HW3

November 8, 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}.

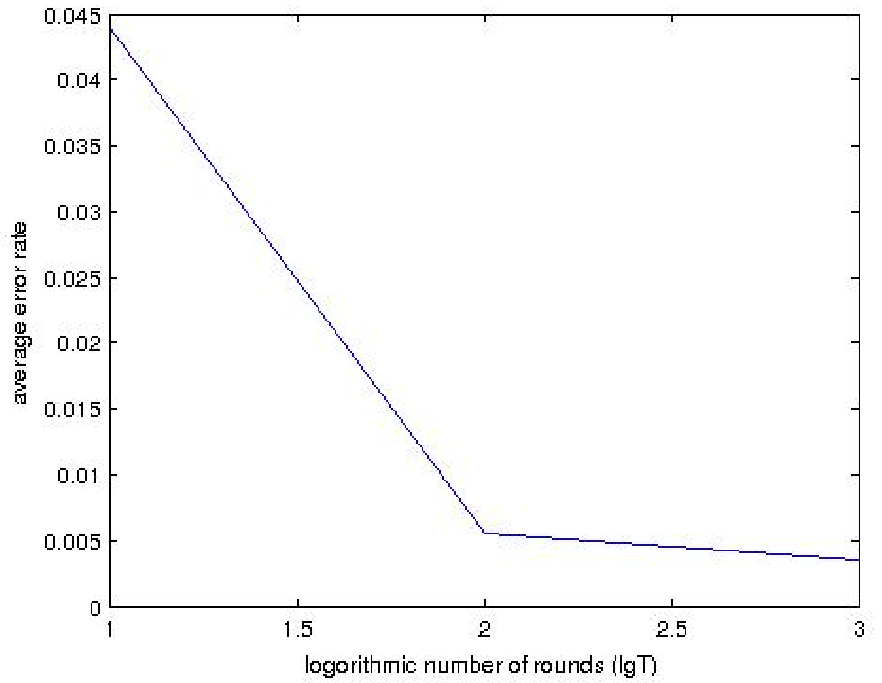

Figure 1: Average cross validation error versus $lg(T)$.

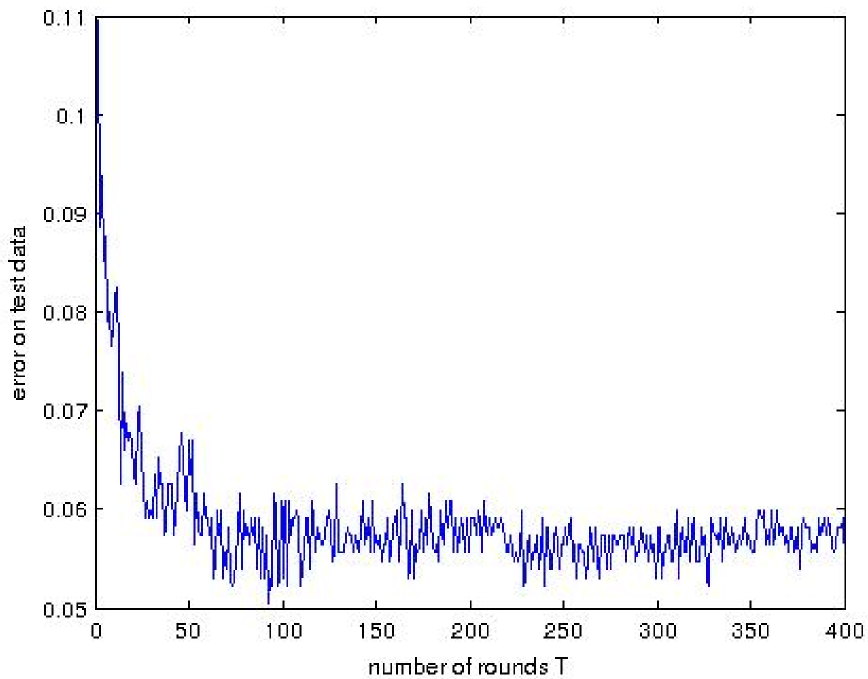

Figure 2: Test error on the test data.

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)

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