Homework #6

6.1 Bayesian Universe

(1) Because (ω,θ) is fixed, $P\big(Z\big|(\omega,\theta)\big) = P(Z)$. Derive as follows:

$$P(Z) = P(\{(x_n, y_n)\}_{n=1}^{N}) = \prod_{n=1}^{N} P(x_n)P(y_n|x_n) = \prod_{n=1}^{N}\big(P(x_n)P(y_n|\rho_n)\big) \quad (5.1.1)$$

By the procedure (c), substitute those parameters in equation (5.1.1),

$$P(x)P(y|\rho) = P(x)\left(\frac{1}{\sqrt{2\pi}}\exp(-(y-\rho)^2)\right) \quad (5.1.2)$$

Combine (5.1.1) and (5.1.2) we get the likelihood $P\big(Z\big|(\omega,\theta)\big)$

$$\prod_{n=1}^{N} P(x_n)P(y_n|\rho_n) = \prod_{n=1}^{N}\left(P(x_n)\left(\frac{1}{\sqrt{2\pi}}\exp(-(y_n-\rho_n)^2)\right)\right) \quad (5.1.3)$$

We look at the maximum likelihood function. Therefore, our goal function should satisfy

$$\arg\max_{(\omega,\theta)} \ln \prod_{n=1}^{N}\left(\frac{1}{\sqrt{2\pi}}\exp(-(y_n-\rho_n)^2)\right)$$

Because those parameter do not correlated to ω, θ can view as constant, the formula equal to

$$\arg\max_{(\omega,\theta)} \sum_{n=1}^{N}(-(y_n-\rho_n)^2) = \arg\min_{(\omega,\theta)} \sum_{n=1}^{N}(y_n-\rho_n)^2$$

That is same as *linear regression* that we did on Problem 2.3-(1).

(2) By the Bayesian theorem, we can derive the posteriori as follow:

$$P\big((\omega,\theta)\big|Z\big) = \frac{P\big(Z\big|(\omega,\theta)\big)P(\omega,\theta)}{P(Z)}$$

Where, the likelihood here derive as 5.1(1), that is

$$P\big(Z\big|(\omega,\theta)\big) = \prod_{n=1}^{N}\left(P(x_n)\left(\frac{1}{\sqrt{2\pi}}\exp(-(y_n-\rho_n)^2)\right)\right)$$

Therefore, posterior is

$$P\big((\omega,\theta)\big|Z\big) = \frac{\prod_{n=1}^{N}\left(P(x_n)\left(\frac{1}{\sqrt{2\pi}}\exp(-(y_n-\rho_n)^2)\right)\right)\frac{1}{\left(\sqrt{2\pi}\right)^{d+1}\sigma^{d+1}}\exp\left(-\left(\frac{\|\omega\|^2+\theta^2}{2\sigma^2}\right)\right)}{Q}$$

To maximum it, our goal function should satisfy

$$\arg\max_{(\omega,\theta)} \ln \prod_{n=1}^{N} \left(P(x_n) \left(\frac{1}{\sqrt{2\pi}} \exp(-(y_n - \rho_n)^2) \right) \right) \exp\left(-\left(\frac{\|\omega\|^2 + \theta^2}{2\sigma^2} \right) \right)$$

That is equal to

$$\arg\max_{(\omega,\theta)} \frac{1}{2} \left(\sum_{n=1}^{N} (-(y_n - \rho_n)^2) - \frac{\|\omega\|^2 + \theta^2}{2\sigma^2} \right)$$

Consider the negation, rewrite the goal function as

$$\arg\min_{(\omega,\theta)} \frac{\frac{1}{2\sigma^2}\|\omega\|^2}{2} + \frac{\frac{1}{2\sigma^2}\theta^2}{2} + \frac{1}{2}\sum_{n=1}^{N}(y_n - \rho_n)^2$$

It is same as the *regularized linear regression,* and the relation between λ, σ is

$$\lambda = \frac{1}{2\sigma^2}$$

(3) We derive the likelihood as follow:

$$\begin{aligned} P(Z|\omega,\theta) &= P(\{(x_n, y_n)\}_{n=1}^{N} | \omega_n, \theta_n) \\ &= \prod_{n=1}^{N} P(x_n|\omega_n,\theta_n) P(y_n|x_n,\omega_n,\theta_n) \\ &= \prod_{n=1}^{N} P(x_n) P(y_n|\rho_n) \end{aligned}$$

The Objective function max $P(Z|\omega,\theta)$ should satisfy the equation

$$\begin{aligned} &\arg\max_{\omega,\theta} \ln \prod_{n=1}^{N} P(x_n) P(y_n|\rho_n) \\ &= \arg\max_{\omega,\theta} \ln \prod_{n=1}^{N} P(y_n|\rho_n) \\ &= \arg\max_{\omega,\theta} \ln \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n\rho_n)} \\ &= \arg\max_{\omega,\theta} \sum_{n=1}^{N} -(1 - y_n\rho_n) \\ &= \arg\min_{\omega,\theta} \sum_{n=1}^{N} (1 - y_n\rho_n) \end{aligned}$$

The *Logistic Regression* is equivalently gives the maximum likelihood estimated of (ω, θ).

6.2 Power of Adaptive Boosting

(1) In the first iteration we get

$$U^{(0)} = \sum_{n=1}^{N} \frac{1}{N} = \frac{1}{N} \sum_{n=1}^{N} 1 = 1$$

(2) Proof by Induction:

(a) Base: When t = 1, denote $B_b = -y_n \alpha_b h_b(x_n)$, by definition

$$U^{(1)} = U^{(2-1)} = \frac{1}{N} \sum_{n=1}^{N} 1 * \exp(B_1) = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{1} \alpha_\tau h_\tau(x_n)\right)$$

(b) Inductive: Suppose that

$$U^{(k)} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{k} \alpha_\tau h_\tau(x_n)\right)$$

We have

$$U^{(k+1)} = U^{(k+2-1)} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{k} \alpha_\tau h_\tau(x_n)\right) * \exp(B_{k+1})$$

$$= \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{k+1} \alpha_\tau h_\tau(x_n)\right)$$

(3) Express $\nu(H) - U^{(T)} = s$, denote that $\sum_{\tau=1}^{T} \alpha_\tau h_\tau(x_n) = v$

$$s = \frac{1}{N} \sum_{n=1}^{N} I\left[y_n \neq \text{sign}\left(\sum_{\tau=1} \alpha_\tau h_\tau(x_n)\right)\right] - \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{T} \alpha_\tau h_\tau(x_n)\right)$$

The sign of s	$y_n \geq 0$	$y_n < 0$
The sign of v ≥ 0	Zero - Nonzero	$1 - (\text{some value} \geq 1)$
The sign of v < 0	$1 - (\text{some value} \geq 1)$	Zero – Nonzero

In each cases we get the result that $s \leq 0$.

(4) Derive the step as follow:

$$
\begin{aligned}
& U^{(t)} \\
&= \sum_{n=1}^{N} u_n \exp\left(-\alpha_t y_n h_t(x_n)\right) \\
&= \sum_{n+} u_n \exp(-\alpha_t) + \sum_{n-} u_n \exp(\alpha_t) \\
&= \sum_{n+} u_n \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_{n-} u_n \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \\
&= \sum_{n+} u_n \epsilon_t \sqrt{\frac{1}{\epsilon_t(1-\epsilon_t)}} + \sum_{n-} u_n (1-\epsilon_t) \sqrt{\frac{1}{\epsilon_t(1-\epsilon_t)}} \\
&= \sqrt{\frac{1}{\epsilon_t(1-\epsilon_t)}} \left(\epsilon_t \sum_{n+} u_n + (1-\epsilon_t) \sum_{n-} u_n\right) \\
&= \sqrt{\frac{1}{\epsilon_t(1-\epsilon_t)}} \left(\epsilon_t (1-\epsilon_t) U^{(t-1)} + (1-\epsilon_t)\epsilon_t U^{(t-1)}\right) \qquad \text{By the definition of } \epsilon_t \\
&= 2\sqrt{\epsilon_t(1-\epsilon_t)}\, U^{(t-1)}
\end{aligned}
$$

(5) Consider the function $E(x) = x(1-x) = -\left(x-\frac{1}{2}\right)^2 + \frac{1}{4}$. In the region $0 \le \epsilon_t \le \epsilon \le \frac{1}{2}$, we have

$$\sqrt{\epsilon_t(1-\epsilon_t)} \le \sqrt{\epsilon(1-\epsilon)}$$

(6) Consider $E(\epsilon) = \sqrt{\epsilon(1-\epsilon)} - \frac{1}{2}\exp\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)$. In the case $\epsilon = 1/2$, we have $E = 0$, than we consider $E'^{(\epsilon)} = (1-2\epsilon)\left(\frac{1}{2}(\epsilon-\epsilon^2)^{-\frac{1}{2}} + exp\left(\frac{1}{2} - 2\epsilon + 2\epsilon^2\right)\right) \le 0$ in case $\epsilon \le \frac{1}{2}$, thus, $E(\epsilon) \le 0$.

(7) Use the equation above:

$$
\begin{aligned}
& U^{(T)} \\
&= \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)} \\
&\le \prod_{t=1}^{T} \sqrt{\epsilon(1-\epsilon)} \\
&\le \prod_{t=1}^{T} \frac{1}{2}\exp\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right) \\
&\le \exp\left(-2T\left(\frac{1}{2}-\epsilon\right)^2\right)
\end{aligned}
$$

(8) Use the fact above,

$$\nu(H) \leq U^{(T)} \leq \exp\left(-2T\left(\frac{1}{2}-\epsilon\right)^2\right)$$

When T grow up, $U^{(T)} \to 0$. And we also know

$$\nu(H) \in \left\{\frac{n}{N}\right\}_{n=0}^{N}$$

Find a T such that

$$T = \frac{1}{2\left(\frac{1}{2}-\epsilon\right)^2}\log N + c = O(\log N)$$

Then we have

$$U^{(T)} < \frac{1}{N}$$

Therefore,

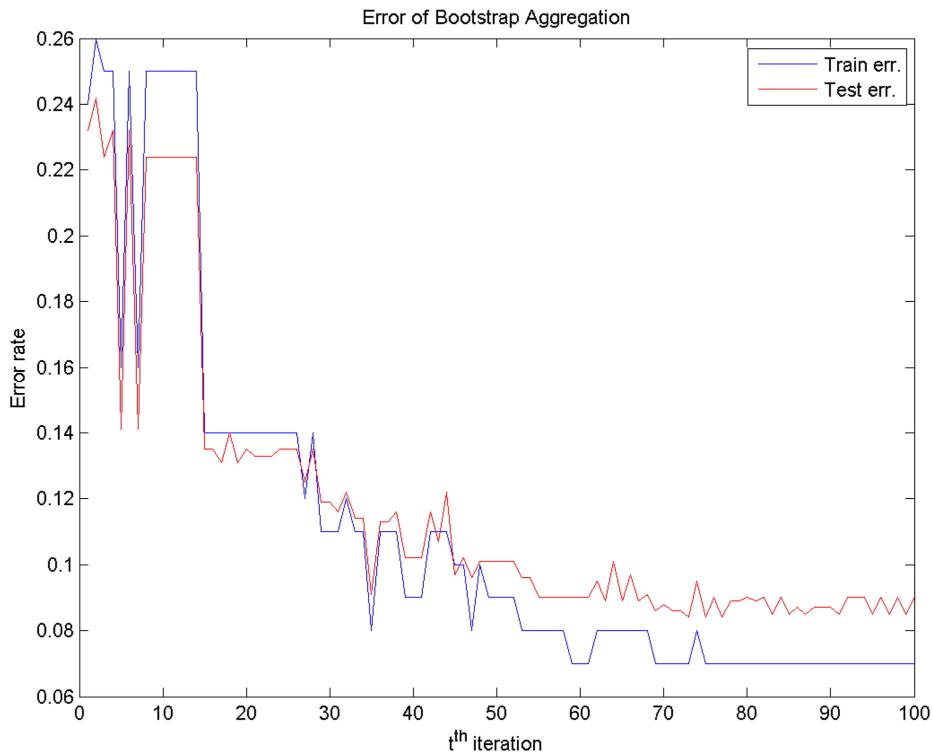
$$\nu(H) = 0$$

6.3 Experiments with Bootstrap Aggregation

(1) The training error is 0.23, and the test error is 0.255. Our brief finding:

- (a) The error always is a fixed value, unless we change θ.
- (b) One thing should be careful, the method that we find θ should cover [-1, 1], for a training data set range in [0, 1].

(2) The figure we found as follow:



Our brief finding:

- (a) Training error and testing error have strong correlated. Even a small pick on training error would reflect on testing error.
- (b) The performance seems better than expected.

(3) Pseudocode:

-1 $D =$ Sort data points increasingly in time $O(N \log N)$

-2 $L^{+} = [0,$ Aggregating the positive value from left to right$]$ in time $O(N)$

-3 $R^{-} = [$Aggregating the negative value from right to left, $0]$ in time $O(N)$

-4 $R^{+} = [$Aggregating the positive value from right to left, $0]$ in time $O(N)$

-5 $L^{-} = [0,$ Aggregating the negative value from left to right$]$ in time $O(N)$

-6 $W^{1} =$ column weighted sum on $[L+; R-]$ in time $O(N)$

-7 $W^{2} =$ column weighted sum on $[R+; L-]$ in time $O(N)$

-8 $2\theta^{1} = (A_1^1 + A_2^1) =$ The argmin value column in W^{1} in time $O(N)$

-9 $2\theta^{2} = (A_1^2 + A_2^2) =$ The argmin value column in W^{2} in time $O(N)$

-10 $\theta = \min(\theta^{1} + \theta^{2})$ in time $O(N)$

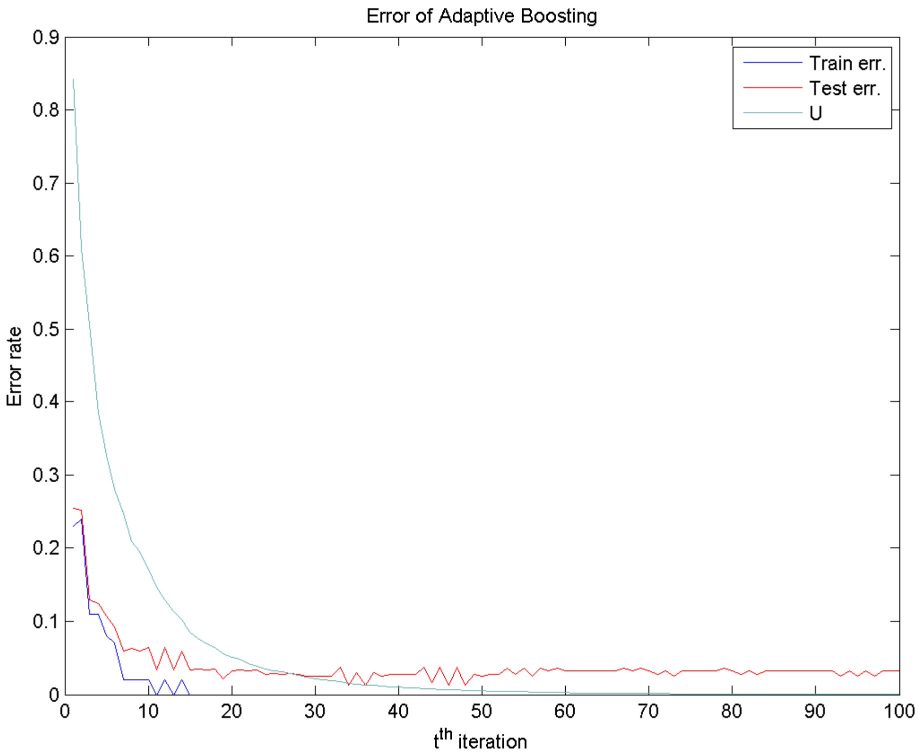
Our object function is

$$\arg\min_{h_{s,i,\theta}} \sum_{n=1}^{N} u_n I\left[y_n \neq h_{s,i,\theta}(x_n)\right]$$

Therefore, we use the pseudocode above and get minimum a good result in time $O(N \log N)$.

6.4 Experiments with Adaptive Boosting

(1) The figure show as follow:



Our brief finding:

(a) Training error `seems' always less than testing error, and it would converge to 0.

(b) U `seems' has an inverse proportion with t.

(2) Brief finding:

(a) Adaptive Boosting algorithm (AdaBoost) has a better result than Bootstrap Aggregation algorithm (Bagging).

(b) AdaBoost has a zero training error performance, but Bagging does not have that.

(c) Both of algorithms start with a not bad error rate, Bagging would go up but AdaBoost won't.