9.520: Statistical Learning Theory and Applications February 22nd, 2010

Generalization Bounds, Intro to Stability

Lecturer: Lorenzo Rosasco *Scribes: J. Kaeli and J. Wiens*

The goal of today's class is to revisit the concept of **generalization bounds** and derive them using a measure of **stability**. We will do this using **concentration inequalities**.

1 Generalization Bounds

A learning algorithm maps a training set $S = \{(x_1, y_1), ..., (x_n, y_n)\} = \{z_1, ..., z_n\}$, where $z = (x, y)$, to a function f_S. It is assumed that this mapping, denoted by $\mathcal{A}$, is both deterministic (the same training set will always result in the same mapping) and independent of the ordering of the points in the training set. The quality of the mapping can be measured using a loss function $V(f_S(x), y) = V(f_S, z)$. We define the expected risk $I[f_S]$ and the empirical risk $I_S[f_S]$ to be:

$$I[f_S] = \mathbb{E}_z \left[V(f_S, z)\right] = \int V(f_S, z)\, d\mu(z)$$

$$I_S[f_S] = \frac{1}{n} \sum_{i=1}^{n} V(f_S, z_i).$$

Ideally, we would like to choose $\mathcal{A}$ so that the expected risk $I[f_S]$ is small. While we can measure the empirical risk $I_S[f_S]$, we cannot measure $I[f_S]$ because it depends on the unknown probability distribution μ. A generalization bound is a (probabilistic) bound on the defect (generalization error),

$$D[f_S] = I[f_S] - I_S[f_S].$$

In the second lecture, the notion of generalization was introduced. We saw that a learning algorithm $\mathcal{A}$ generalizes if for any unknown probability distribution μ, the defect converges to zero (in probability) as the number of training points approaches infinity. This means $I_S[f_S]$ can be used as a "proxy" for $I[f_S]$: if the defect can be bounded and $I_S[f_S]$ can be observed to be small, then with high probability $I[f_S]$ is also small. A probabilistic bound takes the form

$$\mathbb{P}(D[f_S] \geq \epsilon) \leq \delta$$

where both ϵ and δ will go to zero as n goes to infinity. Relating the error ϵ, confidence δ, and number of training points n allows us to say something about the quality our learning algorithm in a probabilistic sense.

Historically, the necessary and sufficient conditions for learning have relied upon finite complexity (justifying that a property holds if the size of the space it occupies can be controlled) and uniform Glivenko-Cantelli classes (measurable functions of i.i.d. variables whose empirical measures converge to their true values) to uphold generalization and consistency [1]. However, suitable notions of stability in learning algorithms turn out to be also a necessary and sufficient condition for learning [2].

2 Stability

Stability, in the context of learning algorithms, means that the function f_S should depend continuously on the training data S. This means that a small perturbation on the training set S should

induce only a small change in the solution function f_S. Typically, we expect that this dependence should decrease with the size of S. We can define, for a training set S, the new training set $S^{i,z}$ which is obtained by replacing the i^{th} point in S with a new point z. An algorithm has uniform stability (or, is β-stable) if

$$\forall (S,z) \in \mathcal{Z}^{n+1},\ \forall i,\ \sup_{z' \in Z} |V(f_S, z') - V(f_{S^{i,z}}, z')| \leq \beta.$$

In words, for all training sets S, and all training points in S (not including z'), the worst possible difference (using all possible z') between the loss function evaluated at z' for the function obtained using training set S and the loss function evaluated at z' for the function obtained using the new training set $S^{i,z}$ will be at most β. This β is typically itself a function of n [2].

This is a strong requirement, implying that the function f_S must be very insensitive to changes in the training set $S^{i,z}$ even when an unlikely or "bad" training set is drawn. This notion is stronger than the one discussed in the second class; yet it is still satisfied by Tikhonov regularization, as we shall see in the next lecture. Once the stability is characterized by β, relating this to the bounds on its performance can be accomplished using concentration inequalities.

3 Concentration Inequalities

The law of large numbers states that the sums of independent random variables have a high probabilty of being near the expected value of the sums. This is true for a relatively large class of functions of independent random variables as well. Concentration inequalities are a way to represent how these functions and variables are distributed around their expectations [3]. In particular, McDiarmid's inequality is a useful formulation.

Let $V_1, \ldots, V_n$ be random variables. If a function F mapping $V_1, \ldots, V_n$ to $\mathbb{R}$ satisfies

$$\sup_{v_1,\ldots,v_n,v_i'} |F(v_1, \ldots, v_n) - F(v_1, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_n)| \leq c_i,$$

then the following statement holds:

$$\mathbb{P}\left(|F(v_1, \ldots, v_n) - \mathbb{E}(F(v_1, \ldots, v_n))| > \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

In words, if the largest difference (over all v_i) between a function operating on a set of random variables and the same function operating on the same set of random variables with one variable changed is less than or equal to some constant c_i, then the probabibility that that same function operating on the same set of random variables will differ from its expectation by more than ϵ is an exponential function of ϵ and c_i.

The above inequality can be used to derive Hoeffding's Inequality. Suppose each $v_i \in [a, b]$, and we define $F(v_1, \ldots, v_n) = \frac{1}{n}\sum_{i=1}^n v_i$, the average of the v_i. Then, $c_i = \frac{1}{n}(b-a)$. Applying McDiarmid's Inequality, we have that

$$\begin{aligned}
\mathbb{P}\left(|F(\mathbf{v}) - \mathbb{E}(F(\mathbf{v}))| > \epsilon\right) &\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (\frac{1}{n}(b-a))^2}\right) \\
&= 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).
\end{aligned}$$

4 Generalization Bounds via McDiarmid's Inequality

As previously stated, our main goal here is to bound the difference between empirical and expected error for a β-stable algorithm $\mathcal{A}$. In other words, for any $\epsilon \geq 0$ our goal is to bound the term:

$$\mathbb{P}(|I_S[f_S] - I[f_S]| > \epsilon)$$

To formulate this bound, we will apply McDiarmid's inequality to the random variable

$$D[f_S] = I[f_S] - I_S[f_S].$$

In order to do so we will first need to bound:

1. how much the random variable $D[f_S]$ can change when changing one example, and
2. the expectation of the random variable $D[f_S]$.

Bounding the Expectation of the Defect

Given a β-stable algorithm $\mathcal{A}$, the expected value of the defect is as follows:

$$\begin{aligned}
\mathbb{E}_S D[f_S] &= \mathbb{E}_S \left[I_S[f_S] - I[f_S] \right] \\
&= \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^{n} V(f_S, z_i) - \int V(f_S, z) d\mu(z) \right] \\
&= \mathbb{E}_S \left[\int \Big(\frac{1}{n} \sum_{i=1}^{n} V(f_S, z_i) - V(f_S, z) \Big) d\mu(z) \right] \\
&= \mathbb{E}_{(S,z)} \left[\frac{1}{n} \sum_{i=1}^{n} V(f_S, z_i) - V(f_S, z) \right] \\
&= \mathbb{E}_{(S,z)} \left[\frac{1}{n} \sum_{i=1}^{n} V(f_{S^{i,z}}, z) - V(f_S, z) \right] \\
&= \mathbb{E}_{(S,z)} \left[\frac{1}{n} \sum_{i=1}^{n} \Big(V(f_{S^{i,z}}, z) - V(f_S, z) \Big) \right] \\
&\leq \mathbb{E}_{(S,z)} \left[\frac{1}{n} \sum_{i=1}^{n} \beta \right] \\
&\leq \beta
\end{aligned}$$

The fifth equality follows by the symmetry of the expectation. The expected value of a training set on a training point doesn't change when we rename the points. Thus, by renaming the i^{th} point in the training set S to z we can replace z_i with z, such that the two loss functions are with respect to the same variable z. This renaming allows us to apply McDiarmid's inequality directly.

Bounding the Deviation of the Defect

Let $\mathcal{A}$ a β-stable learning algorithm with respect to a loss function V where the loss function V is bounded, i.e. for all $x \in \mathcal{X}$, $V(f_S, x) \leq M$ for some $M \geq 0$, then:

$$\begin{aligned}
|D[f_S] - D[f_{S^{i,z}}]| &= \left|I_S[f_S] - I[f_S] - I_{S^{i,z}}[f_{S^{i,z}}] + I[f_{S^{i,z}}]\right| \\
&\leq \left|I[f_S] - I[f_{S^{i,z}}]\right| + \left|I_S[f_S] - I_{S^{i,z}}[f_{S^{i,z}}]\right| \\
&\leq \beta + \frac{1}{n}\left|V(f_S, z_i) - V(f_{S^{i,z}}, z)\right| + \frac{1}{n}\sum_{j \neq i}\left|V(f_S, z_j) - V(f_{S^{i,z}}, z_j)\right| \\
&\leq \beta + \frac{M}{n} + \beta \\
&= 2\beta + \frac{M}{n}
\end{aligned}$$

Based on the above results for the bound on the deviation of the defect we can now apply McDiarmid's inequality: for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|D[f_S] - \mathbb{E}D[f_S]\right| > \epsilon\right) &\leq 2\exp\left(-\frac{2\epsilon^2}{n(2(\beta + \frac{M}{n}))^2}\right) \\
&= 2\exp\left(-\frac{n\epsilon^2}{2(n\beta + M)^2}\right).
\end{aligned}$$

Now, recalling the definition of the defect, $D[f_S] = I[f_S] - I_S[f_S]$, and the bound on the expectation of the defect, we have,

$$\mathbb{P}\left(\left|I[f_S] - I_S[f_S] - \beta\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2(n\beta + M)^2}\right)$$

Letting

$$\delta = 2\exp\left(-\frac{n\epsilon^2}{2(n\beta + M)^2}\right),$$

then, with confidence $1 - \delta$, we have the bound

$$I[f_S] \leq I_S[f_S] + \beta + (\beta n + M)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

5 Convergence

Letting $\beta = \frac{k}{n}$ for some constant k, then with probability $1 - \delta$:

$$I[f_S] \leq I_S[f_S] + \frac{k}{n} + (2k + M)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Notice that the error term on the right hand side is $O\left(\frac{1}{\sqrt{n}}\right)$. Here we see that $\beta = \frac{k}{n}$ is good enough, since once $\beta = O(\frac{1}{n})$, further increases in stability don't effect the rate of convergence. Even when $\beta = 0$ the convergence is still $O\left(\frac{1}{\sqrt{n}}\right)$.

References

[1] R. M. Dudley, E. Gine, and J. Zinn, *Uniform and Universal Glivenko-Cantelli Classes*, Journal of Theoretical Probability, Vol. 4, No. 3, 1991.

[2] O. Bousquet and A. Elisseeff, Stability and Generalization, Journal of Machine Learning Research, Vol. 2, pp.499-526, 2002.

[3] Boucheron, S., G. Lugosi and O. Bousquet: *Concentration Inequalities*. Advanced Lectures on Machine Learning Lecture Notes in Artificial Intelligence 3176, 208-240. (Eds.) Bousquet, O., U. von Luxburg and G. Rätsch, Springer, Heidelberg, Germany. 2004.

[4] Ashish Rastogi, McDiarmid's Inequality, New York University, Computer Science, Machine Learning 2008.