Machine Learning Spring 2019 HW2

Yueh Cheng Liu
National Taiwain University

April 29, 2019

1

$$\nabla F(A,B) = \begin{bmatrix} \frac{\delta F(A,B)}{\delta A} \\ \frac{\delta F(A,B)}{\delta B} \end{bmatrix}$$

$$\begin{aligned}
\frac{\delta F(A,B)}{\delta A} &= \frac{1}{N}\sum_{n=1}^{N} \frac{\delta(-y_n(Az_n+B))}{\delta A}\frac{e^{-y_n(Az_n+B)}}{1+e^{-y_n(Az_n+B)}} \\
&= \frac{1}{N}\sum_{n=1}^{N} p_n \frac{\delta(-y_n(Az_n+B))}{\delta A} \\
&= \frac{1}{N}\sum_{n=1}^{N} -p_n y_n z_n
\end{aligned}$$

and

$$\begin{aligned}
\frac{\delta F(A,B)}{\delta B} &= \frac{\delta F(A,B)}{\delta B} \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{\delta(-y_n(Az_n+B))}{\delta B}\frac{e^{-y_n(Az_n+B)}}{1+e^{-y_n(Az_n+B)}} \\
&= \frac{1}{N}\sum_{n=1}^{N} p_n \frac{\delta(-y_n(Az_n+B))}{\delta B} \\
&= \frac{1}{N}\sum_{n=1}^{N} -p_n y_n
\end{aligned}$$

2

$$H(F) = \begin{bmatrix} \frac{\delta^2 F(A,B)}{\delta A^2} & \frac{\delta^2 F(A,B)}{\delta A \delta B} \\ \frac{\delta^2 F(A,B)}{\delta B \delta A} & \frac{\delta^2 F(A,B)}{\delta B^2} \end{bmatrix}$$

$$\begin{aligned}
\frac{\delta\theta(x)}{\delta x} &= \frac{\delta \frac{e^x}{1+e^x}}{\delta x} \\
&= \frac{\delta(1+e^{-x})^{-1}}{\delta x} \\
&= -e^{-x} \cdot -(1+e^{-x})^{-2} \\
&= \frac{1}{1+e^{-x}} \frac{e^{-}x}{1+e^{-x}} \\
&= \theta(x)(1-\theta(x))
\end{aligned}$$

$$\begin{aligned}
\frac{\delta^2 F(A,B)}{\delta A^2} &= \frac{\delta \frac{1}{N} \sum_{n=1}^{N} -y_n z_n \theta(-y_n(Az_n + B))}{\delta A} \\
&= \frac{1}{N} \sum_{n=1}^{N} -y_n z_n \theta(-y_n(Az_n + B))(1 - \theta(-y_n(Az_n + B))) \frac{\delta(-y_n(Az_n + B))}{\delta A} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 z_n^2 \theta(-y_n(Az_n + B))(1 - \theta(-y_n(Az_n + B))) \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 z_n^2 p_n(1 - p_n)
\end{aligned}$$

$$\begin{aligned}
\frac{\delta^2 F(A,B)}{\delta B^2} &= \frac{\delta \frac{1}{N} \sum_{n=1}^{N} -y_n \theta(-y_n(Az_n + B))}{\delta B} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 \theta(-y_n(Az_n + B))(1 - \theta(-y_n(Az_n + B))) \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n(1 - p_n)
\end{aligned}$$

$$
\begin{aligned}
\frac{\delta^2 F(A,B)}{\delta A \delta B} = \frac{\delta^2 F(A,B)}{\delta B \delta A} &= \frac{\delta \frac{1}{N}\sum_{n=1}^{N} -y_n z_n \theta(-y_n(Az_n+B))}{\delta B} \\
&= \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n \theta(-y_n(Az_n+B))(1-\theta(-y_n(Az_n+B))) \\
&= \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n)
\end{aligned}
$$

3

$e^{-x} = 0$ when $x \to \infty$. The target function of soft margin SVM with rbf kernel will be

$$
\min_\alpha \lim_{\gamma\to\infty} \frac{1}{2}\sum_{n,m} \alpha_n \alpha_m y_n y_m e^{-\gamma||x_n - x_m||^2} - \sum_n \alpha_n = \min_\alpha -\sum_n \alpha_n
$$

with constraint $\sum_n y_n \alpha_n = 0$ and $0 \le \alpha_n \le C$.

Since the number of postive and negative samples are the same, we can choose $\alpha_n = C$ for all n to achieve smallest target function. The optimal $\boldsymbol{\alpha}$ will be all-C vector.

4

Given $N = 2$,

$$
\begin{aligned}
w_1 x_1 + w_0 = x_1 - x_1^2 \\
w_1 x_2 + w_0 = x_2 - x_2^2
\end{aligned}
$$

Solve w_0 and w_1

$$
w_1(x_1 - x_2) = (x_1 - x_1^2) - (x_2 - x_2^2)
$$

$$
w_1 = \frac{(x_1 - x_2)(1 - x_1 - x_2)}{x_1 - x_2} = 1 - x_1 - x_2
$$

$$w_0 = x_1 - x_1^2 - (1 - x_1 - x_2)x_1 = x_1x_2$$

Since $x_1, x_2 \in \text{Uinform}(0, 1)$, $\mathbb{E}\, 1 - x_1 - x_2 = 0$ and $\mathbb{E}\, x_1x_2 = 0.25$

$$\bar{g}(x) = \mathbb{E}\, w_1x + w_0 = 0.25$$

5

$$\begin{aligned}(\tilde{y_n} - w^T\tilde{x_n})^2 &= u_n(y_n - w^T x_n)^2 \\ &= (\sqrt{u_n}y_n - \sqrt{u_n}w^T x_n)^2\end{aligned}$$

so that $(\tilde{x_n}, \tilde{y_n}) = (\sqrt{u_n}x_n, \sqrt{u_n}y_n)$

6

Let ϵ_t be the weighted error at step t and $k_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$.

$\epsilon_1 = 0.22$ and $k_1 = \sqrt{\frac{1-\epsilon_1}{\epsilon_1}}$

$$\begin{aligned}\frac{u_+^{(2)}}{u_-^{(2)}} &= \frac{u_+^{(1)}/k_1}{u_-^{(1)} * k_1} \\ &= \frac{1}{k_1^2} = \frac{0.22}{0.78} = 0.282\end{aligned}$$

7

For integers between $[-M, M]$ has $2M$ interval. s can be $+1$ or -1 and d different feature to choose. Plus the 2 decision stumps where $g(x) = +1$ and $g(x) = -1$ for all x which are not effected by d. The number of different decision stump is $2d \cdot 2M + 2 = 4dM + 2 = 42$.

8

Given i, s, there are $|x_i - x_i'|$ number of θ that makes $s \cdot \text{sign}(x_i - \theta)s \cdot \text{sign}(x_i' - \theta) = -1$ and $2M - |x_i - x_i'|$ number of θ that makes $s \cdot \text{sign}(x_i - \theta)s \cdot \text{sign}(x_i' -$

$\theta) = +1$. Combining the result of problem 7,

$$\begin{aligned} K_{ds}(x, x') &= 2 + 2\sum_{i=1}^{d}(2M - |x_i - x'_i| - |x_i - x'_i|) \\ &= 2 + 4dM - 4\sum_{i=1}^{d}|x_i - x'_i| \end{aligned}$$

9

$\lambda = 50.0$ has the minimum $E_{in} = 0.315$

10

$\lambda = 0.05, 0.5, 5$ has the minimum $E_{out} = 0.36$

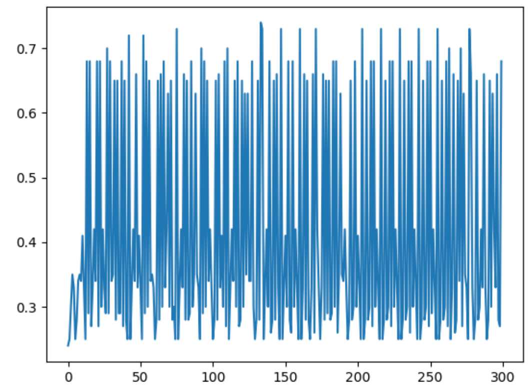
11

$\lambda = 50.0$ has the minimum $E_{in} = 0.31$. The result is a little smaller then a single ridge regression. This is probably because of bagging.

12

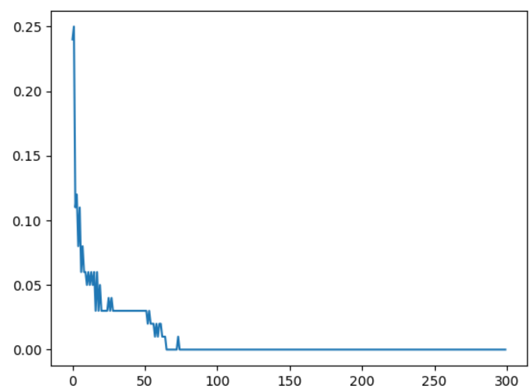
$\lambda = 0.05, 0.5$ has the minimum $E_{out} = 0.36$. The result is the same as 10. The boostrapping and bagging technique does not help E_{out} in this case.

13



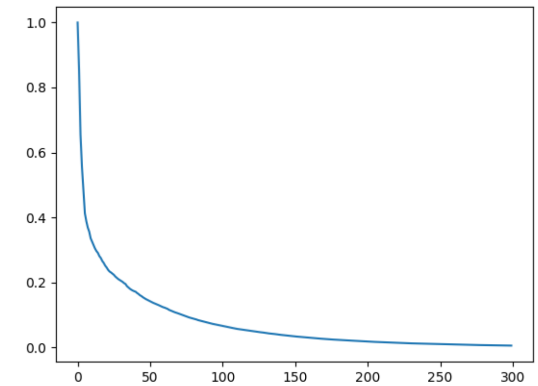
$E_{in}(g_T) = 0.68$. $E_{in}(g_t)$ is low at first. After a few steps, the weights of some hard examples (or noises) are larger making the model to fit on those samples. This increases $E_{in}(g_t)$ and lower the weight of those examples, causing the decrease of $E_{in}(g_{t+1})$. This repeating routine makes the bouncing curve.

14



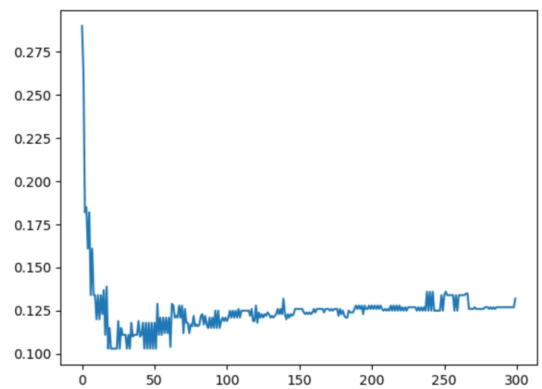
$E_{in}(G_T) = 0$. Following the theoretical guarantee of AdaBoost, $E_{in}(G_t)$ decreases to 0 fastly.

15



$U_T = 0.0055$. By theoretical prove, U_t will be decreasing and become almost 0.

16



$E_{out}(G_T) = 0.132$. $E_{out}(G_t)$ first decrease to almost 0.1 and then increase a litle due to overfitting.

17

$$\epsilon_t = \frac{\sum_i u_{t,i}[y_i \neq h_t(x_i)]}{\sum_i u_{t,i}} = \frac{\sum_i u^-_{t,i}}{\sum_i u_{t,i}}$$

and

$$1 - \epsilon_t = \frac{\sum_i u^+_{t,i}}{\sum_i u_{t,i}}$$

$$\begin{aligned}
U_{t+1} &= \sum_i u^+_{t,i}\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_i u^-_{t,i}\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \\
&= \sqrt{\epsilon_t(1-\epsilon_t)}(\frac{\sum_i u^+_{t,i}}{1-\epsilon_t} + \frac{\sum_i u^-_{t,i}}{\epsilon_t}) \\
&= \sqrt{\epsilon_t(1-\epsilon_t)}2\sum_i u_{t,i} \\
&= 2U_t\sqrt{\epsilon_t(1-\epsilon_t)}
\end{aligned}$$

$$\begin{aligned}
\epsilon(1-\epsilon) &= -(\epsilon^2 - \epsilon + \frac{1}{4}) + \frac{1}{4} \\
&= -(\epsilon - \frac{1}{2})^2 + \frac{1}{4}
\end{aligned}$$

Thus, $\sqrt{\epsilon(1-\epsilon)}$ will be smaller when ϵ is a way from $\frac{1}{2}$, and $\sqrt{\epsilon_t(1-\epsilon_t)} \leq \sqrt{\epsilon(1-\epsilon)}$ since $\epsilon_t \leq \epsilon < \frac{1}{2}$

18

$$\begin{aligned}
E_{in}(G_T) &\leq U_{T+1} \\
&\leq 2U_T\sqrt{\epsilon(1-\epsilon)} \\
&= U_1(2\sqrt{\epsilon(1-\epsilon)})^T \\
&\leq (e^{-2(\frac{1}{2}-\epsilon)^2})^T
\end{aligned}$$

$e^{-2T(\frac{1}{2}-\epsilon)^2}$ will decrease as T gets larger.

Because T is discrete, we only need to prove that after $T = O(\log N)$ $E_{in}(G_T)$ becomes $\frac{1}{N}$. Let

$$(e^{-2(\frac{1}{2}-\epsilon)^2})^T = \frac{1}{N}$$

$$T = \frac{-\log N}{\log e^{-2(\frac{1}{2}-\epsilon)^2}} = \frac{\log N}{2(\frac{1}{2}-\epsilon)^2} = O(\log N)$$