CSE522, Winter 2011, Learning Theory Lecture 13 - 03/08/2011

Algorithmic Stability

Lecturer: Ofer Dekel *Scribe: Thach Nguyen*

The goal of this lecture is to establish risk bounds that depend on the learning algorithm A instead of the hypothesis class $\mathcal{H}$. In particular, we would prove results that look like

With probability at least $1-\delta$ *over the sample set* $S \sim \mathcal{D}^m$,

$$|\ell\left(A(S); \mathcal{D}\right) - \ell\left(A(S); S\right)| \leq \epsilon$$

i.e. we compare the performance of the same function, $A(S)$, on two different distributions, the uniform distribution on the sample set S and the real distribution $\mathcal{D}$. Contrast this to the approach we took in the Rademacher and VC theory, where we compared the performance of two functions, the output of the algorithm $A(S)$ and the best hypothesis h^*, on the same distribution $\mathcal{D}$.

Notice two simple facts. First, $A(S)$ is a random variable, with the randomness comes from S. Second,$A(S)$ depends on S, so we cannot use Hoefding's inequality to compare $\ell\left(A(S); \mathcal{D}\right)$ and $\ell\left(A(S); S\right)$.

1 Uniform stability

Before getting to the formal definitions, we introduce some notations

- $S^{\setminus i} = S \setminus \{(x_i, y_i)\}$, i.e. the set of $m-1$ samples where the ith sample is removed.
- $S^i = S^{\setminus i} \cup \{(x'_i, y'_i)\}$ for some *worst* (x'_i, y'_i).

Definition 1 (Uniform stabiliy). *Algorithm A has* uniform stability β *with respect to a loss function ℓ if for all S and all $i \in [m]$,*

$$\max_{x,y} \left| \ell\left(A(S); (x,y)\right) - \ell\left(A(S^{\setminus i}); (x,y)\right) \right| \leq \beta$$

i.e. the algorithm is "stable" with respect to removing a single sample at all points.

Note that β depends on m and we would want β_m to be around $\frac{1}{m}$.

Remark 1. *There are weaker notions of stability such as*

- ***Error stability:*** *For all S, for all $i \in [m]$: $|\ell(A(S); \mathcal{D}) - \ell(A(S^{\setminus i}); \mathcal{D})| \leq \beta$, i.e. the average difference is small. This is a very week notion of stability.*
- ***Hypothesis stability:*** *For all $i \in [m]$, $\mathbf{E}_{S,(x,y)}\left[\left|\ell\left(A(S); (x,y)\right) - \ell(A(S^{\setminus i}); (x,y))\right|\right] \leq \beta$*

Remark 2. *By triangle inequality, uniform stability β implies that for S, for all $i \in [m]$*

$$\max_{x,y} \left| \ell(A(S); (x,y)) - \ell(A(S^i); (x,y)) \right| \leq 2\beta$$

Now we show that uniform stability implies a bound in the form stated in the first paragraph.

Theorem 2. *Suppose that $\ell \in [0, C]$ and A has uniform stability β. Furthermore, suppose that A is* anonymous, *i.e. $A(S) = A(S')$ if S and S' contains the same elements (but in different orders). Let Z be the random variable defined by $Z = \ell\left(A(S); \mathcal{D}\right) - \ell(A(S); S)$. Then w.p. at least $1-\delta$,*

$$Z \leq 2\beta + m(4\beta + C/m)\sqrt{\frac{\log(1/\delta)}{2m}}$$

Proof. We will show that Z is concentrated, and then show that $\mathbf{E}[Z]$ is small. For the first part, note that by triangle inequality:

$$\left|\ell\left(A(S);\mathcal{D}\right)-\ell\left(A(S^i);\mathcal{D}\right)\right| \leq \left|\ell\left(A(S);\mathcal{D}\right)-\ell\left(A(S^{\setminus i});\mathcal{D}\right)\right| + \left|\ell\left(A(S^{\setminus i};\mathcal{D}\right)-\ell\left(A(S^i);\mathcal{D}\right))\right| \leq 2\beta$$

Also:

$$\begin{aligned}\left|\ell\left(A(S);S\right)-\ell\left(A(S^i);S^i\right)\right| &\leq \frac{1}{m}\sum_{j\neq i}\left|\ell\left(A(S);(x_j,y_j)\right)-\ell\left((;A)(S^i);(x_j,y_j)\right)\right| \\ &\quad+\frac{1}{m}\left|\ell(A(S);(x_i,y_i))-\ell(A(S^i);(x'_i,y'_i))\right| \\ &\leq \frac{m-1}{m}2\beta+\frac{C}{m}\leq 2\beta+\frac{C}{m}\end{aligned}$$

Hence, $\left|Z-Z^i\right| \leq 4\beta+\frac{C}{m}$ where Z^i denotes the random variable where S is replaced by S^i. This implies that $\max_S Z-\min_S Z \leq m(4\beta+C/m)$; therefore $Z \leq \mathbf{E}[Z]+m(4\beta+C/m)$.

To bound $\mathbf{E}[Z]$, we will need some identities

- $\mathbf{E}_S\left[\ell\left(A(S);(x_j,y_j)\right)\right] = \mathbf{E}_{S,(x',y')}\left[\ell\left(A(S^j);(x',y')\right)\right]$. This identity holds because for (x',y') drawn from $\mathcal{D}$, $\mathbf{Pr}[S]=\mathbf{Pr}\left[S^j\right]$.
- $\mathbf{E}_{S,(x',y')}\left[\ell\left(A(S^j);(x',y')\right)\right] = \mathbf{E}_{S,(x',y')}\left[\ell\left(A(S^i);(x',y')\right)\right]$. To see that this identity holds, let S' be the set that contains the same elements as S but with (x_i,y_i) and (x_j,y_j) exchange their order. Then since A is anonymous, $A(S)=A(S')$. The identity then forllow from the fact that $\mathbf{Pr}[S]=\mathbf{Pr}[S']$ and the previous identity.

With these identities, we have:

$$\begin{aligned}\mathbf{E}[Z] &= \mathbf{E}_S\left[\ell\left(A(S);\mathcal{D}\right)-\ell(A(S);S)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))\right]-\mathbf{E}_S\left[\ell\left(A(S);S\right)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))\right]-\mathbf{E}_S\left[\frac{1}{m}\sum_{j=1}^{m}\ell\left(A(S);(x_j,y_j)\right)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))\right]-\frac{1}{m}\sum_{j=1}^{m}\mathbf{E}_S\left[\ell\left(A(S);(x_j,y_j)\right)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))\right]-\frac{1}{m}\sum_{j=1}^{m}\mathbf{E}_{S,(x',y')}\left[\ell\left(A(S^j);(x',y')\right)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))\right]-\frac{1}{m}m\mathbf{E}_{S,(x',y')}\left[\ell\left(A(S^i);(x',y')\right)\right] \\ &= \mathbf{E}_{S,(x',y')}\left[\ell(A(S);(x',y'))-\ell\left(A(S^i);(x',y')\right)\right] \\ &\leq 2\beta\end{aligned}$$

Now, by McDiamid's inequality, we have that with probability at least $1-\delta$,

$$Z \leq \mathbf{E}[Z]+m(4\beta+c/m)\sqrt{\frac{\log(1/\delta)}{m}} \leq 2\beta+m(4\beta+c/m)\sqrt{\frac{\log(1/\delta)}{m}}.$$

□

The implication of this theorem is that if A is uniformly stable and has good empirical risk, we have good bound on A's risk even if the hypothesis class $\mathcal{H}$ is bad.

2 Regularized ERM

We give an example algorithm that has uniform stability. Consider the case where $y \in \{-1, 1\}$, is convex and differentiable. For simplicity, assume that $\ell(h; (x, y)) = \ell(yh(x))$, so ℓ is λ-lipschitz. (log loss is a nice loss here; hinge loss is not differentiable, but there's a fix for that.) The algorithm we will use is Regularized ERM, which outputs the hypothesis that minimize the following quantity:

$$R(h) = \ell(h; S) + c\psi(h)$$

where $\psi(h)$ is some convex and differentiable function from $\mathcal{H}$ to $\mathbb{R}$ and c is some constant.

Consider the following quantities

- $\bar{h} = \operatorname{argmin}_h R(h)$
- $R^{\setminus i}(h) = \ell\left(h; S^{\setminus i}\right) + c\psi(h)$
- $\bar{h}^{\setminus i} = \operatorname{argmin}_h R^{\setminus i}(h)$

Note that $\bar{h}^{\setminus i}$ is the output of the algorithm on $S^{\setminus i}$. Therefore $\beta = \max_{x,y} \left|\ell\left(\bar{h}; (x, y)\right) - \ell\left(\bar{h}^{\setminus i}; (x, y)\right)\right|$ is exactly the quantity we want to bound. We have:

$$\beta = \max_{x,y} \left|\ell\left(\bar{h}; (x, y)\right) - \ell\left(\bar{h}^{\setminus i}; (x, y)\right)\right| \leq \max_{x,y} \lambda \left|y\bar{h}(x) - y\bar{h}^{\setminus i}(x)\right| = \max_{x} \lambda \left|\bar{h}(x) - \bar{h}^{\setminus i}(x)\right|. \tag{1}$$

Now we consider a concrete hypothesis class $\mathcal{H} = \{w \in \mathbb{R}^n\}$ and domain $\mathcal{X} = \{x \in R^n : \|x\| \leq X\}$. Let $\psi(w) = \|w\|^2$.

To prove a concrete bound on β, we will use the Bregman divergence defined as follows.

Definition 3. *The Bregman divergence of a function f is defined by $B_f(v'\|v) = f(v') - f(v) - \langle v' - v, \nabla f(v)\rangle$.*

Lemma 4. *If f is convex then $B_f \geq 0$.*

Lemma 5. $cB_\psi(\bar{h}^{\setminus i}\|\bar{h}) + cB_\psi(\bar{h}\|\bar{h}^{\setminus i}) \leq \frac{2\lambda}{m} \max_x \left|\bar{h}^{\setminus i}(x) - \bar{h}(x)\right|$.

Proof. We have:

$$\begin{aligned}
cB_\psi(\bar{h}^{\setminus i}\|\bar{h}) + cB_\psi(\bar{h}\|\bar{h}^{\setminus i}) &\leq cB_R(\bar{h}^{\setminus i}\|\bar{h}) + cB_{R^{\setminus i}}(\bar{h}\|\bar{h}^{\setminus i}) \\
&= R(\bar{h}^{\setminus i}) - R(\bar{h}) + R^{\setminus i}(\bar{h}) - R^{\setminus i}(\bar{h}^{\setminus i}) \\
&= \frac{1}{m}\left(\ell\left(\bar{h}^{\setminus i}; (x_i, y_i)\right) - \ell\left(\bar{h}; (x_i, y_i)\right)\right) + \frac{1}{m(m-1)} \sum_{j \neq i} \left(\ell\left(\bar{h}; (x_j, y_j)\right) - \ell\left(\bar{h}^{\setminus i}; (x_j, y_j)\right)\right) \\
&\leq \frac{\lambda}{m}\left|\bar{h}^{\setminus i}(x_i) - \bar{h}(x_i)\right| + \frac{\lambda}{m(m-1)} \sum_{j \neq i} \left|\bar{h}(x_j) - \bar{h}^{\setminus i}(x_j)\right| \\
&\leq \frac{\lambda}{m} \max_x \left|\bar{h}^{\setminus i}(x) - \bar{h}(x)\right| + \frac{\lambda}{m(m-1)}(m-1) \max_x \left|\bar{h}(x) - \bar{h}^{\setminus i}(x)\right| \\
&\leq \frac{2\lambda}{m} \max_x \left|\bar{h}(x) - \bar{h}^{\setminus i}(x)\right|
\end{aligned}$$

where the first inequality forllows from the facts that $B_{f+g} = B_f + B_g$, that ℓ is convex and Lemma 4. □

Now note that for our definition of ψ, $B_\psi(w\|w') = \|w - w'\|^2$ is symmetric. Thus, the lemma implies

$$2c\|\bar{w}^{\setminus i} - \bar{w}\|^2 \leq \frac{2\lambda}{m}\left|\bar{h}^{\setminus i}(x_i) - \bar{h}(x_i)\right| \leq \frac{2\lambda}{m} X \|\bar{w}^{\setminus i} - \bar{w}\|$$

Hence, we have

$$\|\bar{w}^{\setminus i} - \bar{w}\| \leq \frac{\lambda X}{mc}$$

By (1), we have

$$\beta \leq \max_x \lambda \left| \bar{h}(x) - \bar{h}^{\setminus i}(x) \right| \leq \lambda X \|\bar{w} - \bar{w}^{\setminus i}\| \leq \lambda X \cdot \frac{\lambda X}{mc} = \frac{\lambda^2 X^2}{mc} = O\left(\frac{1}{m}\right) \tag{2}$$

This shows that the regularized ERM algorithm has uniform convergence for this setting. Choosing an appropriate c so as to make sure that the algorithm still has good empirical loss yields a good bound on its risk.