Generalization Bounds and Stability

Lorenzo Rosasco Tomaso Poggio

9.520 Class 6

February, 23 2011

About this class

Goal To recall the notion of generalization bounds and show how they can be derived from a stability argument.

Plan

- Generalization Bounds
- Stability
- Generalization Bounds Using Stability

Learning Algorithms

A learning algorithm $\mathcal{A}$ is a map

$$S \mapsto f_S$$

where $S = (x_1, y_1). \ldots (x_n, y_n)$.

We assume that:

- $\mathcal{A}$ is deterministic,
- $\mathcal{A}$ does not depend on the ordering of the points in the training set.

How can we measure quality of f_S?

Error Risks

Recall that we've defined the expected risk:

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

and the empirical risk:

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

Note: we will denote the loss function as $V(f, z)$ or as $V(f(x), y)$, where $z = (x, y)$. For example:

$$\mathbb{E}_z\left[V(f, z)\right] = \mathbb{E}_{(x,y)}\left[V(f_S(x), y)\right]$$

Generalization Bounds

Goal

Choose $\mathcal{A}$ so that $I[f_S]$ is small $\Longrightarrow$ $I[f_S]$ depends on the unknown probability distribution.

Approach

We can measure $I_S[f_S]$. A **generalization bound** is a (probabilistic) bound on the defect (generalization error)

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

If we can bound the defect and we can observe that $I_S[f_S]$ is small, then $I[f_S]$ is likely to be small.

Properties of Generalization Bounds

A probabilistic bound takes the form

$$\mathbb{P}(I[f_S] - I_S[f_S] \geq \epsilon) \leq \delta$$

or equivalenty with confidence $1 - \delta$

$$I[f_S] - I_S[f_S] \leq \epsilon$$

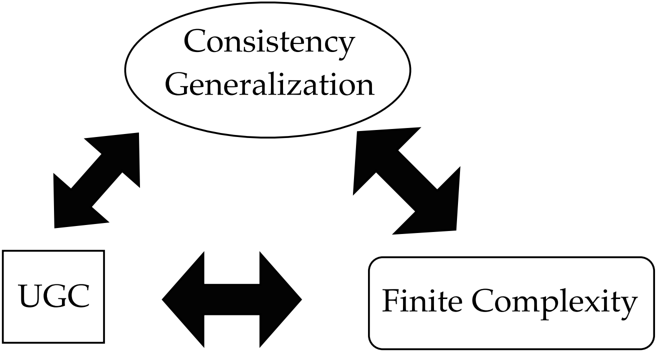
Properties of Generalization Bounds (cont.)

Complexity

A historical approach to generalization bounds is based on controlling the complexity of the hypothesis space (covering numbers, VC-dimension, Rademacher complexities)

Necessary and Sufficient Conditions for Learning

ERM



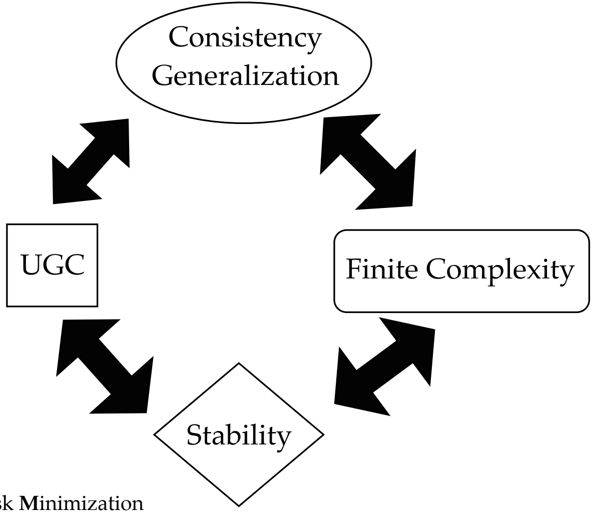
Empirical **R**isk **M**inimization
Uniform **G**livenko **C**antelli

Stability

As we saw in class 2, the basic idea of stability is that a good algorithm should not change its solution much if we modify the training set slightly.

Necessary and Sufficient Conditions for Learning (cont.)

ERM



Empirical **R**isk **M**inimization
Uniform **G**livenko **C**antelli

We explain this approach to generalization bounds, and show how to apply it to Tikhonov Reguarization in the next class.

Note that we will consider a stronger notion of stability, than the one discussed in class 2. Tikhonov regularization satisfies this stronger notion of stability.

Uniform Stability

notation: S training set, $S^{i,z}$ training set obtained replacing the i-th example in S with a new point $z = (x, y)$.

Definition

We say that an algorithm $\mathcal{A}$ has **uniform stability** β (is β-stable) if

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

Uniform Stability (cont.)

- Uniform stability is a strong requirement: a solution has to change very little even when a very unlikely ("bad") training set is drawn.
- the coefficient β is a function of n, and should perhaps be written β_n.

Stability and Concentration Inequalities

Given that an algorithm $\mathcal{A}$ has stability β, how can we get bounds on its performance?
$\Longrightarrow$ Concentration Inequalities, in particular, McDiarmid's Inequality.

Concentration Inequalities show how a variable is concentrated around its mean.

McDiarmid's Inequality

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

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

then the following statement holds:

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

McDiarmid's Inequality

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:

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

Example: 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

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

We will use β-stability to apply McDiarmid's inequality to the defect $D[f_S] = I[f_S] - I_S[f_S]$.

2 steps

1. bound the expectation of the defect
2. bound how much the defect can change when we replace an example

Bounding The Expectation of The Defect

Note that $\mathbb{E}_S = \mathbb{E}_{(z_1,\ldots,z_n)}$.

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

The second 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.

Bounding The Deviation of The Defect

Assume that there exists an upper bound M on the loss.

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

Applying McDiarmid's Inequality

By McDiarmid's Inequality, for any ϵ,

$$\mathbb{P}\left(|D[f_S] - \mathbb{E}D[f_S]| > \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(2(\beta + \frac{M}{n}))^2}\right) =$$

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

A Different Form Of The Bound

Let

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

Solving for ϵ in terms of δ, we find that

$$\epsilon = (n\beta + M)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

We can say that with confidence $1 - \delta$,

$$D[f_S] \leq \mathbb{E}D[f_S] + (n\beta + M)\sqrt{\frac{2\ln(2/\delta)}{n}}$$

But $\mathbb{E}D[f_S] \leq \beta$......

A Different Form Of The Bound

Let

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

Solving for ϵ in terms of δ, we find that

$$\epsilon = (n\beta + M)\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

We can say that with confidence $1-\delta$,

$$D[f_S] \leq \mathbb{E}D[f_S] + (n\beta + M)\sqrt{\frac{2\ln(2/\delta)}{n}}$$

But $\mathbb{E}D[f_S] \leq \beta$......

A Different Form Of The Bound (cont.)

Finally, recalling the definition, of the defect we have with confidence $1 - \delta$,

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

Convergence

Note that if $\beta = \frac{k}{n}$ for some k, we can restate our bounds as

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

and with probability $1-\delta$,

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

Fast Convergence

For the uniform stability approach we've described, $\beta = \frac{k}{n}$ (for some constant k) is "good enough". Obviously, the best possible stability would be $\beta = 0$ — the function can't change at all when you change the training set. An algorithm that always picks the same function, regardless of its training set, is maximally stable and has $\beta = 0$. Using $\beta = 0$ in the last bound, with probability $1 - \delta$,

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

The convergence is still $O\left(\frac{1}{\sqrt{n}}\right)$. So once $\beta = O(\frac{1}{n})$, further increases in stability don't change the rate of convergence.

Summary

We define a notion of stability (β- stability) for learning algorithms and show that generalization bound can be obtained using concentration inequalities (McDiarmid's inequality).

Uniform stability of $O\left(\frac{1}{n}\right)$ seems to be a strong requirement.

Next time, we will show that Tikhonov regularization possesses this property.