Homework 1

Problem 1 Some may wonder why the Perceptron algorithm enjoys such appealing convergence properties. One high level answer lies in the empirical risk function that we are trying to optimize. In this homework problem, we will prove that the empirical risk associated with the hinge loss is convex. Recall that the empirical risk function is defined as:

$$\mathcal{L}(w) = \frac{1}{N}\sum_{i=1}^{N} \text{loss}(f(x_i), y_i; w),$$

where—in this problem and in class—we use the hinge loss and the linear classifier so that the loss function: $\text{loss}(f(x_i), y_i; w) = \max\{1 - y_i\langle w, x_i\rangle, 0\}$. The data point $y_i \in \{+1, -1\}$ and $x_i \in \mathbb{R}^d$. The unknown parameter $w \in \mathbb{R}^d$.

1. First prove that each individual loss function $\varphi_i(w) = \text{loss}(f(x_i), y_i; w)$ is a convex function in w. You can in fact prove that for any two convex functions $\phi(w), \psi(w)$, the new function $g(w) = \max\{\phi(w), \psi(w)\}$ is also convex (Note that the linear function $(1 - y_i\langle w, x_i\rangle)$ is always convex in w, regardless of y_i and x_i).
2. Then prove that the average of convex functions is also a convex function.

Problem 2 Suppose that we minimize the average squared loss for the linear classification problem discussed in class:

$$\mathcal{L}(w) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(f(x_i) - y_i)^2,$$

where $f(x) = \langle w, x\rangle$, $w \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, and $y \in \{-1, 1\}$.

How does it solve the linear classification problem? In particular, assume that we are in the under-parameterized regime such that $N \geq 2d$, and that the linear space spanned by $\{x_1 \ldots, x_N\}$ is of rank d, and answer the following questions.

1. What's the minimizer $\hat{w}$ of the average squared loss (empirical risk) given a training data set of $\{(x_1, y_1), \ldots, (x_N, y_N)\}$? What's the training error (average squared loss over the training data set) if we use this minimizer?

2. Can you either prove that this minimizer solves the linear classification problem or find an example that it does not solve the linear classification problem? Recall that in the linear classification problem, we are looking for $\hat{w}$ so that $y_i = \mathrm{sgn}(\langle \hat{w}, x_i \rangle)$.

Hints: Consider assembling $\{x_1 \ldots, x_N\}$ into a matrix $X \in \mathbb{R}^{d \times N}$. A solution $\hat{w}$ to the optimization of this $\mathcal{L}$ on $\mathbb{R}^d$ satisfies: $\nabla \mathcal{L}(\hat{w}) = 0$.

MACHINE LEARNING

ASSIGNMENT-01

1. First let's prove that linear function is convex.

Now, a function $f: X \to R$ is convex if it satisfies:

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$$

for all $x, y \in X$ and $t \in [-1, 1]$.

Now, in case of a linear function $\boxed{z(x) = cx}$, the inequality simply holds with equality:

$$z(tx + (1-t)y) = c(tx + (1-t)y) = ctx + c(1-t)y = tz(x) + (1-t)z(x)$$

for any $x, y \in X$ and any t

Hence Proved

Now, let's prove that constant function is convex

In case of a constant function $\boxed{z(x) = c}$, so putting it in the eqn

$$z(tx + (1-t)y) = c(t + (1-t))$$

Now, $tz(x) + (1-t)z(y) = tc + (1-t)c$

$$= \cancel{tc} + c - \cancel{tc}$$

$$= c$$

Hence Proved

(a) $y_i(\omega) = \text{loss}(f(x_i), y_i; \omega)$
$= \max\{1 - y_i\langle\omega, x_i\rangle, 0\}$

Given : $\phi_i(\omega) = 1 - y_i\langle\omega, x_i\rangle$

$\boxed{\text{Taking} \Rightarrow y_i(x_i) = 0}$

(i) To prove :- $\phi(w)$ is convex (linear)

Proof - $\boxed{\phi(\omega) \text{ is a linear function in } \omega}$ and linear functions are convex (given).

(ii) So now we have to prove :- $\psi_i(x)$ is convex (constant)

Proof :- $\boxed{\psi_i(x) \text{ is a constant function}}$ and constant functions are convex

(iii) Now, if $\phi(\omega)$ is convex and $\psi(\omega)$ is convex, then $g(\omega) = \max\{\phi(w), \psi(w)\}$ is convex

Proof :- $g(\omega_1) = \max\{\phi(\omega_1), \psi(\omega_1)\}$

$g(\omega_2) = \max\{\phi(\omega_2), \psi(\omega_2)\}$

$\omega_\lambda = \lambda\omega_1 + (1-\lambda)\omega_2$

$g(\omega_\lambda) = \max\{\phi(\omega_\lambda), \psi(\omega_\lambda)\}$

because ϕ is convex

By definition: $\phi(\omega_\lambda) \leq \lambda\phi(\omega_1) + (1-\lambda)\phi(\omega_2)$

Similarly ψ is convex

So again : $\psi(\omega_\lambda) \leq \lambda\psi(\omega_1) + (1-\lambda)\psi(\omega_2)$

$$g(\omega_\lambda) \le \max\{\lambda\phi(\omega_1) + (1-\lambda)\phi(\omega_2),\ \lambda(\psi(\omega_1)) + (1-\lambda)\psi(\omega_2)\}$$

Now, we know $\rightarrow$

$$\max\{a+b, c+d\} \le \max\{a,c\} + \max\{b,d\}$$

$$g(\omega_\lambda) \le \max\{\lambda\phi(\omega_1), \lambda\psi(\omega_1)\} + \max\{(1-\lambda)\phi(\omega_2), (1-\lambda)\psi(\omega_2)\}$$

$$= \lambda\max\{\phi(\omega_1), \psi(\omega_1)\} + (1-\lambda)\max\{\phi(\omega_2), \psi(\omega_2)\}$$

$$= \lambda g(\omega_1) + (1-\lambda)g(\omega_2)$$

$$\Rightarrow g(\omega_\lambda) \le \lambda g(\omega_1) + (1-\lambda)g(\omega_2) \quad \forall \lambda \in (0,1)$$

where $\omega_\lambda = \lambda\omega_1 + (1-\lambda)\omega_2$

$\Rightarrow$ $\boxed{\psi_i(\omega) = \text{loss}(f(x_i), y_i; \omega) \text{ is convex using above 2 proofs}}$

(b) To prove Average of 2 convex functions is also convex

Now $\text{avg}(f_1(x), f_2(x)) = \dfrac{f_1(x) + f_2(x)}{2}$

So, first let's prove: if $f_1(x)$ & $f_2(x)$ are convex functions then $(f_1+f_2)(x)$ is also a convex function

Proof :- $g(x) = (f_1 + f_2)(x)$

$g(x) = f_1(x) + f_2(x)$
$g(y) = f_1(y) + f_2(y)$

$z = \lambda x + (1-\lambda) y \quad - (1)$
$g(z) = f_1(z) + f_2(z)$

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

because both f_1 & f_2 are convex
$= \lambda (f_1(x) + f_2(x)) + (1-\lambda)(f_1(y) + f_2(y))$
$= \lambda g(x) + (1-\lambda) g(y)$

$\Rightarrow g(z) \leq \lambda g(x) + (1-\lambda) g(y) \cdot \quad \lambda \in (0,1)$
By (1) "z" $\Rightarrow \lambda(x) + (1-\lambda) y$

$\Rightarrow$ Sum of 2 convex function is convex by Mathematical Induction ;
Similarly $\rightarrow$ Sum of 'n' f(args) or convex func^n is convex

Now, let's prove if $f(x)$ is convex function, then $g(x) = \frac{f(x)}{c}$ is also convex for any constant 'c'

~~Proof - Simple Coming back~~

Proof - $g(x) = \frac{f(x)}{c}$ $\qquad g(z) = \frac{f(z)}{c}$

$g(y) = \frac{f(y)}{c}$ $\qquad \boxed{z = \lambda(x) + (1-\lambda) y}$

$$g(z) = \frac{1}{C}\left(\lambda f(x) + (1-\lambda) f(y)\right)$$

$$g(z) \leq \lambda g(x) + (1-\lambda) g(y)$$

$\Rightarrow g(z)$ is convex

Re - iterating our question :-

$$\psi_i = \text{loss}\,(f(x_i), y_i, w) = \max\{1 - y_i (w_i, x_i), 0\}$$

$\Rightarrow \psi_i$ is convex from above proof

Now, $\Sigma \psi_i$ is convex from above and

also $\frac{1}{N} \sum_{i=1}^{N} \psi_i$ is convex

Hence proved

2) (a) $\mathcal{L}(W) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(f(x_i) - y_i \right)^2 , \quad w \in \mathbb{R}^d$

$$f(x) = \langle w, x \rangle$$

$$= \frac{1}{N} \sum \frac{1}{2} (x_i^T w - y_i)^2$$

$$= \frac{1}{2N} \sum_{i=1}^{N} e_i^{1}(\omega) \quad \left(e = x_i^T w - x_i \right)$$

Now, writing in matrix form, $e = [e_1 e_2 e_3 \cdots]$

$$= \frac{1}{2N} e^T e$$

$$= \frac{1}{2N} (Y - XW)^T (Y - XW)$$

$$= \frac{1}{2N} (Y^T - W^T X^T)(Y - XW)$$

$$= \boxed{\frac{1}{2N} (Y^T Y - 2W^T X^T Y + W^T X^T X W)}$$

Taking gradient to find the minimum since its a convex function and equating it to 0, to find minima.

$$\nabla \mathcal{L}(\omega) = \frac{1}{2n} (\nabla Y^T Y - 2 \nabla \omega^T X^T Y + \omega^T X^T X W)$$

$$= \frac{1}{2N} (0 - 2X^T Y + 2X^T X W)$$

$$= X^T X W - X^T Y$$

Now, setting gradient to 0 to find optimal ω i.e $\hat{\omega}$,

$$\Rightarrow X^T X w - X^T Y = 0$$

$$\Rightarrow \boxed{\hat{w} = (X^T X)^{-1} X^T Y}$$

where $X = [x_1, x_2, \ldots]$
$Y = [y_1, y_2, \ldots]$

This is our minimizer $\hat{w}$

Training Error → Putting $\hat{w}$ in eqⁿ of Loss func

$$\Rightarrow \frac{1}{2N}\left(y^Ty - 2\left((X^TX)^{-1}X^Ty\right)^T X^Ty + ((X^TX)^{-1}X^Ty)^T X^T X ((X^TX)^{-1}X^Ty)\right)$$

(b) We will be proving that the minimizer doesn't solve the linear classification problem :-

Now here to prove the above it's enough to prove that $Y_i w^T x_i \geq 0$ which would mean that the point (x_i, y_i) is correctly classified.

Now, $w \Rightarrow (X^TX)^{-1}X^Ty$

Let $X \Rightarrow \begin{bmatrix} 840 & 840 \\ 840 & -840 \\ -840 & -840 \\ -840 & 840 \\ 3 & 7 \end{bmatrix}$ $Y = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$

$$X^TX = \begin{bmatrix} 2822409 & 21 \\ 21 & 2822449 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 3.5430\ldots e^{-7} & -2.6361\ldots e^{-12} \\ -2.6361\ldots e^{-12} & 3.5430\ldots e^{-07} \end{bmatrix}$$

$(X^T X)^{-1} X^T \Rightarrow$

$$\begin{bmatrix} 2.97\times e^{-4}, & 2.97\times e^{-4}, & -2.97\times e^{-4}, & -2.97\times e^{-4}, & 1.06\times e^{-6} \\ 2.97\times e^{-4}, & -2.97\times e^{-4}, & -2.97\times e^{-4}, & 2.97\times e^{-4}, & 2.48\times e^{-6} \end{bmatrix}$$

Now,

$$W = (X^T X)^{-1} X^T Y$$
$$= [1.189\times e^{-3}, \quad -2.488\times e^{-6}]$$

So, solving for $X = [3,7]$, $Y = -1$

$\Rightarrow \quad X W^T Y = -0.4767$ which is smaller than 0

Hence, we proved that it wasn't able to correctly classify $[3,7]$.