Exercise 5 :

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^{n} w_j^t}, \quad i = 1, 2, \ldots, n \qquad (1)$$

$$\epsilon_t = \epsilon_t(h_t) = \sum_{i=1}^{n} p_i^t \cdot |h_t(x_i) - y_i| \qquad (2)$$

$$\beta_t = \frac{\epsilon_t}{1 - \epsilon_t} \qquad (3)$$

$$w_i^{t+1} = w_i^t \, \beta_t^{1 - |h_t(x_i) - y_i|} \qquad (4)$$

Since, $y_i \in \{0, 1\}$

and Assuming $h_t(x_i) \in \{0, 1\}$

$$\therefore |h_t(x_i) - y_i| = 0 \text{ OR } 1$$

$$\epsilon_{t+1}(h_t) = \sum_{i=1}^{n} p_i^{t+1} \cdot |h_t(x_i) - y_i| \qquad (5)$$

$$\Rightarrow \epsilon_{t+1}(h_t) = \sum_{h_t(x_i) = y_i} p_i^{t+1} \cdot 0 + \sum_{h_t(x_i) \neq y_i} p_i^{t+1} \cdot 1$$

$$\Rightarrow \boxed{\epsilon_{t+1}(h_t) = \sum_{h_t(x_i) \neq y_i}^{n} p_i^{t+1}} \quad \text{--- (6)}$$

From (1),

$$p_i^{t+1} = \frac{w_i^{t+1}}{\sum_{j=1}^{n} w_j^{t+1}}$$

From (4),

$$\Rightarrow p_i^{t+1} = \frac{w_i^t \, \beta_t^{1 - |h_t(x_i) - y_i|}}{\sum_{j=1}^{n} w_j^t \cdot \beta_t^{1 - |h_t(x_i) - y_i|}}$$

$$\Rightarrow p_i^{t+1} = \frac{w_i^t \, \beta_t^{1 - |h_t(x_i) - y_i|}}{\sum_{h_t(x_j) = y_j}^{n} w_j^t \beta_t + \sum_{h_t(x_j) \neq y_j}^{n} w_j^t}$$

$$\Rightarrow \sum_{h_t(x_i) \neq y_i}^{n} p_i^{t+1} = \frac{\sum_{h_t(x_i) \neq y_i}^{n} w_i^t \, \beta_t^{1 - |h_t(x_i) - y_i|}}{\sum_{h_t(x_j) = y_j}^{n} w_j^t \beta_t + \sum_{h_t(x_j) \neq y_j}^{n} w_j^t}$$

$$\Rightarrow \sum_{h_t(x_j) \neq y_j}^{n} p_i^{t+1} = \frac{\sum_{h_t(x_i) \neq y_i}^{n} w_i^t}{\sum_{h_t(x_j)=y_j}^{n} w_j^t \beta_t + \sum_{h_t(x_j) \neq y_j}^{n} w_j^t} \quad (7)$$

From (2),

$$\epsilon_t = \sum_{h_t(x_i) \neq y_i}^{n} p_i^t = \frac{\sum_{h_t(x_i) \neq y_i}^{n} w_i^t}{\sum_{j=1}^{n} w_j^t}$$

$$\Rightarrow \sum_{h_t(x_i) \neq y_i}^{n} w_i^t = \epsilon_t \left[\sum_{j=1}^{n} w_j^t \right] \quad (8)$$

We know that,

$$1 = \sum_{i=1}^{n} p_i^t = \sum_{h_t(x_i)=y_i}^{n} p_i^t + \sum_{h_t(x_i) \neq y_i}^{n} p_i^t$$

$$\Rightarrow 1 = \sum_{h_t(x_i)=y_i} p_i^t + \epsilon_t$$

$$\Rightarrow \sum_{h_t(x_i)=y_i} p_i^t = 1 - \epsilon_t \quad (9)$$

$$\sum_{h_t(x_i)=y_i}^{n} p_i^t = \frac{\sum_{h_t(x_i)=y_i}^{n} w_i^t}{\sum_{j=1}^{n} w_j^t}$$

Thus from (9),

$$\sum_{h_t(x_i)=y_i} w_i^t = (1-\epsilon_t) \cdot \left[\sum_{j=1}^{n} w_j^t\right] \quad \text{—(10)}$$

From (6),

$$\epsilon_{t+1}(h_t) = \sum_{h_t(x_i)\neq y_i}^{n} p_i^{t+1}$$

$$= \frac{\sum_{h_t(x_i)\neq y_i}^{n} w_i^t}{\sum_{h_t(x_j)=y_j}^{n} w_j^t \beta_t + \sum_{h_t(x_j)\neq y_j}^{n} w_j^t}$$

$$= \frac{(\epsilon_t) \cdot \sum_{j=1}^{n} w_j^t}{\beta_t \cdot (1-\epsilon_t) \cdot \sum_{j=1}^{n} w_j^t + \epsilon_t \cdot \sum_{j=1}^{n} w_j^t}$$

$$\Rightarrow \quad \epsilon_{t+1}(h_t) = \frac{\epsilon_t}{\beta_t \cdot (1-\epsilon_t) + \epsilon_t}$$

From (3),

$$\epsilon_{t+1}(h_t) = \frac{\epsilon_t}{\frac{\epsilon_t}{1-\epsilon_t} \cdot (1-\epsilon_t) + \epsilon_t}$$

$$\Rightarrow \quad \boxed{\epsilon_{t+1}(h_t) = \frac{1}{2}}$$