CIS 520: Problem Set #5

Due on October 30, 2017

Lyle Ungar, Shivani Agarwal

Eric Oh

Collaborators: Jiarui Lu

Problem 1

Perceptron vs. Winnow

Solution

(a) Note that for this setting of sparse target vector $\mathbf{u}$ and dense feature vectors $\boldsymbol{x_t}$,

- $R_\infty = 1$
- $\|\boldsymbol{u}\|_1 = k$
- $\|\boldsymbol{u}\|_2 = k^{\frac{1}{2}}$
- $\|\boldsymbol{x_t}\|_2 = d^{\frac{1}{2}} = R$

The respective error bounds are then given by

- Winnow : $\sum_{t=1}^{T} I(\hat{y}_t \neq y_t) \leq 2\left(\frac{R_\infty^2 \|\boldsymbol{u}\|_1^2}{\gamma^2}\right)\ln(d) = 2\left(\frac{k^2}{\gamma^2}\right)\ln(d)$
- Perceptron : $\sum_{t=1}^{T} I(\hat{y}_t \neq y_t) \leq \frac{R^2\|\boldsymbol{u}\|_2^2}{\gamma^2} = \frac{dk}{\gamma^2}$

Since $k \ll d$, the d terms dominate in the error bounds. From them, $\ln(d) < d$, meaning that the **Winnow** algorithm has a better error bound.

(b) Note that for this setting of dense target vector $\mathbf{u}$ and sparse feature vectors $\boldsymbol{x_t}$,

- $R_\infty = 1$
- $\|\boldsymbol{u}\|_1 = d$
- $\|\boldsymbol{u}\|_2 \leq 2\sqrt{d}$
- $\|\boldsymbol{x_t}\|_2 = k^{\frac{1}{2}} = R$

The respective error bounds are then given by

- Winnow : $\sum_{t=1}^{T} I(\hat{y}_t \neq y_t) \leq 2\left(\frac{R_\infty^2 \|\boldsymbol{u}\|_1^2}{\gamma^2}\right)\ln(d) = 2\left(\frac{d^2}{\gamma^2}\right)\ln(d)$
- Perceptron : $\sum_{t=1}^{T} I(\hat{y}_t \neq y_t) \leq \frac{R^2\|\boldsymbol{u}\|_2^2}{\gamma^2} \leq \frac{4kd}{\gamma^2}$

Since $k \ll d$, the d terms dominate in the error bounds. From them, $d^2\ln(d) > d$, meaning that that **Perceptron** algorithm has a better error bound.

(c) If the problem has only non-negative features, the Winnow algorithm is not meaningful. This is because the weight updating is a multiplicative factor of exponentials, meaning that the weights will always be positive and since the features are all positive, the Winnow algorithm will always predict a positive label.

Problem 2

Multiclass Boosting

Solution

Problem 2 continued on next page. . .

(a)

$$\begin{aligned}
D_{t+1}(i) &= \frac{D_t(i)\exp(-\alpha_t \widetilde{h}_{t,y_i}(x_i))}{Z_t} \\
&= \frac{D_{t-1}(i)\exp(-\alpha_{t-1}\widetilde{h}_{t-1,y_i}(x_i))\exp(-\alpha_t \widetilde{h}_{t,y_i}(x_i))}{Z_t \times Z_{t-1}} \\
&\vdots \\
&= \frac{D_1(i)\sum_{t=1}^{T}\exp(-\alpha_t \widetilde{h}_{t,y_i}(x_i))}{\prod_{t=1}^{T} Z_t} \\
&= \frac{\frac{1}{m}\exp(-F_{t,y_i}(x_i))}{\prod_{t=1}^{T} Z_t}
\end{aligned}$$

(b)
- If $H(x_i) = y_i$, then it follows that $I(H(x_i) \neq y_i) = 0 \leq I(F_{t,y_t} \leq 0)$ by definition.
- If $H(x_i) \neq y_i$, then it follows that $I(H(x_i) \neq y_i) = 1$. It remains to show that $F_{t,y_i} \leq 0$ must be true. To see this, note that $\sum_{k=1}^{K} F_{t,k}(x_i) = -(K-2)\sum_{t=1}^{T}\alpha_t \leq 0$ due to $\alpha_t \geq 0$ for all t. Then, take any two distinct $1 \leq a, b \leq K$ and consider $F_{t,a}(x_i)$ and $F_{t,b}(x_i)$. Consider the form of $\widetilde{h}_{t,k}(x_i)$. It is a vector that contains $+1$ in the predicted class for x_i and is -1 in all other elements. Thus, $F_{t,a}(x_i) + F_{t,b}(x_i)$ is negative. Then it follows that since the sum over K is negative, both $F_{t,a}(x_i)$ and $F_{t,b}(x_i)$ are negative as well. WLOG, let $y_i = a$, it then follows that $F_{t,y_i} \leq 0$ and the inequality follows.

(c) From (a), note that we can write

$$\begin{aligned}
D_{t+1}\prod_{t=1}^{T} Z_t &= \frac{1}{m}\exp(-F_{t,y_t}(x_i)) \\
\sum_{i=1}^{m} D_{t+1}\prod_{t=1}^{T} Z_t &= \frac{1}{m}\sum_{i=1}^{m}\exp(-F_{t,y_t}(x_i)) \\
\prod_{t=1}^{T} Z_t &= \frac{1}{m}\sum_{i=1}^{m}\exp(-F_{t,y_t}(x_i))
\end{aligned}$$

where the last line follows from the sum of the weights being 1. Then it follows that

$$\begin{aligned}
\mathrm{er}_s[H] &= \frac{1}{m}\sum_{i=1}^{m} I(H(x_i) \neq y_i) \\
&\leq \frac{1}{m}\sum_{i=1}^{m} I(F_{t,y_t}(x_i) \leq 0) \quad \text{(from (b)} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\exp(-F_{t,y_t}(x_i)) \\
&= \prod_{t=1}^{T} Z_t
\end{aligned}$$

(d) Note from the definition of α_t, we can write

$$\exp(-\alpha_t) = \sqrt{\frac{\mathrm{er}_t}{1-\mathrm{er}_t}} \quad \text{and} \quad \exp(-\alpha_t) = \sqrt{\frac{1-\mathrm{er}_t}{\mathrm{er}_t}}$$

From the definition of Z_t, we have

$$\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i)\exp(-\alpha_t \widetilde{h}_{t,y_i}(x_i)) \\
&= \sum_{i=1}^{m} D_t(i)\left[\exp(-\alpha_t)I(h_t(x_i)=y_i)) + \exp(\alpha_t)I(h_t(x_i)\neq y_i))\right] \\
&= \exp(-\alpha_t)\sum_{i=1}^{m} D_t(i)I(h_t(x_i)=y_i) + \exp(\alpha_t)\sum_{i=1}^{m} D_t(i)I(h_t(x_i)\neq y_i) \\
&= \sqrt{\frac{\mathrm{er}_t}{1-\mathrm{er}_t}}(1-\mathrm{er}_t) + \sqrt{\frac{1-\mathrm{er}_t}{\mathrm{er}_t}}(\mathrm{er}_t) \\
&= \sqrt{\mathrm{er}_t(1-\mathrm{er}_t)} + \sqrt{(1-\mathrm{er}_t)\mathrm{er}_t} \\
&= 2\sqrt{\mathrm{er}_t(1-\mathrm{er}_t)}
\end{aligned}$$

(e) Assuming $\mathrm{er}_t \le \frac{1}{2} - \gamma$, we have

$$\begin{aligned}
\mathrm{er}_s[H] &\le \prod_{t=1}^{T} Z_t \quad \text{(from (c)}\\
&= \prod_{t=1}^{T} 2\sqrt{\mathrm{er}_t(1-\mathrm{er}_t)} \quad \text{(from (d)} \\
&\le 2^T \prod_{t=1}^{T} \sqrt{\left(\frac{1}{2}-\gamma\right)\left(\frac{1}{2}+\gamma\right)} \\
&= 2^T \prod_{t=1}^{T} \frac{4}{4}\sqrt{\left(\frac{1}{4}-\gamma\right)^2} \\
&= \prod_{t=1}^{T} \sqrt{1-4\gamma^2} \\
&= (1-4\gamma^2)^{\frac{T}{2}} \\
&\le \exp(-2T\gamma^2)
\end{aligned}$$

where the last line follows from $1 - x \le \exp(-x)$.

Problem 3

Loss-Based Performance Measures

Solution

1. $h(x) = \mathrm{sign}(\hat{\eta}(x) - 0.2) = \left\{ \begin{array}{cc} +1 & \text{for } \hat{\eta}(x) > 0.2 \\ -1 & \text{otherwise} \end{array} \right\}$

2.
   - Expected loss predicting $+1$: $\eta(x)\cdot 0 + (1-\eta(x))\cdot 1 = 1 - \eta(x)$
   - Expected loss predicting -1 : $\eta(x)\cdot 1 + (1-\eta(x))\cdot 0 = \eta(x)$
   - Expected loss abstaining : $\eta(x)\cdot 0.4 + (1-\eta(x))\cdot 0.4 = 0.4$.

   The expected loss for $+1$ is less than that of abstaining when $1-\eta(x) < 0.4$, or $\eta(x) > 0.6$. The expected loss for -1 is less than that of abstaining when $\eta(x) < 0.4$. All other predicted values, we would abstain. Thus, our decision rule is given by $h(x) = \left\{ \begin{array}{cc} +1 & \text{for } \hat{\eta}(x) > 0.6 \\ -1 & \text{for } \hat{\eta}(x) < 0.4 \\ ? & \text{for } 0.4 \leq \hat{\eta}(x) \leq 0.6 \end{array} \right\}$ If the cost of abstaining were to decrease to 0.2, intuitively the likelihood of abstaining would increase. The decision rule is given below and matches our intuition.

$$h(x) = \left\{ \begin{array}{cc} +1 & \text{for } \hat{\eta}(x) > 0.8 \\ -1 & \text{for } \hat{\eta}(x) < 0.2 \\ ? & \text{for } 0.2 \leq \hat{\eta}(x) \leq 0.8 \end{array} \right\}$$

3. **Patient 1**:

   - Expected loss predicting NR : 0.6(0) + 0.3(9) + 0.1(10) = 3.7
   - Expected loss predicting PR : 0.6(4) + 0.3(0) + 0.1(1) = 2.5
   - Expected loss predicting CR : 0.6(5) + 0.3(1) + 0.1(0) = 3.3

   Thus, we would predict PR.
   Patient 2:

   - Expected loss predicting NR : 0.1(0) + 0.3(9) + 0.6(10) = 8.7
   - Expected loss predicting PR : 0.1(4) + 0.3(0) + 0.6(1) = 1
   - Expected loss predicting CR : 0.1(5) + 0.3(1) + 0.6(0) = 0.8

   Thus, we would predict CR.

   Under 0-1 loss, we would predict the label with the highest probability: NR for patient 1 and CR for patient 2.

Problem 4

Learning Theory
Solution

1. (c) - generalization error
2. (b) - test error ; (c) - cross-validation error
3. (a) - obtaining high confidence bounds on generalization error ; (d) - model selection
4. (b) - consistent for some data distributions
5. (b) SVM with RBF kernel and suitably chosen C parameter; (c) Linear logistic regression with L_2 regularization and suitably chosen λ parameter; (d) - Logistic regression with RBF kernel, RKHS regularization, and suitably chosen λ parameter