Name : Han Nguyen . TXN200004

Assignment 1

Warm-up

Given $f(x)$, $f_1(x)$, $f_2(x)$ are convex

(1) Prove $g(x) = w \cdot f(x)$ is convex :

$$g(\lambda x + (1-\lambda)y) = w f(\lambda x + (1-\lambda)y)$$
$$\leq w[\lambda f(x) + (1-\lambda) f(y)]$$
$$= \lambda w f(x) + (1-\lambda) w f(y)$$
$$= \lambda g(x) + (1-\lambda) g(x)$$

$\Rightarrow g(x)$ is a convex function

(2). Prove $g(x) = f_1(x) + f_2(x)$ is convex

$$g(\lambda x + (1-\lambda)y) = f_1(\lambda x + (1-\lambda)y) + f_2(\lambda x + (1-\lambda)y)$$
$$\leq \lambda f_1(x) + (1-\lambda) f_1(y) + \lambda f_2(x) + (1-\lambda) f_2(y)$$
$$= \lambda [f_1(x) + f_2(x)] + (1-\lambda)[f_1(y) + f_2(y)]$$
$$\geq \lambda g(x) + (1-\lambda) g(x)$$

$\Rightarrow g(x)$ is a convex function

(3). $g(x) = \max(f_1(x), f_2(x))$ is convex

$$g(\lambda x + (1-\lambda)y) = \max[f_1(\lambda x + (1-\lambda)y), f_2(\lambda x + (1-\lambda)y)]$$
$$\leq \max[\lambda f_1(x) + (1-\lambda) f_1(y), \lambda f_2(x) + (1-\lambda) f_2(y)]$$
$$\leq \max[\lambda \max(f_1(x), f_2(x)) + (1-\lambda) \max(f_1(y), f_2(y)$$

$$, \lambda \max(f_2(x), f_1(x)) + (1-\lambda)\max(f_2(y), f_1(y))$$

$$= \max[\lambda g(x) + (1-\lambda) g(y), \lambda g(x) + (1-\lambda) g(y)]$$

$$= \lambda g(x) + (1-\lambda) g(y)$$

→ $g(x)$ is a convex function

(4). $g(x) = \exp(f(x))$ is convex

$$g(\lambda x + (1-\lambda)y) = \exp(f(\lambda x + (1-\lambda)y))$$

$$\leq \exp(\lambda f(x) + (1-\lambda) f(y)) \quad (1)$$

- Since $\log(x)$ is a concave

$$\log(\lambda a + (1-\lambda) b) \geq \lambda \log(a) + (1-\lambda)\log(b)$$

Let $a = \exp(f(x))$, $b = \exp(f(y))$

$$\Rightarrow \log(\lambda \cdot \exp(f(x)) + (1-\lambda)\cdot \exp(f(y)))$$

$$\geq \lambda f(x) + (1-\lambda) f(y) \quad (2)$$

- From (1) and (2)

$$\Rightarrow g(\lambda x + (1-\lambda)y) \leq \exp(\log(\lambda \cdot \exp(f(x)) + (1-\lambda)\cdot \exp(f(y))))$$

$$= \lambda \cdot \exp(f(x)) + (1-\lambda)\exp(f(y))$$

$$= \lambda g(x) + (1-\lambda) g(y)$$

$$\Rightarrow g(x) = \exp(f(x)) \text{ is a convex function}$$

Problem 1 :

① Prove $\sum_{m \in \{1 \dots M\}} | a^T x^{(m)} + b - y^{(m)} |$ is convex

$= | a^T x^{(1)} + b - y^{(1)} | + | a^T x^{(2)} + b - y^{(2)} | + \dots + | a^T x^{(M)} + b - y^{(M)} |$

Since $| a^T x^{(i)} + b - y^{(i)} |$ is convex then using ② from Warm-up we can argue that

$\sum_{m \in \{1 \dots M\}} | a^T x^{(m)} + b - y^{(m)} |$ is a convex function

② [Problem 1_2.py] → code file. Hand written logic below

- $a^T x^{(m)} + b - y^{(m)} < 0$

$$\rightarrow \begin{cases} \nabla a = -x^{(m)} \\ \frac{\partial}{\partial b} = -1 \end{cases}$$

- $a^T x^{(m)} + b - y^{(m)} > 0$

$$- \begin{cases} \nabla a = x^{(m)} \\ \frac{\partial}{\partial b} = 1 \end{cases}$$

- $a^T x^{(m)} + b - y^{(m)} = 0$

$$\Rightarrow \begin{cases} \nabla a = t\, x^{(m)} \text{ for } t \in [-1, 1] \\ \frac{\partial}{\partial b} = t \text{ for } t \in [-1, 1] \end{cases}$$

. Let's take $t = 0$

Then (*) $g_a = \sum_m \left(x^{(m)} \cdot \text{sign}\left(a^T x^{(m)} + b - y^{(m)}\right)\right)$

$\Rightarrow$ This mean total sum of products $x^{(m)}$ * sign at $x^{(m)}$

If we represent x as vector of all given values x

and s is the vector of corresponding sign $(xa + b - y)$

$\Rightarrow$ $g_a = X^T \cdot S$

(*) $g_b = \sum_m \left(\text{sign}\left(a^T x^{(m)} + b - y^{(m)}\right)\right)$

Same strategy applied

$\Rightarrow$ $g_b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}^T \cdot S$ (sum of all values in s)

- Update functions:

$$a^{(t+1)} = a^{(t)} - \gamma_t g_a = a^{(t)} - \gamma_t \cdot X^T \cdot S$$

$$b^{(t+1)} = b^{(t)} - \gamma_t g_b = b^{(t)} - \gamma_t \cdot S$$

- How to pick step size γ_t

Since absolute function is not convex smoothly we need to pick a diminishing step size

$\gamma_t = \frac{1}{1+t}$ could work but I want it to shrink a bit slower so I pick $\gamma_t = \frac{1}{1+\sqrt{t}}$

When to stop?

$g^{(t)} = \begin{bmatrix} g_a \\ g_b \end{bmatrix}$ I will stop when $\|g^{(t)}\|_2 \leq \varepsilon$

Pick $\varepsilon = 10^{-6}$

. Also, to not let it run forever I will set the max iteration to 10^5.

(3) Pros:

- Less punishing on outliner $|c| \leq c^2$

Cons:

- Not easy to solve since it's not differentiable
- Picking bad step size can result in the function not converge

Problem 2

1. Consider feature map $\phi(x_1, x_2) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix}$

a) Prove.

A circle in $\mathbb{R}^2$ space can be represented by the equation

$$(x_1 - a)^2 + (x_2 - b)^2 = r^2$$

$$\Leftrightarrow \quad x_1^2 + x_2^2 - 2ax_1 - 2bx_2 + a^2 + b^2 - r^2 = 0 \quad (1)$$

Consider our hypothesis with the given feature map

$$w^T \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix} + b = 0 \qquad w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$\Leftrightarrow \quad w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_3 x_2^2 + b = 0$$

$$\Leftrightarrow \quad w_3 (x_1^2 + x_2^2) + w_1 x_1 + w_2 x_2 + b = 0 \qquad (2)$$

From (1) and (2)

Let $\begin{cases} w_3 = 1 \\ w_2 = -2b \\ w_1 = -2a \end{cases} \Rightarrow (1) \Leftrightarrow (2)$

Conclusion: with feature map $\begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix}$ we can represent any circle in $\mathbb{R}^2$

b) No not all input in the feature space correspond to a circle since

If $w_3 = 0 \rightarrow w^T \phi(x_1, x_2) + b = 0$

$\Leftrightarrow \quad w_1 x_1 + w_2 x_2 + b = 0$

↳ This is a line

But went $w_3 \neq 0$

$\rightarrow w^T \phi(x_1, x_2) + b = 0$

$$\Leftrightarrow \left(x_1 + \frac{w_1}{2w_3}\right)^2 + \left(x_2 + \frac{w_2}{2w_3}\right)^2 = \frac{w_1^2 + w_2^2}{4 w_3^2} - \frac{b}{w_3}$$

⇒ This can only be a circle when

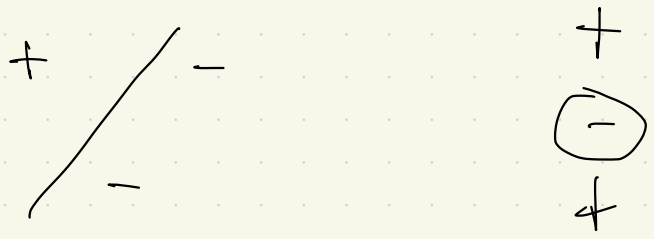
$$\frac{w_1^2 + w_2^2}{4 w_3^2} > \frac{b}{w_3}$$

(2). Consider 2 data points

→ this is easily separatable with a line

- Consider 3 data points

→ can surely be separatable by a line or a circle



- Consider 4 data points → this can fail to converge

in our feature map

→ Can't be separated by a line or circle

⇒ Minimum data points that can make the function not converge is 4

(3) $\frac{1}{M}\sum_{m=1}^{M} \max\{0, -y^{(m)}\cdot(w^T x^{(m)} + b)\}$

i) let $f_{w,b}(\underline{\Phi} x^{(m)}) = w^T \underline{\Phi} x^{(m)} + b$

$$\rightarrow \nabla_w = -\frac{1}{M}\sum_{m=1}^{M} y^{(m)}\left(\underline{\Phi}(x^{(m)})\right)\cdot 1_{-y^{(m)} f_{w,b}(\underline{\Phi}(x^{(m)}))\geq 0}$$

$1_{\rightarrow}$ = 1 if incorrectly
= 0 if correct

$$\nabla_b = -\frac{1}{M}\sum_{m=1}^{M} y^{(m)}\cdot 1_{-y^{(m)} f_{w,b}(\underline{\Phi}(x^{(m)}))}$$

Step size : $\gamma_t = 1$

$w_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ $b_0 = 0$

Update functions:

- $w^{(t+1)} = w^t + \frac{1}{M}\sum_{m=1}^{M} y^{(m)}\,\underline{\Phi}(x^{(m)})\cdot 1_{-y^{(m)} f_{w,b}(\underline{\Phi}(x^{(m)}))\geq 0}$

- $b^{(t+1)} = b^{t} + \frac{1}{M}\sum_{m=1}^{M} y^{(m)} \cdot 1_{-y^{(m)} \cdot f_{w,b}(\Phi(x^{(m)})) \geq 0}$

$\Rightarrow w^{(t+1)} = w^{t} + \nabla w$

$\rightarrow b^{(t+1)} = b^{(t)} + \nabla b$

[Problem 2.3i.py]

ii) For stochastic subgradient descent

Instead of computing $\frac{1}{M}\sum$ each iteration

We will pick an m row iteratively

Update function . i is the row we pick $\gamma_t = 1$

- $w^{(t+1)} = w^{t} + y^{(i)} \cdot \Phi(x^{(i)}) \cdot 1_{-y^{(i)} f_{w,b}(\Phi(x^{(i)})) \geq 0}$

- $b^{(t+1)} = b^{t} + y^{(i)} \cdot 1_{-y^{(i)} f_{w,b}(\Phi(x^{(i)})) \geq 0}$

[Problem 2.3ii.py]

iii) When using a fixed step size the rate of convergence for perceptron locs won't change as we change the stepsize.

[Problem 2.3iii.py]

3 step size $\begin{cases} 10^{-10} \\ 1 \\ 10^{10} \end{cases}$ -> converges at around 840^{th} iteration