CS 260: Machine Learning **Spring 2020**

Homework 4

Hand Out: May.18 *Due: May.25*

1. **Exercise 10.1**

   Let $\epsilon, \delta \in (0,1)$, and pick k "chunks" of size $m_{\mathcal{H}}(\epsilon/2)$ according to the hint. Then we can apply A on each of these chunks and obtain $\widehat{h}_1, \cdots, \widehat{h}_k$. Note that the probability that

   $$\min_{i\in[k]} L_{\mathcal{D}}(\widehat{h}_i) \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2$$

   is at least $1 - \delta_0^k \ge 1 - \delta/2$ according to the definition of k.

   Next, we apply an ERM over the class $\widehat{\mathcal{H}} = \{\widehat{h}_1, \cdots, \widehat{h}_k\}$ with the training data being the last chunk of size $\lceil 2\log(4k/\delta)/(\epsilon/2)^2 \rceil$, and denote the output hypothesis by $\widehat{h}$. Then based on Corollary 4.6, we can derive that with probability at least $1 - \delta/2$,

   $$L_{\mathcal{D}}(\widehat{h}) \le \min_{i\in[k]} L_{\mathcal{D}}(\widehat{h}_i) + \epsilon/2 \le \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon,$$

   which concludes our proof. ■

2. **Exercise 12.1**

   Let $\mathcal{H}$ be the class of homogenous halfspaces in $\mathbb{R}^d$. Let $\mathbf{x} = \mathbf{e}_1, y = 1$ and consider the sample $S = \{(\mathbf{x}, y)\}$. Let $\mathbf{w} = -\mathbf{e}_1$. Then, $\langle \mathbf{w}, \mathbf{x} \rangle = -1$, and thus $L_S(h_{\mathbf{w}}) = 1$. Still, $\mathbf{w}$ is a local minima. Let $\epsilon \in (0,1)$. For every $\mathbf{w}'$ with $||\mathbf{w}' - \mathbf{w}|| \le \epsilon$, by Cauchy-Schwartz inequality, we have

   $$\begin{aligned} \langle \mathbf{w}', \mathbf{x} \rangle &= \langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}' - \mathbf{w}, \mathbf{x} \rangle \\ &= -1 - \langle \mathbf{w}' - \mathbf{w}, \mathbf{x} \rangle \\ &\le -1 + ||\mathbf{w}' - \mathbf{w}||_2 ||\mathbf{x}||_2 \\ &\le -1 + 1 \\ &= 0 \end{aligned}$$

   Hence, $L_S(\mathbf{w}') = 1$ as well. ■

3. **Exercise 12.2**

   Convexity: Note that the function $g : \mathbb{R} \to \mathbb{R}$, define by $g(a) = \log(1 + \exp(a))$ is convex. To see this, note that g'' is non-negative. The convexity of ℓ (or more accurately, of $\ell(\cdot, z)$ for all z) follows now from Claim 12.4.

   Lipschitzness: The funciton $g(a) = \log(1 + \exp(a))$ is 1-Lipschitz, since $|g'(a)| = \frac{\exp(a)}{1+\exp(a)} = \frac{1}{\exp(-a)+1} \le 1$. Hence, by Claim 12.7, ℓ is B - Lipschitz.

Smoothness: We claim that $g(a) = \log(1 + \exp(a))$ is 1/4-smooth. To see this, note that

$$
\begin{aligned}
g''(a) &= \frac{\exp(-a)}{(\exp(-a) + 1)^2} \\
&= (\exp(a)(\exp(-a) + 1)^2)^{-1} \\
&= \frac{1}{2 + \exp(a) + \exp(-a)} \\
&\leq 1/4
\end{aligned}
$$

Combine this with the mean value theorem, to conclude that g' is 1/4-Lipschitz. Using Claim 12.9, we conclude that ℓ is $B^2/4$ - smooth.

Boundness: The norm of each hypothesis is bounded by B according to the assumptions.

All in all, we conclude that the learning problem of linear regression is Convex-Smooth-Bounded with parameters $B^2/4$, B, and Convex-Lipschitz-Bounded with parameters B, B. ■

4. **Exercise 12.3**

   Fix some $(\mathbf{x}, y) \in \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}'||_2 \leq R\} \times \{-1, 1\}$. Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$. For $i \in [2]$, let $\ell_i = \max\{0, 1 - y\langle \mathbf{w}_i, \mathbf{x}\rangle\}$. We wish to show that $|\ell_1 - \ell_2| \leq R||\mathbf{w}_1 - \mathbf{w}_2||_2$. If both $y\langle \mathbf{w}_1, \mathbf{x}\rangle \geq 1$ and $y\langle \mathbf{w}_2, \mathbf{x}\rangle \geq 1$, then $|\ell_1 - \ell_2| = 0 \leq R||\mathbf{w}_1 - \mathbf{w}_2||_2$. Assume now that $|\{i : y\langle \mathbf{w}_i, \mathbf{x}\rangle < 1\}| \geq 1$. Assume w.l.o.g that $1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle \geq 1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle$. Hence,

$$
\begin{aligned}
|\ell_1 - \ell_2| &= \ell_1 - \ell_2 \\
&= 1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle - \max\{0, 1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle\} \\
&\leq 1 - y\langle \mathbf{w}_1, \mathbf{x}\rangle - (1 - y\langle \mathbf{w}_2, \mathbf{x}\rangle) \\
&= y\langle \mathbf{w}_2 - \mathbf{w}_1, \mathbf{x}\rangle \\
&\leq ||\mathbf{w}_1 - \mathbf{w}_2||\ ||\mathbf{x}|| \\
&\leq R||\mathbf{w}_1 - \mathbf{w}_2||
\end{aligned}
$$

   ■

5. **Exercise 13.3**

   Let $h^* \in \text{argmin}_{h \in \mathcal{H}} L_D(h)$ (for simplicity we assume that h^* exists). We have

$$
\begin{aligned}
&\mathbb{E}_{S\sim D^m}[L_D(A(S)) - L_D(h^*)] \\
&= \mathbb{E}_{S\sim D^m}[L_D(A(S)) - L_S(A(S)) + L_S(A(S)) - L_D(h^*)] \\
&= \mathbb{E}_{S\sim D^m}[L_D(A(S)) - L_S(A(S))] + \mathbb{E}_{S\sim D^m}[L_S(A(S)) - L_S(h^*)] \\
&= \mathbb{E}_{S\sim D^m}[L_D(A(S)) - L_S(A(S))] + \mathbb{E}_{S\sim D^m}[L_S(A(S)) - L_D(h^*)] \\
&\leq \epsilon_1(m) + \epsilon_2(m)
\end{aligned}
$$

   ■

6. **Exercise 14.2**

   Plugging the definition of η and T into Theorem 14.13 we obtain

$$\begin{aligned}
\mathbb{E}[L_D(\bar{\mathbf{w}})] &\le \frac{1}{1-\frac{1}{1+3/\epsilon}}\left(L_D(\mathbf{w}^*)+\frac{||\mathbf{w}^*||^2(1+3/\epsilon)\epsilon^2}{24B}\right) \\
&\le (1+3/\epsilon)\epsilon/3\left(L_D(\mathbf{w}^*)+\frac{(1+3/\epsilon)\epsilon^2}{24}\right) \\
&= (1+\epsilon/3)\left(L_D(\mathbf{w}^*)+\frac{\epsilon(\epsilon+3)}{24}\right) \\
&= L_D(\mathbf{w}^*)+\frac{\epsilon}{3}L_D(\mathbf{w}^*)+\frac{(1+\epsilon/3)\epsilon(\epsilon+3)}{24} \\
&\le L_D(\mathbf{w}^*)+\frac{\epsilon}{3}+\frac{(1+\epsilon/3)\epsilon(\epsilon+3)}{24} \\
&\le L_D(\mathbf{w}^*)+\epsilon
\end{aligned}$$

   The penultimate inequality holds because $L_D(\mathbf{w}^*) \le L_D(\mathbf{0}) \le 1$, the last inequality holds because $\epsilon \in (0,1)$. Thus, we conclude the proof. ■

7. **Exercise 14.3**

   - Clearly, $f(\mathbf{w}^*) \le 0$. If there is strictly inequality, then we can decrease the norm of $\mathbf{w}^*$ while still having $f(\mathbf{w}^*) \le 0$. But $\mathbf{w}^*$ is chosen to be of minimal norm and therefore equality mush hold. In addition, any $\mathbf{w}$ for which $f(\mathbf{w}) < 1$ must satisfy $1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle < 1$ for every i, which implies that it separates the examples.
   - A sub-gradient of f is given by $-y_i\mathbf{x}_i$, where $i \in \mathrm{argmax}\{1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\}$.
   - The resulting algorithm initialize $\mathbf{w}$ to be the all zeros vector and at each iteration finds $i \in \mathrm{argmin}_i\, y_i\langle \mathbf{w}, \mathbf{x}_i\rangle$ and update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i$. The algorithm mush have $f(\mathbf{w}^{(t)}) < 0$ and after $||\mathbf{w}^*||^2 R^2$ iterations. The algorithm is almost identical to the Batch Perceptron algorithm with two modifications. First, the Batch Perceptron updates with any example for which $y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle \le 0$, while the current algorithm chooses the example for which $y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle$ is minimal. Second, the current algorithm employs the parameter η. However, it is easy to verify that the algorithm would not change if we fix $\eta = 1$ (the only modification is that $\mathbf{w}^{(t)}$ would be scaled by $1/\eta$).

8. AdaBoost. We provide two versions of solution, based on two definitions of the hypothesis class:

$$\mathcal{H}_{\mathrm{DS}} = \{\mathbf{x} \mapsto \mathrm{sign}(\theta - x_i)\cdot b : \theta \in \mathbb{R}, i \in [d], b \in \{1, -1\}\}$$

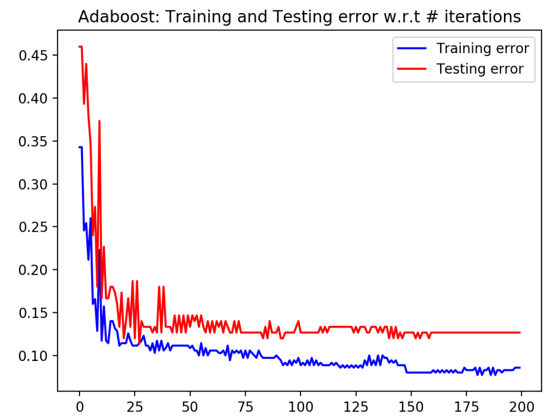
   or

$$\mathcal{H}_{\mathrm{DS}} = \{\mathbf{x} \mapsto \mathrm{sign}(\theta - x_i)\cdot b : \theta \in \mathbb{R}, i \in [d], b \in \{1\}\}$$

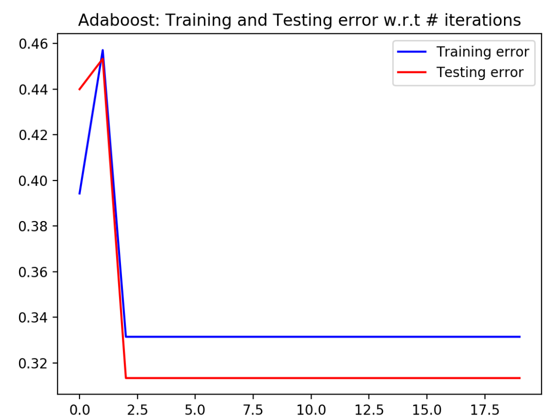
   Answer may vary with implementation.

- The training error of AdaboostClassifier implemented by scikit-learn is 0.1057; The testing error of AdaboostClassifier implemented by scikit-learn is 0.1467
- If $b \in \{1, -1\}$, the training error of a single decision stump with uniform distribution is 0.3429. The testing error is 0.4600.

  If $b \in \{1\}$, the training error of a single decision stump with uniform distribution is 0.3943. The testing error is 0.4400.
- If $b \in \{1, -1\}$, the training error and testing error w.r.t the number of iterations is

  If $b \in \{1\}$, the training error and testing error w.r.t the number of iterations is

- If $b \in \{1, -1\}$, the training error of AdaBoost is 0.0857, the testing error is 0.1267.

  If $b \in \{1\}$ the training error of AdaBoost is 0.3314, the testing error is 0.3133. $\blacksquare$