TTIC 31020: Introduction to Machine Learning

Problem Set 3

Hung Le Tran

20 Jan 2024

Problem 3.1 (Problem 1)

(a) We have chosen the training set S with the assumption that (x_i, y_i)'s are i.i.d. It follows that

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[\mathbb{1}\{\mathcal{F}(S_{-i})(x_i) \neq y_i\}] &= \mathbb{E}_{S_{-i}\sim\mathcal{D}^{m-1}}\left[\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}\left[\mathbb{1}\{\mathcal{F}(S_{-i})(x_i) \neq y_i\}\right]\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^{m-1}}[L_{\mathcal{D}}(F(S))] \\
\Rightarrow \mathbb{E}_{S\sim\mathcal{D}^m}[LOOCV_S(\mathcal{F})] &= \mathbb{E}_{S\sim\mathcal{D}^m}\left[\frac{|\{i : \mathcal{F}(S_{-i})(x_i) \neq y_i\}|}{m}\right] \\
&= \frac{1}{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i=1}^{m}\mathbb{1}\{\mathcal{F}(S_{-i})(x_i) \neq y_i\}\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}[\mathbb{1}\{\mathcal{F}(S_{-i})(x_i) \neq y_i\}] \\
&= \frac{m}{m}\mathbb{E}_{S\sim\mathcal{D}^{m-1}}[L_{\mathcal{D}}(\mathcal{F}(S))] \\
&= \mathbb{E}_{S\sim\mathcal{D}^{m-1}}[L_{\mathcal{D}}(\mathcal{F}(S))]
\end{aligned}$$

as required.

(b) $\mathcal{A}$ enjoys mistake bound M on sequences realized by $\mathcal{H}$, and $S \sim \mathcal{D}^m$ as given here is realized by $\mathcal{H}$. Therefore it can make at most $\max\{T, M\}$ mistakes. It follows that $\tilde{\mathcal{A}}$ will run for at most M iterations, since if it reaches M iterations, $\mathcal{A}$ can no longer make mistakes and the `while` condition exits.

(c) Let N be the number of iterations $\tilde{\mathcal{A}}$ would run on S. This means that after collecting N samples, say S', then $A(S')$ no longer makes mistakes on the remaining $(m+1) - N$ samples. Then, these $(m+1) - N$ samples, when they are validation points, do not contribute to $LOOCV_S(\tilde{\mathcal{A}})$ at all, since S_{-i} (for (x_i, y_i) among those points) would include the N samples, allowing $\tilde{\mathcal{A}}(S_{-i})$ to successfully predict (x_i, y_i). Therefore,

$$LOOCV_S(\tilde{\mathcal{A}}) \leq \frac{N}{m+1}$$

(divided by $(m+1)$ because S has $(m+1)$ samples)

(c) Combining,

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(\tilde{\mathcal{A}}(S))] &= \mathbb{E}_{S\sim\mathcal{D}^{m+1}}[LOOCV_S(\tilde{\mathcal{A}})] \\
&\leq \frac{N}{m+1} \\
&\leq \frac{M}{m+1}
\end{aligned}$$

Therefore to get $< \varepsilon$, we need

$$\frac{M}{m+1} < \varepsilon \Rightarrow m > \frac{M}{\varepsilon} - 1$$

Problem 3.2 (Problem 2)

Part I

(a) WTS
$$\frac{\langle w^o, w_{t+1}\rangle}{\|w^o\|} \geq M_t\gamma$$

Indeed, when $t = 0, 0 \geq 0$.

Suppose that
$$\frac{\langle w^o, w_t\rangle}{\|w^o\|} \geq M_{t-1}\gamma$$

Then if w_{t+1} doesn't update (i.e. did not make mistake on (x_t, y_t)), then M_t does not update too, and the inequality is trivially satisfied. When w_{t+1} does update:
$$\begin{aligned}
\frac{\langle w^o, w_{t+1}\rangle}{\|w^o\|} &= \frac{\langle w^o, w_t\rangle}{\|w^o\|} + \frac{\langle w^o, y_t\varphi(x_t)\rangle}{\|w^o\|} \\
&\geq M_{t-1}\gamma + \gamma \\
&= M_t\gamma
\end{aligned}$$

as required.

(b) WTS
$$\|w_{t+1}\| \leq \sqrt{M_t}$$

The base case is trivial.

Induction hypothesis gives us $\|w_t\| \leq \sqrt{M_{t-1}}$.

Then if w_{t+1} doesn't update then $M_t = M_{t-1}$, the inequality also satisfies.

When it does update, i.e., $y_t\langle w_t, \varphi(x_t)\rangle \leq 0$, then
$$\begin{aligned}
\|w_{t+1}\|^2 &= \|w_t + y_t\varphi(x_t)\|^2 \\
&= \|w_t\|^2 + \|y_t\varphi(x_t)\|^2 + 2\langle w_t, y_t\varphi(x_t)\rangle \\
&\leq M_{t-1} + 1 + 0 = M_t \\
\Rightarrow \|w_{t+1}\| &\leq \sqrt{M_t}
\end{aligned}$$

as required.

(c) Combining both:
$$M_t\gamma \leq \frac{\langle w^o, w_{t+1}\rangle}{\|w^o\|} \leq \|w_{t+1}\| \leq \sqrt{M_t}$$

which implies
$$M_t \leq \frac{1}{\gamma^2}$$

as required.

Part II

(a) By assumption, finite S is realizable by some linear predictor, say, one that corresponds to weight w_0.

Realizability implies that for all $(x_i, y_i) \in S$,

$$y_i\langle w_0, \varphi(x_i)\rangle > 0 \Rightarrow \frac{y_i\langle w_0, \varphi(x_i)\rangle}{\|w_0\|} > 0$$

The minimum of finite positive numbers is positive, so

$$\min_{(x_i,y_i)\in S} \frac{y_i\langle w_0, \varphi(x_i)\rangle}{\|w_0\|} > 0$$

$\gamma(S)$ is then the supremum of a set that contains a positive number (the one above, since w_0 is in the set of possible weights), and is therefore positive.

(b) We know that $M_t \le \frac{1}{\gamma(S)^2}\ \forall\, t$ so $\mathsf{PERCEPTRON}$ has mistake bound $M = \frac{1}{\gamma(S)^2}$.

From question 1, we know that the number of iterations is bounded by $M = \frac{1}{\gamma(S)^2}$.

(c) I would linearly iterate through S to find $(x, y) \in S$ that satisfies $y\langle w, \varphi(x)\rangle \le 0$.

Required runtime per iteration: $O(md)$.

Overall runtime: $O\left(\frac{md}{\gamma(S)^2}\right)$

(d) For this section, let us conventionally denote $\operatorname{sign}(0) = +1$.

Let $S = \{(1, +1), (1, -1)\}$, $d = 1$. We start with $w_1 = 0$.

$$\begin{aligned} w_1 = 0, \operatorname{sign}(\langle w_1, x_2\rangle) = \operatorname{sign}(0) = +1 \ne y_2 \Rightarrow w_2 = 0 + y_2x_2 = -1 \\ w_2 = -1, \operatorname{sign}(\langle w_2, x_1\rangle) = \operatorname{sign}(-1) = -1 \ne y_1 \Rightarrow w_3 = -1 + y_1x_1 = 0 \end{aligned}$$

and we're back to the starting point of the loop. Therefore $\widetilde{\mathsf{PERCEPTRON}}$, iterating $\mathsf{PERCEPTRON}$ repeatedly, will then never terminate.

Part III

(a) Mistake bound:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\widetilde{\mathsf{PERCEPTRON}}(S))\right] \le \frac{M}{m+1} \le \frac{1}{(m+1)\gamma^2}$$

Therefore to ensure expected generalization error at most ε, want:

$$\frac{1}{(m+1)\gamma^2} \le \varepsilon \Rightarrow m \ge \frac{1}{\varepsilon\gamma^2} - 1$$

(b) We made the assumption that $\mathcal{D}$ is separable with margin γ.