Mathematical Foundations of Machine Learning (CS 4783/5783)

Lecture 14: Boosting and Online Learning

Boosting is one of the most widely (in both theory and practice) approaches in machine learning. Even in the deep learning era, boosting based algorithms still reign supreme for a large number of problems in practice (see kaggle competitions). On the theory side, booting results are used as a tool for proving various results in varied set of topics beyond ML. In this lecture we will see the problem of boosting for binary classification and will solve it using online learning results as a somewhat sort of black-box.

1 Weak Learners and Boosting

We will first consider the training error question in boosting, then in later section look into its generalization. Boosting is the problem of using the notion of weak learning and boost the performance of weak learners to a strong learner. What does this mean? So here is the problem.

Given a binary classification training dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ we will assume access to a black box weak learning algorithm $\mathcal{A}$ that takes the sample S and some distribution over the n points in the sample and comes with the following guarantee.

Assumption 1 (γ-weak learners)**.** *Given a sample S of size n, an algorithm $\mathcal{A}$ is said to be a weak learner if for any distribution $q \in \Delta([n])$, the output $h = \mathcal{A}(S, q)$ satisfies:*

$$\mathbb{E}_{i \sim q}\left[\mathbf{1}\{h(x_i) \neq y_i\}\right] \leq \frac{1}{2} - \gamma$$

That is, for the sample S, the weak learner for any distribution on the sample points, can output a classifier whose accuracy is γ better than random guessing. Typically, the weak learning algorithm $\mathcal{A}$ will be a simple classifier, line a half-space or a low depth decision tree etc. Also note that typically learning algorithms only take as input samples, not sample and a distribution over the samples. However, given sample S and distribution q, we can feed a typical learning algorithm with points in S sampled according to q. The weak learning guarantee just translates (well almost) to finding a weak learner on the sample the algorithm received, that is a hypothesis that barely does better than random guessing on the sample it received.

Now the question of boosting is the following:

For sample S, assume there exists a weak learner $\mathcal{A}$. In this case, is it true that there exists a strong classifier whose classification error on S is 0?

We will show the the answer to the above question is a yes and that we can in fact, with a rather small number of calls to weak learner obtain such a classifier. The classic result of boosting is due to Yoav Freund and Robert Schapire. They presented the Adaboost algorithm in their seminal work "A decision-theoretic generalization of on-line learning and an application to boosting" in 1997. In this lecture notes however, we will show a different algorithm for boosting by showing a tight

2 Boosting Via Experts Algorithm

In this section we will answer the boosting question in the affirmative by showing an interesting connection to Online learning, especially the experts problem. The high level idea is this: we think of each sample point as an expert and will use an online learning algorithm to pick distributions over the sample points so as to minimize regret. We will think of the loss on each sample point as the reward for that expert. So the experts algorithm is trying to maximize rewards or in other words trying to pick distributions over sample points that maximize the loss. As an adversary for the experts problem we will pick the weak learner that will in turn try to (in a weak sense) do reasonably well for the distributions presented by the experts algorithm. In this case, we will show how the regret guarantee of the experts algorithm will give us strong learning. Here is the black-box boosting algorithm based on some online learning algorithm for teh experts problem.

```
Input: Training set $S = (x_1, y_1), \dots, (x_n, y_n)$, weak learner $\mathcal{A}$, experts algorithm over $n$ experts. $\eta = \sqrt{2\log(n)/T}$

For $t = 1$ to $T$
    1. For every $i \in [n]$, $q_t(i) = \frac{\exp(-\eta \sum_{s=1}^{t-1} \ell_t(i))}{\sum_{i=1}^{n} \exp(-\eta \sum_{s=1}^{t-1} \ell_t(i))}$ .
    2. Use weak-learner to obtain $h_t = \mathcal{A}(S, q_t)$
    3. For expert $i$, define $\ell_t(i) = 1 - \mathbf{1}\{h_t(x_i) \neq y_i\}$
End For

Return hypothesis $\hat{y}_{\text{boost},T}(x) = \text{sign}\left(\sum_{t=1}^{T} h_t(x)\right)$
```

In the above, we are turning the online learning algorithm of exponential weights on its heals. We are treating each sample point x_t, y_t as an expert. So we have n experts for the n samples. q_t is the distribution returned by the experts algorithm at round t which is a distribution over the n samples points. Now let us see what weak learning + regret bound for experts problem gives us:

Theorem 2. *Let $\mathcal{A}$ was a γ weak learner for the sample S. Then after $T \geq \frac{2\log n}{\gamma^2}$ we have that:*

$$\forall i \in [n], \hat{y}_{\text{boost},T}(x_i) = y_i$$

That is, after the prescribed number of iterations, our training error is 0.

Proof. By regret bound of exponential weights algorithm, we have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{i \sim q_t}\left[\ell_t(i)\right] \leq \min_{i \in [n]} \frac{1}{T}\sum_{t=1}^{T} \ell_t[i] + \sqrt{\frac{2\log n}{T}}$$

However, $\ell_t[i] = 1 - \mathbf{1}\{h_t(x_i) \neq y_i\}$ and so:

$$\begin{aligned}
-\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{n} q_t[j]\mathbf{1}\{h_t(x_j) \neq y_j\} &\leq \min_{i \in [n]} -\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{h_t(x_i) \neq y_i\} + \sqrt{\frac{2\log n}{T}} \\
&= -\max_{i \in [n]} \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{h_t(x_i) \neq y_i\} + \sqrt{\frac{2\log n}{T}}
\end{aligned}$$

Rearranging we get that:

$$\begin{aligned}\max_{i\in[n]}\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{h_t(x_i)\neq y_i\} &\le \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{n} q_t[j]\mathbf{1}\{h_t(x_j)\neq y_j\}+\sqrt{\frac{2\log n}{T}}\\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{j\sim q_t}\left[\mathbf{1}\{h_t(x_j)\neq y_j\}\right]+\sqrt{\frac{2\log n}{T}}\\ &= \frac{1}{2}-\gamma+\sqrt{\frac{2\log n}{T}}\end{aligned}$$

Now if $T \ge \frac{2\log n}{\gamma^2}$ then$-\gamma+\sqrt{\frac{2\log n}{T}} < 0$ and so,

$$\max_{i\in[n]}\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{h_t(x_i)\neq y_i\} < \frac{1}{2}$$

In other words, for any sample i in our set of n samples, the average zero-one error of the classifier $h_1,\ldots,h_T$ is strictly smaller than 1/2. This means that at least half of the classifiers are right! Hence taking majority would always give the correct classifier. Hence,

$$\max_{i\in[n]}\mathbf{1}\{\hat{y}_{\mathrm{boost},T}(x_i)\neq y_i\}=0$$

This proves the theorem. □

The above result is remarkable because it says that existence of weak learning where we need to barely surpass random guess implies we can get a classifier whose training error is in fact 0. The result of course also suggests that weak learning assumption is not as mild as it seems because it says that existence of weak learner implies we can easily find a perfect classifier. In fact, weak learning assumption as it turns out implies having a perfect classifier with margin in certain sense.

3 Generalization Analysis for Boosting

In the previous section we only talked about training error. But in reality we are of course interested in test error. To think about test error its useful to think about which set our base weak learners come from. Assume the weak learning algorithm picks hypothesis from a base set of simple hypothesis $\mathcal{H}\subset\mathcal{Y}^{\mathcal{X}}$. While we can easily lift this analysis to infinite $\mathcal{H}$ using VC theory, for simplicity, say $\mathcal{H}$ has finite cardinality. Now if we ran boosting for T rounds, what is cardinality of the model class $\hat{y}_{\mathrm{boost},T}$ belongs to?

Well our boosting algorithm picks each $h_t\in\mathcal{H}$ and we pick $h_1,\ldots,h_T$ each based on sample from $\mathcal{H}$. Hence the cardinality of the model class to which $\hat{y}_{\mathrm{boost},T}$ belongs to is at most $|\mathcal{H}|^T$. Assuming $T > T_0$, we get the bound:

$$\mathbb{E}_S\left[L_D(\hat{y}_{\mathrm{boost},T})\right]\le\sqrt{\frac{T\log|\mathcal{H}|}{n}}$$

Now if we had used the experts algorithm, and we pick $T=T_0=\frac{\log n}{\gamma^2}$ then we conclude that

$$\mathbb{E}_S\left[L_D(\hat{y}_{\mathrm{boost},T_0})\right]\le\sqrt{\frac{\log(n)\log|\mathcal{H}|}{\gamma^2 n}}$$

Thus, if weak learning assumption is satisfied for every sample S drawn from distribution w.r.t. some hypothesis class $\mathcal{H}$, then we obtain the above bound on test classification error of the boosting algorithm. In fact, we can get the above bound in high probability whenever weak learning assumption holds in high probability over draw of samples S w.r.t. hypothesis $\mathcal{H}$. Of course, if $\mathcal{H}$ has VC dimension d then we can replace $|\mathcal{H}|$ by $(n/d)^d$ and obtain bound of form $\frac{\log n}{\gamma}\sqrt{\frac{d}{n}}$ on test loss of boosting algorithm.