EE 6180: Advanced Topics in Artificial Intelligence	October 21, 2020
Lecture 13	
Lecturer: Abhishek Sinha	*Scribe: Sourav Sahoo*

1 Boosting as Expert's Problem

Suppose we have access to a weak learning oracle $\mathcal{A}(S, p)$. Let $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be the training set. We also have access to a probability distribution p. A weak-learning oracle gives us the following result:

$$a(\mathcal{A}) = \sum_{i=1}^{N} p(i)\mathbf{1}(h(x_i) = y_i) \geq \frac{1}{2} + \gamma \tag{1}$$

where $a(\mathcal{A})$ is the accuracy of the oracle, $h = \mathcal{A}(S, p)$ be the hypothesis, $h(x_i)$ is the prediction of hypothesis h for x_i and $0 < \gamma \ll 1$. Now we want to design an algorithm $\mathcal{A}'$ such that $a(\mathcal{A}') \geq 1 - \epsilon$, for any $\epsilon > 0$.

We try to tackle the boosting problem from an expert's problem perspective. Assume we have N experts, $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. Let at time step t, the Hedge algorithm gives a probability distribution p_t. Suppose the expert i is said to incur a loss of 1 if $h_t(x_i) = y_i$, i.e, $l_t(i) = \mathbf{1}(h_t(x_i) = y_i)$ and $h_t = \mathcal{A}(S, p_t)$. Let the output of classifier be $H(x) = \text{sign}\left(\sum_{i=t}^{T} h_t(x)\right)$. Let (x_j, y_j) be any expert and $R_T = \mathcal{O}(\sqrt{T \ln N})$ be the total regret. So, we have the following result:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)\mathbf{1}(h_t(x_i) = y_i) &\leq \sum_{t=1}^{T} \mathbf{1}(h_t(x_j) = y_j) + R_T \\
\implies \sum_{t=1}^{T} \frac{1}{2} + \gamma &\overset{(1)}{\leq} \sum_{t=1}^{T} \mathbf{1}(h_t(x_j) = y_j) + R_T \\
\implies \frac{1}{2} + \gamma &\leq \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(h_t(x_j) = y_j) + \frac{R_T}{T}
\end{aligned}$$

As the output of the classifier is made using majority vote, if the first term of the $RHS \geq \frac{1}{2}$, then we can be sure that the algorithm classifier correctly classifies any arbitrarily chosen x_j. To ensure that, we need to show:

$$\begin{aligned}
\gamma &\geq \frac{R_T}{T} \\
\implies \gamma &\geq \frac{C\sqrt{T \ln N}}{T} \\
\implies T &\geq \mathcal{O}\left(\frac{\ln N}{\gamma^2}\right)
\end{aligned}$$

So, if we can run the algorithm for $T \geq \mathcal{O}\left(\frac{\ln N}{\gamma^2}\right)$ rounds, then all the examples can be classified correctly. This result was proven by Freund and Schapire (1997). Multiple variants of boosting such as AdaBoost (Hastie et al., 2009), XGBoost (Chen and Guestrin, 2016) are commonly used now in practice.

2 Follow the Perturbed Leader (FTPL)

We will first describe the set up for this problem. Suppose at time step t, we predict $x_t \in \mathcal{X}$. The adversary outputs $\theta_t \in \mathbb{R}^d$ and $\|\theta_t\|_\infty \leq 1$. Then, the loss incurred is $\langle \theta_t, x_t \rangle$. So, the regret upto time

T is given as:

$$R_T = \sum_{t=1}^{T} \langle x_t, \theta_t \rangle - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle x, \theta_t \rangle \tag{2}$$

The objective is to minimize R_T. The FTPL algorithm is described as follows:

Algorithm 1 Follow the Perturbed Leader

---

procedure FTPL
Input: $\eta > 0, \mathcal{X}$
    $\Theta_1 \leftarrow 0$
    for $t = 1, \ldots, T$ **do**
        sample $\gamma_t \sim \mathcal{N}(\mathbf{0}, I_d)$
        predict: $x_t \leftarrow \operatorname{argmin}_x \langle x, \Theta_t + \eta\gamma_t \rangle$
        $\Theta_{t+1} = \Theta_t + \theta_t$

Lemma 1 (Stein's Lemma)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $g(X)$ be a function that is differentiable almost everywhere and $\mathbb{E}[g(X)] < \infty$, then,*

$$\sigma^2 \mathbb{E}[g'(X)] = \mathbb{E}[(X - \mu) g(X)] \tag{3}$$

Proof.

$$\begin{aligned}
\mathbb{E}[(X-\mu)g(X)] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x)(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \left[-g(x)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(x)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \right] \\
&= \frac{1}{\sigma\sqrt{2\pi}} \sigma^2 \int_{-\infty}^{\infty} g'(x)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \sigma^2 \mathbb{E}[g'(X)]
\end{aligned}$$

■

References

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.