HW3

November 16, 2016

A. Boosting

1. I use 10-fold cross validation to test the average error, the result is shown in Fig 1. k is selected 3 as maximum.

   Next, I set $T^* = 400$ and plot the error on the test data, shown in Fig 2.

   Obviously, AdaBoost outperforms SVM (5% vs 15% error), which shows weak learner can be trained robust.

2. First, we don't care about what values α_t and Z_t are. In each iteration, we still choose $h_t \in \mathbb{H}$ with the smallest error $1_{y_i h_t(x_i)<0}$. So the algorithm structure is the same, the only difference is the exact values.

   The normalized factor

$$
\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{i:y_i h_t(x_i)=1} D_t(i) e^{-\alpha} + \sum_{i:y_i h_t(x_i)=0} D_t(i) + \sum_{i:y_i h_t(x_i)=-1} D_t(i) e^{\alpha_t} \\
&= \epsilon_t^1 e^{-\alpha_t} + \epsilon_t^0 + \epsilon_t^{-1} e^{\alpha_t} \\
&\text{(choose } \alpha_t \text{ s.t. } \min Z_t) \text{ we get } \alpha_t = \frac{1}{2} \ln \frac{\epsilon_t^1}{\epsilon_t^{-1}} \\
&= 2\sqrt{\epsilon_t^1 \epsilon_t^{-1}} + \epsilon_t^0
\end{aligned}
$$

   (a) The objective function is

$$F(\boldsymbol{\alpha}) = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{s=i}^{n} \alpha_s h_s(x_i)}.$$

   And let e_t be the tth unit vector in $\mathbb{R}^n$. We need to find the greatest gradient each iteration. Like the original AdaBoost,

$$F(\boldsymbol{\alpha}_{t-1} + \eta \boldsymbol{e}_t) = \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{s=i}^{n} \alpha_s h_s(x_i) - y_i \eta h_t(x_i)}.$$

   Then

$$
\begin{aligned}
F'(\boldsymbol{\alpha}_{t-1}, \boldsymbol{e}_t) &= \frac{1}{m} \sum_{i=1}^{m} -y_i h_t(x_i) e^{-y_i \sum_{s=i}^{n} \alpha_s h_s(x_i)} \\
&= -\frac{1}{m} \sum_{i=1}^{m} y_i h_t(x_i) m D_t(i) \prod_{s=1}^{t-1} Z_s \\
&= -\left[\sum_{i:y_i h_t(x_i)=1} D_t(i) + 0 - \sum_{i:y_i h_t(x_i)=-1} D_t(i) \right] \prod_{s=1}^{t-1} Z_s \\
&= (\epsilon_t^{-1} - \epsilon_t^1) \prod_{s=1}^{t-1} Z_s
\end{aligned}
$$

   As we find the direction with greatest gradient, and ϵ_t^{-1} is the error rate, we will pick h_t with the smallest ϵ_t^{-1}.

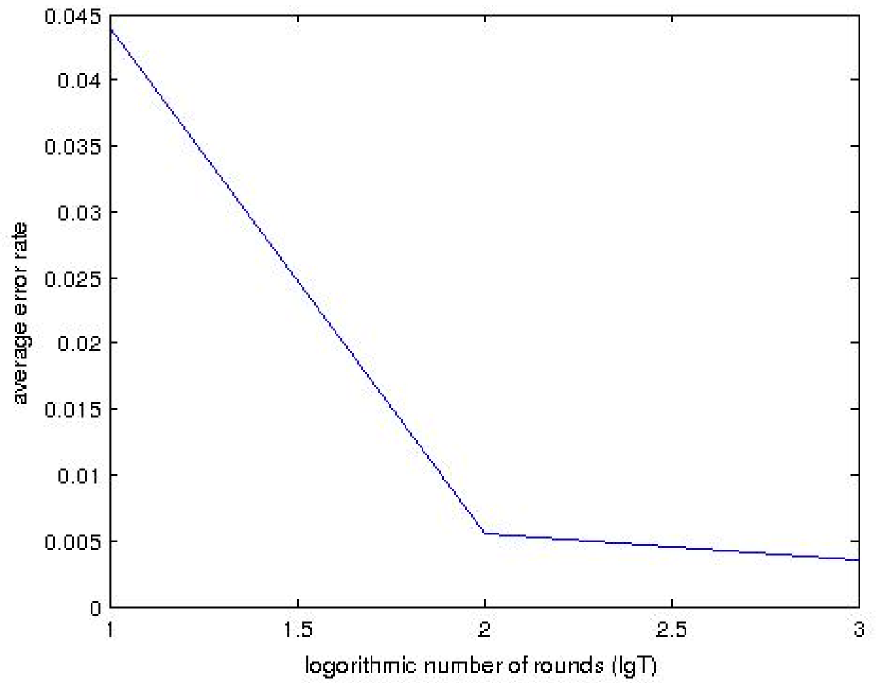

Figure 1: Average cross validation error versus $lg(T)$.

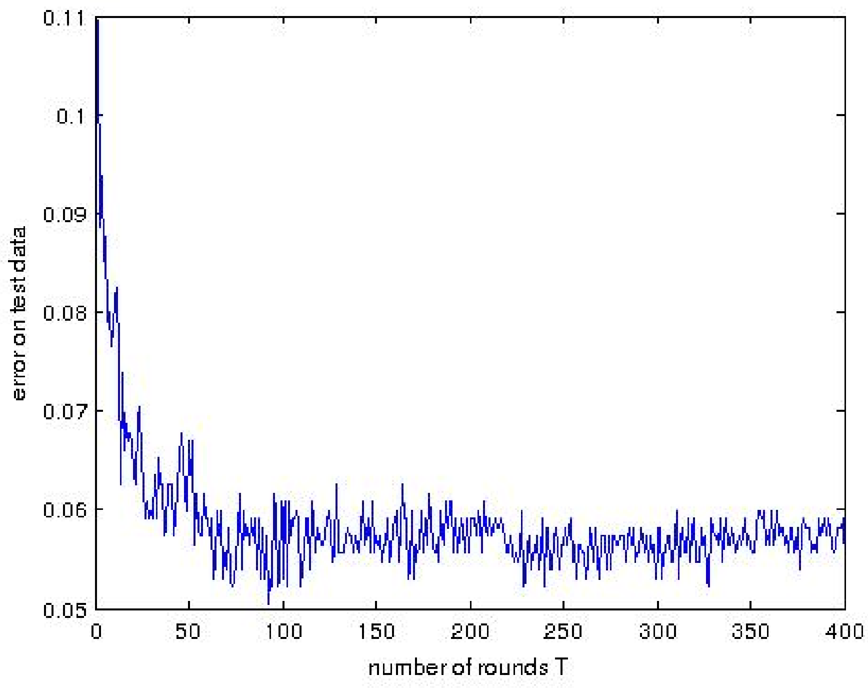

Figure 2: Test error on the test data.

For the step size η,

$$\begin{aligned}
\frac{dF(\boldsymbol{\alpha}_{t-1}+\eta \boldsymbol{e}_t)}{d\eta} = 0 &\Leftrightarrow -\sum_{i=1}^{m} y_i h_t(x_i) e^{-y_i \sum_{s=1}^{t-1} \alpha_s h_s(x_i)} e^{-\eta y_i h_t(x_i)} = 0 \\
&\Leftrightarrow \sum_{i=1}^{m} y_i h_t(x_i) D_t(i) m \prod_{s=1}^{t-1} Z_s e^{-\eta y_i h_t(x_i)} = 0 \\
&\Leftrightarrow \sum_{i=1}^{m} y_i h_t(x_i) D_t(i) e^{-\eta y_i h_t(x_i)} = 0 \\
&\Leftrightarrow \epsilon_t^1 e^{-\eta} - \epsilon_t^{-1} e^{\eta} = 0 \\
&\Leftrightarrow \eta = \frac{1}{2} \ln \frac{\epsilon_t^1}{\epsilon_t^{-1}}
\end{aligned}$$

which is the same as α_t as discussed previously.

(b) Edge can still be defined as

$$\gamma_t(D) = \frac{1}{2}\sum_{i=1}^{m} y_i h_t(x_i) D(i) = \frac{1}{2}(\epsilon_t^1 - \epsilon_t^{-1}.)$$

Then the weak learning assumption would be: $\exists \gamma > 0$ s.t. $\forall D$ and $\forall h_t$, $\gamma_t(D) > \gamma$ holds. i.e. the best edge $\gamma^* > 0$

(c)

$H \in \{-1, 0, 1\}^X$
function AdaBoost3($S = (x_1, y_1), \cdots, (x_m, y_m)$)
  for $i \leftarrow 1$ to m **do**
    $D_1(i) = \frac{1}{m}$
  end for
  for $t \leftarrow 1$ to T **do**
    $h_t \leftarrow$ base classifier in H with small error ϵ_t^{-1}
    $\alpha_t \leftarrow \frac{1}{2}\ln\frac{\epsilon_t^1}{\epsilon_t^{-1}}$
    $Z_t = 2\sqrt{\epsilon_t^1 \epsilon_t^{-1}} + \epsilon_t^0$
    for $i \leftarrow 1$ to m **do**
      $D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$
    end for
    $f_t = \sum_{i=1}^{t} \alpha_s h_s$
  end for
  return f_T
end function

(d)

$$\begin{aligned}
\hat{R}(h) &= \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(x_i) < 0} \\
&\le \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f(x_i)} \\
&\le \frac{1}{m}\sum_{i=1}^{m} D_{T+1}(i) m \prod_{t=1}^{T} Z_t \\
&= \prod_{t=1}^{T} Z_t \\
&= \prod_{t+1}^{T} \left[2\sqrt{\epsilon_t^1 \epsilon_t^{-1}} + \epsilon_t^0 \right]
\end{aligned}$$

B. On-line Learning

1. The $max(x, 0)$ function is differentiable except at point 0. The set B is all the vectors that have no positive component.

$$\frac{\partial \Phi}{\partial x_i} = \frac{2}{\alpha}\left[\sum_{i=1}^{N}(x_i)_+^\alpha\right]^{\frac{2}{\alpha}-1} \alpha(x_i)_+^{\alpha-1} \tag{1}$$

Since $\alpha > 2$, the summation term on the right-hand side of the above equation is non-differentiable if it is 0, i.e., $\boldsymbol{x}$ is non-positive. But the set $\mathbb{R}^N - B$ excludes the zero situation, so the first-order derivative is differentiable, which means Φ is twice differentiable.

2. Similar to Eq. 1,

$$\nabla\Phi(\boldsymbol{R}_{t-1}) = 2\left[\sum_{i=1}^{N}(\boldsymbol{R}_{t-1,i})_+^\alpha\right]^{\frac{2}{\alpha}-1} (\boldsymbol{R}_{t-1})_+^{\alpha-1} \tag{2}$$

Because $\boldsymbol{R}_{t-1} \notin B$, we get $\sum_{i=1}^{N}(\boldsymbol{R}_{t-1,i})_+^\alpha > 0$, and

$$\nabla\Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t \le 0 \Leftrightarrow (\boldsymbol{R}_{t-1})_+^{\alpha-1} \cdot \boldsymbol{r}_t \le 0$$

Using $r_{t,i} = L(\hat{y}_t, y_t) - L(y_{t,i}, y_t)$, $w_{t,i} = (R_{t-1,i})_+^{\alpha-1}$, and $\hat{y}_t = \frac{\sum_{i=1}^{N} w_{t,i} y_{t,i}}{\sum_{i=1}^{N} w_{t,i}} = E[y_{t,i}]$

$$\begin{aligned}
(\boldsymbol{R}_{t-1})_+^{\alpha-1} \cdot \boldsymbol{r}_t &= \sum_{i=1}^{N}(R_{t-1,i})_+^{\alpha-1} r_{t,i} \\
&= \sum_{i=1}^{N} w_{t,i}(L(\hat{y}_t, y_t) - L(y_{t,i}, y_t)) \\
&= \sum_{i=1}^{N} w_{t,i}(L(E[y_{t,i}], y_t) - L(y_{t,i}, y_t)) \\
&\le \sum_{i=1}^{N} w_{t,i}(E[L(y_{t,i}, y_t)] - L(y_{t,i}, y_t)) \\
&= \sum_{i=1}^{N} w_{t,i} \frac{\sum_{i=1}^{N} w_{t,i}(E[L(y_{t,i}, y_t)] - L(y_{t,i}, y_t))}{\sum_{i=1}^{N} w_{t,i}} \\
&= \sum_{i=1}^{N} w_{t,i} E[E[L(y_{t,i}, y_t)] - L(y_{t,i}, y_t)] \\
&= 0
\end{aligned}$$

□

3. As shown in Eq. 2,

$$\begin{aligned}
\nabla^2\Phi(\boldsymbol{u}) &= \nabla\left(2\left[\sum_{i=1}^{N}(u_i)_+^\alpha\right]^{\frac{2}{\alpha}-1} (\boldsymbol{u})_+^{\alpha-1}\right) \\
&= 2(2-\alpha)\left[\sum_{i=1}^{N}(u_i)_+^\alpha\right]^{\frac{2}{\alpha}-2} (\boldsymbol{u})_+^{\alpha-1}(\boldsymbol{u}^T)_+^{\alpha-1} + 2(\alpha-1)\left[\sum_{i=1}^{N}(u_i)_+^\alpha\right]^{\frac{2}{\alpha}-1} (\boldsymbol{u}_i)_+^{\alpha-2} \qquad (3) \\
&\le 2(\alpha-1)\boldsymbol{\Lambda} \qquad (4)
\end{aligned}$$

In Eq. 3, the first term is a positive semi-definite matrix as $(\boldsymbol{u})_+^T(\boldsymbol{u})_+$ is symmetric. Because the summation term is positive for any $\boldsymbol{u}$ and $\alpha > 2$, the first term is non-positive. The second term is a diagonal matrix, where

$$\Lambda_{ii} = \left[\sum_{i=1}^{N}(u_i)_+^\alpha\right]^{\frac{2}{\alpha}-1} (u_i)_+^{\alpha-2} = \left(\frac{(u_i)_+}{||\boldsymbol{u}_+||_\alpha}\right)^{\alpha-2} \triangleq \lambda_i^{\alpha-2}$$

is identity matrix. [Note $\sum_i \lambda_i^\alpha = 1$] That's how Eq. 4 is obtained.

Therefore,

$$\begin{aligned}
\boldsymbol{r}^T[\nabla^2\Phi(\boldsymbol{u})]\boldsymbol{r} &\leq 2(\alpha-1)\boldsymbol{r}^T\boldsymbol{\Lambda}\boldsymbol{r} \\
&= 2(\alpha-1)\sum_{i=1}^{N}\lambda_i^{\alpha-2}r_i^2 \\
(R_i = r_i^2) &= 2(\alpha-1)(\boldsymbol{\lambda}^{\alpha-2}\cdot\boldsymbol{R}) \\
&\leq 2(\alpha-1)||\boldsymbol{\lambda}^{\alpha-2}||_{\frac{\alpha}{\alpha-2}}||\boldsymbol{R}||_{\frac{\alpha}{2}} \\
&= 2(\alpha-1)\left(\sum_{i=1}^{N}\lambda_i^{\alpha}\right)^{\frac{\alpha-2}{\alpha}}\left(\sum_{i=1}^{N}r_i^{\alpha}\right)^{\frac{2}{\alpha}} \\
&= 2(\alpha-1)||\boldsymbol{r}||_\alpha^2
\end{aligned}$$

□

4.

$$\begin{aligned}
\Phi(\boldsymbol{R}_t)-\Phi(\boldsymbol{R}_{t-1}) &= \sum_{n=1}^{\infty}\frac{\Phi^{(n)}(\boldsymbol{R}_{t-1})}{n!}(\boldsymbol{R}_t-\boldsymbol{R}_{t-1})^n(\text{Taylor expansion}) \\
&\leq \nabla\Phi(\boldsymbol{R}_{t-1})\cdot(\boldsymbol{R}_t-\boldsymbol{R}_{t-1})+(\boldsymbol{R}_t-\boldsymbol{R}_{t-1})^T\frac{\nabla^2\Phi(\boldsymbol{R}_{t-1})}{2}(\boldsymbol{R}_t-\boldsymbol{R}_{t-1}) \\
&= \nabla\Phi(\boldsymbol{R}_{t-1})\cdot\boldsymbol{r}_t+\boldsymbol{r}_t^T\frac{\nabla^2\Phi(\boldsymbol{R}_{t-1})}{2}\boldsymbol{r}_t \\
(\text{Results of Q2 and Q3}) &\leq (\alpha-1)||\boldsymbol{r}_t||_\alpha^2
\end{aligned}$$

□

5. In that case,

$$\Phi(\boldsymbol{R}\in B)=\mathbf{0}$$

6.

$$\begin{aligned}
\Phi(\boldsymbol{R}_T) &= \sum_{t=1}^{T}(\Phi(\boldsymbol{R}_t)-\Phi(\boldsymbol{R}_{t-1}))+\Phi(\boldsymbol{R}_0) \\
&\leq (\alpha-1)\sum_{t=1}^{T}||\boldsymbol{r}_t||_\alpha^2 \\
&\leq (\alpha-1)\sum_{t=1}^{T}\left(\sum_{i=1}^{N}M^\alpha\right)^{\frac{2}{\alpha}} \\
&= (\alpha-1)TM^2N^{\frac{2}{\alpha}} \qquad (5)
\end{aligned}$$

7. Note that

$$R_{t,i}=\sum_{i=1}^{t}r_{t,i}=r_{t,i}+R_{t-1,i}\geq R_{t-1,i}$$

and $\Phi(\boldsymbol{x})$ is a non-deacreasing function with respect to α.
Therefore

$$\Phi(\boldsymbol{R}_T)\geq\lim_{\alpha\to\infty}||(\boldsymbol{R}_T)_+||_\alpha^2=(\max_{1\leq i\leq N}R_{T,i})^2=R_T^2 \qquad (6)$$

8. Combine Eq. 5 and Eq. 6, we have

$$R_T^2\leq(\alpha-1)TM^2N^{\frac{2}{\alpha}}$$

We pick an α s.t. $\frac{dR_T}{d\alpha}=0$ and do some approximation, we get

$$\alpha\approx 2\ln N$$