기계학습 Assignment #2. 2016-10764 이경형

problem 1.

(a) $E_{in}(h)$이 최소가 되는 곳에서, $\nabla E_{in}(h)=0$이다.

$\therefore \nabla E_{in}(h)=\sum_{n=1}^{N} 2(h-y_n)=0$, $\sum_{n=1}^{N} h=\sum_{n=1}^{N} y_n$, 양변을 N으로 나누면, $h_{mean}=\frac{1}{N}\sum_{n=1}^{N} y_n$ 이다.

(b) $E_{in}(h)$에 대해, $h>y_n$에서 $|h-y_n|=h-y_n$,

$h<y_n$에서 $=-h+y_n$이다. h에 대해 미분하면,

$\frac{dE_{in}(h)}{dh}=\begin{cases}1, & \text{for } h>y_n \\ -1, & \text{for } h<y_n\end{cases}$ 이고, $E_{in}'(h)=0$이기 위해서는 $h<y_n$과 $h>y_n$의 갯수가 같아야 한다.

$\therefore$ 또 $\frac{dE_{in}}{dh}=0$인 h는 h_{med}이다.

problem 2.

(a). $1-y_n w^T x_n>0$이면, $e_n(w)=1-y_n w^T x_n$.

$=1-\sum_i w_i - y_n \sum_i w_i x_{ni}$ 에서, 상수 y_n, x_{ni}에 대해 미분 가능하다. $1-y_n w^T x<0$에서 $e_n(w)=0$이므로 연속이고 미분 가능하다.

$1-y_n w^T x_n=0$일 때, $\nabla e_n(w)=-y_n x_n$에 대해, 미분 가능하기 위해 모든 방향에서의 미분 극한값이 같아야 하므로, $-y_n x_n=0$이어야 한다. 이 때, y_n은 $+1, -1$이므로, $x_n=0$이고, 이 경우 $y_n w^T x_n \neq 1$이다. 따라서 $y_n=w^T x_n$에서는 미분 가능하기 위한 조건이 모순되므로 이 점에서만 미분 불가능하다.

(b). $y_n=\text{sign}(w^T x_n)$이면, $[[\text{sign}(w^T x_n)\neq y_n]]=0$이고, $y_n w^T x_n>0$에 대해, $0\le e_n(w)=\max(0, 1-y_n w^T x_n)$이므로 이 때는 upper bound이다.

$y_n\neq \text{sign}(w^T x_n)$이면, $[[\text{sign}(w^T x_n)\neq y_n]]=1$이고, $y_n w^T x_n<0$에 대해, $1-y_n w^T x_n>1$이므로. 이 때의 $e_n(w)>1$에 대해 두 경우 모두 upper bound임을 알 수 있다.

$\therefore E_{in}(w)=\frac{1}{N}\sum_{n=1}^{N}[[\text{sign}(w^T x_n)\neq y_n]]\le \frac{1}{N}\sum_{n=1}^{N} e_n(w)$

로 in sample error의 upper bound임을 알 수 있다.

(모든 n에 대해 $[[\text{sign}(w^T x_n)\neq y_n]]\le e_n(w)$이므로).

(c) minimum $\frac{1}{N}\sum_{n=1}^{N} e_n(w)$를 찾기 위해 gradient descent method를 사용하자.

$\nabla e_n(w)=\begin{cases}0, & \text{for } y_n w^T x>1 \\ -y_n x_n, & \text{for } y_n w^T x<1\end{cases}$.

$\therefore \frac{1}{N}\sum_{n=1}^{N} e_n(w)$의 gradient는 위 식은 n에 대해 더하고 N으로 나눈 것과 같다. 이를 기반으로 SGD를 짜면 다음과 같다.

Random value $w(0)$에 대해,

for t in $(0, T)$:

$V_t=\nabla(\frac{1}{N}\sum_{n=1}^{N} e_n(w(t)))=\frac{1}{N}\sum_{n=1}^{N}\nabla e_n(w(t))$

$=\frac{1}{N}\sum_{<S>} -y_n x_n$.

(이 때 $<S>$는 $y_n w^T x_n<1$인 집합 n).

$w(t+1)=w(t)-\eta V_t$. (η는 설정된 learning rate)

return w.

problem 3. 뒤쪽에 프린트하여 첨부합니다.

problem 4.

(a). D에 대해 $g^{(D)}(x)$들의 mean을 구하는 것으로, 다음과 같이 나타낼 수 있다.

$\bar{g}(x)=E_D[g^{(D)}(x)]=\frac{1}{K}\sum_{i=1}^{K} g^{(D_i)}(x)$.

(b) $E_D[(g^{(D)}(x)-f(x))^2]=E_D[(g^{(D)}-\bar{g}+\bar{g}-f)^2]$

$=E_D[(g^{(D)}-\bar{g})^2+(\bar{g}-f)^2-2(g^{(D)}-\bar{g})(\bar{g}-f)]$.

이 때, $E_D(\bar{g}-f)$의 경우 D와 independent하므로 E_D를 취해도 같다. 또한 $E_D[g^{(D)}-\bar{g}]=E[g^{(D)}]-\bar{g}$

$=\bar{g}-\bar{g}=0$이므로, 준식은 다음과 같다.

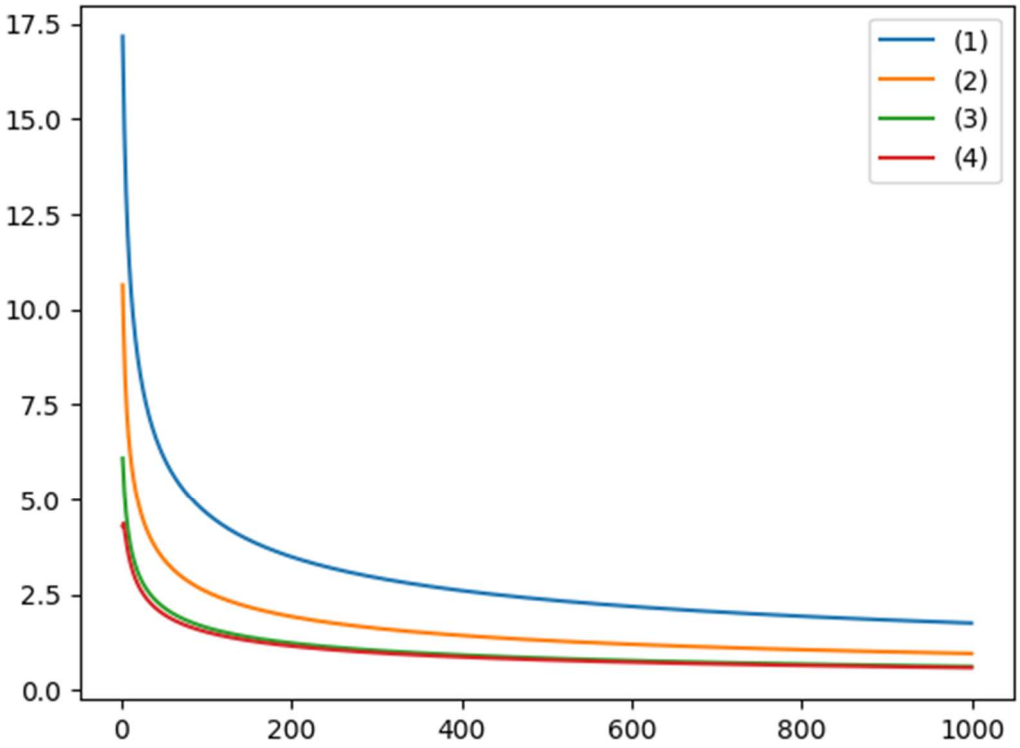
$=E_D[(g^{(D)}-\bar{g})^2]+(\bar{g}-f)^2=\text{var}(x)+\text{bias}(x)$.

(c) $E_D[E_{out}(g^{(D)})]=E_x[E_D[(g^{(D)}(x)-f(x))^2]]$.

$=E_x[\text{var}(x)+\text{bias}(x)]=E_x[\text{var}(x)]+E_x[\text{bias}(x)]$

$=\text{bias}+\text{var}$ ((b)에서 증명한 것을 사용)

Problem 3



위 그림은 python의 matplotlib module을 이용하여 각 error bound를 plot한 것이다. d_{VC} = 50, δ = 0.05에 대해, perceptron을 가정하여 $m_H(2N)$ <= (2N)^50 + 1을 이용하였다.

그림에서 확인할 수 있듯, Devroye bound가 가장 빠르게 수렴하는 것을 확인할 수 있다. 이때 Parrondo and van den Broke와 비슷한 수준을 보인다.

problem 5.

(a). $Xw+b=\begin{bmatrix}-0.5\\-4.5\\0.7\end{bmatrix}$ 이므로,

$y_n(w^Tx_n+b)$가 $n=1$에서 0.5이다.

$\therefore \rho=0.5$.

(b) ~~(b_n, W_n)~~ $(b', w')=\frac{1}{\rho}(b, w)$라 할때,

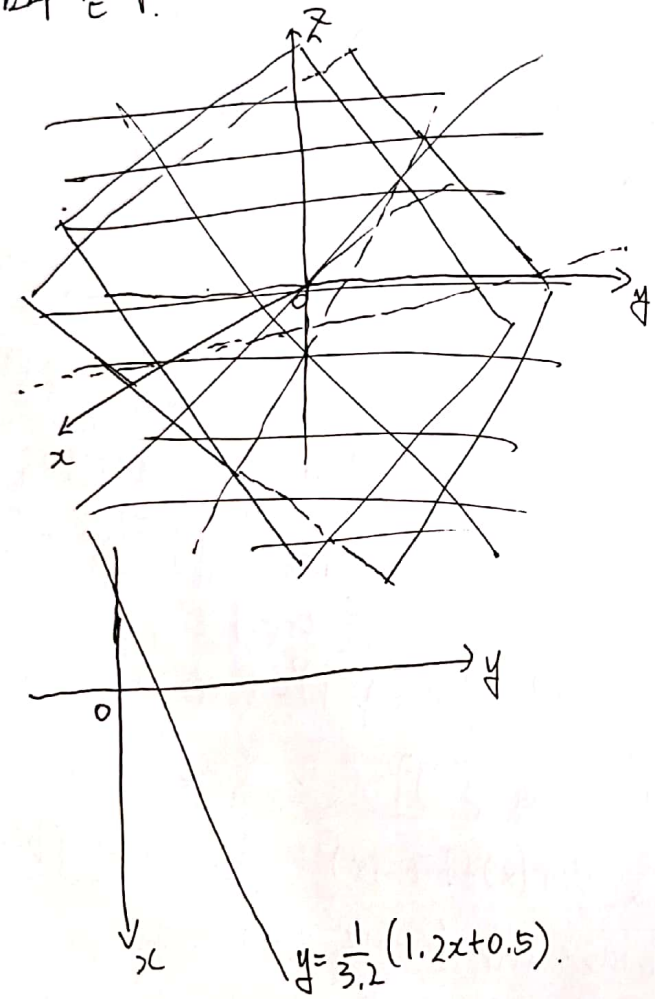
$b'=-1$, $w'=\begin{bmatrix}2.4\\-6.4\end{bmatrix}$이고,

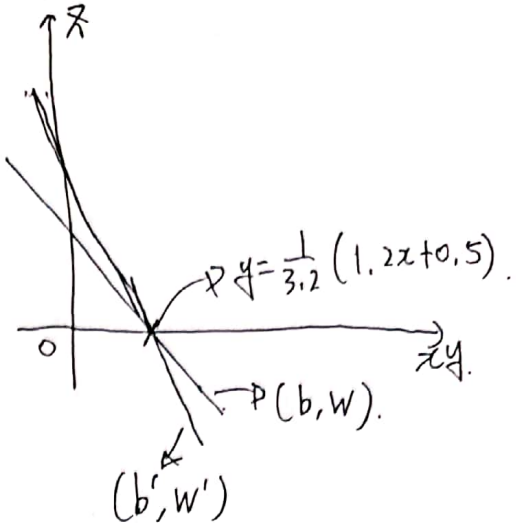
$Xw'+b'=\begin{bmatrix}-1\\-9\\1.4\end{bmatrix}$이다.

$\therefore$ 이때 $\rho=\min y_n(w^Tx_n+b)=1$이다.

(c) ~~$y=Wx+$~~ $y=w^Tx+b$에 대해

(w, b)와 (w', b')에 대해 hyperplane은 다음과 같다.





위 그림은 각각 xy평면, plane을 옆에서 본 그림이다. 그림과 같이 X_n에 대해 같은 seperator로 작용하며, w^Tx_n+b값이 상수배가 되는 것을 확인할 수 있다.