Uniform Stability and Generalization in Statistical Learning Theory - a tutorial with Tikhonov Regularization

Brando Miranda

January 2015

Abstract

We provide a tutorial of the foundations of Uniform Stability and Generalization in Learning according to the MIT class 9.520 Statistical Learning Theory and Applications [2] through the example of Tikhonov Regularization. In particular, our main contributions are: 1) to provide a new presentation of the proof 2) provide clear motivations and underlying intuition for pedagogy and 3) provide an example with Tikhonov Regularization for concreteness and pedagogy.

1 Introduction

The notion of [6] is one which captures how much a predictive function changes when the training set is changed slightly. When an algorithm is uniformly stable, intuitively, it means that the algorithm is resistant, in the "worst case", to a change in the training set. In other words, for all training sets and any changes to this training set, the predictive accuracy should not change in a noticeable way. Hence, since the algorithm didn't change that much, then its error shouldn't change (too much) either. In this article we will show that if a learning algorithm L is uniformly stable in a precise sense, then it will generalize. i.e. the error in the training set will be a true reflection of the quality of our predictions on unseen data points. In other words, if the algorithm is resistant to small perturbations in the training set as defined by uniform stability, then we can mathematically show that the learning algorithm will be able to make good predictions on sample points that are inside our training set as well as samples that are not.

2 Motivation

Why would someone want their learning algorithm to be stable in any sense? One can intuitively motivate the concept of stability with the analogy of

a scientific theory, where the power of our predictor lies in the predictive power of the scientific theory. If we had a good scientific theory and thus a good predictor, then it should have been able to see general trends in the data that inspired this theory. In other words, small changes to the data that inspired this theory, should not affect the theory as a whole to much, because the theory was able to extract the general trends effectively in a predictive manner. If it was a good theory in the first place, then small changes in the training data should not affect its success to predict unseen data samples. Similarly, if we have a learning algorithm that is able to generalize on unseen data points, then if the algorithm was "good" at learning from the data it had, then small perturbations on that data should not affect its predictive power too much. This intuitive idea can actually be made mathematically rigorous with uniform stability. If a learning algorithm is (uniformly) stable then, from the previous argument, it makes sense that it should have good predictive power on data it has already seen and data it has not yet seen. i.e. if the learning algorithm is stable, it should be able to generalize.

3 Formal Definition of Uniform Stability

An algorithm L has β with respect to the loss function V if the following holds:

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

A probabilistic version of uniform stability β is:

$$\forall S \in Z^m, \forall i \in \{1, ..., m\}, \mathbb{P}_S\{\sup_{z \in Z} |V(f_S, z) - V(f_{S^i}, z)| \leq \beta\} \geq 1 - \delta$$

4 Summary of Results for Uniform Stability and Generalization

If a learning algorithm L is uniformly stable and also has a bounded loss function, then with confidence $1 - \delta$ the difference between the empirical risk and the generalization error will be upper bounded as follows:

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

Specifically, if one has a uniformly stable learning algorithm with $\beta_n = O\left(\frac{1}{n}\right)$, then the upper bound becomes:

$$I[f_S] - I_S[f_S] \leq O\left(\frac{1}{\sqrt{n}}\right)$$

which clearly approaches zero as n approaches infinity. Therefore, the empirical risk is equal to the generalization error for sufficiently large and thus, the empirical risk is a good proxy for the generalization error. In this case, we say that the learning algorithm generalizes.

4.1 Remarks for main results for Uniform Stability and Generalization

Why minimizing the empirical risk is a good idea for large training sets: Notice that this result is basically saying that for sufficiently large n, the empirical risk and the generalization error are approximately equal. Therefore, this means that given enough training data, finding a predictor that minimizes the empirical risk will actually also minimize the generalization error. Therefore, for a stable learning algorithm, minimizing the empirical risk is actually a good procedure for minimizing generalization (since the two are approximately equal for sufficiently large n). Unfortunately, this is obviously an asymptotic bound and therefore, large values of n are needed. However, this result does justify why minimizing the empirical risk for large training sets might be a good idea.

Justification for bounded loss function: One immediate argument that one could hold against such a theoretical result is an argument against the bounded loss function. In reality, we never choose a bounded loss function. For example, the squared loss, is not bounded for all values on the real line. However, the important argument made to justify such an argument is that in reality, we will never actually expect to observe every value of the real line. Within a realistic domain of values that any reasonable training set might have, we will observe finite bounded numbers z. Therefore, since we don't expect to get any value of the real line, we approximately have a bounded loss function.

5 Example of a Uniformly Stable algorithm: Tikhonov Regularization

Recall tikhonov regularization to be:

$$f_S^\lambda = arg\min_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^{n}V(f(x_i),y_i)+\lambda||f||^2_{\mathbb{R}^k}\right)$$

it can be shown that tikhonov regularization is uniformly stable. If that is true then tikhonov regularization is proved to generalize. To prove that Tikhonov regularization is stable we only need to show these three statements to be true:

1. we assume that the loss is Lipschitz continuous: $|V(f_i(x),y')-V(f_2(x),y')| \leq L||f_1-f_2||_\infty$
2. we need that the hypothesis class to be over reproducing kernel Hilbert spaces (RKHS): $||f-f'||_\infty \leq \kappa||f-f'||_{\mathbb{R}^K}$ for any $f,f'\in\mathcal{H}$
3. finally we need the following lemma to hold: $||f_S^\lambda - f_{S^{i,z}}^\lambda||^2_{\mathbb{R}^K} \leq \frac{L||f_S^\lambda - f_{S^{i,z}}^\lambda||_\infty}{\lambda n}$

If the above holds and the loss function is upper bounded by M, then the generalization bound has the following form:

$$|I[f_S^\lambda]-I_S[f_S^\lambda]| \leq \frac{L^2\kappa^2}{\lambda n}+(\frac{2L^2\kappa^2}{\lambda n}+M)\sqrt{\frac{2ln(\frac{2}{\delta})}{n}}$$

Therefore, with confidence $1-\delta$, tikhonov regularization generalizes as n goes to infinity.

5.1 Remarks on Bound for Tikhonov Regularization

Notice that keeping λ fixed as n increases, the generalization tightens as $O\left(\frac{1}{\sqrt{n}}\right)$. However, fixing λ keeps our hypothesis spaced fixed. However, as we get more data, we want λ to get smaller. However, if λ gets smaller too quickly, then the bounds have the potential to become vacuous.

5.2 Main Result: Tikhonov Regularization is Stable

Theorem 5.1. *If a learning algorithm L is uniformly stable, then as the number of training points approaches infinity the empirical risk approaches the generalization error with high probability i.e.* $\lim_{n\to\infty} I_S[f_S] = I[f_S]$ *with high probability. Therefore, our goal will be to show that if uniform stability holds for L, then the difference between the empirical error and the generalization error will go to zero with high probability i.e.:*

$$Pr[|I_S[f_S] - I[f_S]| \leq \epsilon(n)] \geq 1-\delta$$

where $\epsilon(n)$ *will be an upper bound that approaches zero as n approaches infinity.*

Proof. Let's begin the proof by using the fact that our learning algorithm is uniformly stable. If the learning algorithm$L$ is uniformly stable, then using McDiarmid's inequality, and setting the generalization error to be the functional yields: $\forall(S,z) \in Z^{n+1}, \forall i \in \{1,...,n\}, \sup_{z\in Z} |I[f_S] - I[f_{S^{i,z}}]| \leq \beta \implies P(|I[f_S] - \mathbb{E}_S[I[f_S]]| \geq \epsilon) \leq 2e^{\left(\frac{-2\epsilon^2}{n\beta^2}\right)}$ Which says that the generalization error will be close to the expected generalization error over training sets with high probability. If we change the above probabilistic statement to its confidence form and demand to have $1-\delta$ confidence that $I[f_S]$ and $\mathbb{E}_S[I[f_S]]$ are close, then after the bound equal to δ and some lines of algebra, its easy to verify that ϵ must be at most: $\epsilon = n\beta\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ Therefore, we have with confidence $1-\delta$ that: $|I[f_S] - \mathbb{E}_S[I[f_S]]| \leq n\beta\sqrt{\frac{ln(\frac{2}{\delta})}{2n}} \implies I[f_S] \leq \mathbb{E}_S[I[f_S]] + n\beta\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ The above statement is an upper bound on the generalization error, which can be turned into the desired bound (i.e. one bounding the difference of the empirical risk and generalization error) easily by subtracting the empirical risk $I_S[f_S]$ from both sides of the inequality yielding: $I[f_S] - I_S[f_S] \leq \mathbb{E}_S[I[f_S]] - I_S[f_S] + n\beta\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ the LHS is exactly what we need to conclude the proof, however we need to finish upper bounding the RHS, specifically we need to upper bound $\mathbb{E}_S[I[f_S]] - I_S[f_S]$ to conclude the proof. The upper bound for $\mathbb{E}_S[I[f_S]] - I_S[f_S]$ is exactly: $\mathbb{E}_S[I[f_S]] - I_S[f_S] \leq \beta + (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ where M is the upper bound on the loss function $V(f,z)$. If that is true then the proof concludes that: $I[f_S] - I_S[f_S] \leq \beta + (2n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ Which gives the desired

upper bound and as long as the upper bound decreases as n increases, then the generalization error and empirical risk can be made arbitrarily close. We will finish the proof by proving what we need in lemma 5.2. □

We will finish the proof by proving what we need in the following lemma:

Lemma 5.2. *For a bounded loss function $V(f,z)$ (with upper bound M) and a uniformly stable learning algorithm the following upper bound holds with confidence $1-\delta$*

$$\mathbb{E}_S[I[f_S]] - I_S[f_S] \leq \beta + (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$$

Proof. Notice that the above statement is the difference of $I_S[f_S]$ and the expected value of a different quanity. This suggests that setting $I_S[f_S]$ to be the functional in McDiarmids inequality might be good first step to yield the above result. Thus, we wull use McDiarmid's inequality and let the functional F in McDiarmid's be $I_S[f_S]$. If that is our choice of functional then we must show the following first $\forall i, \sup_{S,z} |I_S[f_S] - I_{S^{i,z}}[f_{S^{i,z}}]| \leq c_i$ if we want the probabilistic upper bound on $\mathbb{E}_S[I_S[f_S]]$ to hold. Let's search for the upper bound c_i by considering the LHS of the above inequality and expanding it: $\forall i, \sup_{S,z} |I_S[f_S] - I_{S^{i,z}}[f_S^{i,z}]| = |\frac{1}{n}\sum_{i=1}^{n} V(f_s, z_i) - (\frac{1}{n}\sum_{j\neq i} V(f_{S^{i,z}}, z_j) + \frac{1}{n}V(f_{S^{i,z}}, z))|$ by triangle inequality and pairing up the samples points $z_i \neq z_j$ in the summations we get: $\leq \frac{1}{n}\sum_{j\neq i} |V(f_S, z_j) - V(f_{S^{i,z}}, z_j)| + \frac{1}{n}|V(f_S, z_i) - V(f_{S^{i,z}}, z)|$ We divided the terms that way because the first term $V(f_S, z_j) - V(f_{S^{i,z}}, z_j)$ can be upper bounded by β because of the stability of our learning algorithm. Thus, $V(f_S, z_j) - V(f_{S^{i,z}}, z_j) \leq \beta$ and the second term can only be bounded by the the upper bound of the loss function (because we are evaluating the loss at two different points with different training points): $V(f_S, z_i) - V(f_{S^{i,z}}, z) \leq M$ Combining both terms yields the desired upper bound (to use McDirmad's afterwards): $\leq \frac{n-1}{n}\beta + \frac{M}{n} \leq \beta + \frac{M}{n}$ With this last result we can apply McDiarmid's inequality and the constants c_i at the beginning of the lemma are $\beta + \frac{M}{n}, \forall i$. Thus we have: $\forall i, \sup_{S,z} |I_S[f_S] - I_{S^{i,z}}[f_S^{i,z}]| \leq \beta + \frac{M}{n} \implies Pr[|I_S[f_S] - \mathbb{E}_S[I_S[f_S]]| \geq \epsilon] \leq 2e^{\frac{-2n\epsilon^2}{(n\beta+M)^2}}$ Switching the above bound to its confidence form and requiring it to have confidence $1-\delta$ yields ϵ to be: $\epsilon = (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ Therefore, from the probabilistic bound implied by McDiarmid's inequality we have: $|I_S[f_S] - \mathbb{E}_S[I_S[f_S]]| \leq (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}} \implies I_S[f_S] \leq \mathbb{E}_S[I_S[f_S]] + (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ Which isn't exactly what we wanted to prove by the lemma. However we can subtract $\mathbb{E}_S[I[f_S]]$ from both sides of the inequality to get it in the form of the original lemma statement. Doing that yields: $I_S[f_S] - \mathbb{E}_S[I[f_S]] \leq \mathbb{E}_S[I_S[f_S]] - \mathbb{E}_S[I[f_S]] + (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ Which will conclude the proof of the lemma if we can upper bound $\mathbb{E}_S[I_S[f_S]] - \mathbb{E}_S[I[f_S]]$. Which is actually simple to show its upper bounded is β. First notice that: $\mathbb{E}_S[I_S[f_S]] - \mathbb{E}_S[I[f_S]] = \mathbb{E}_z[\mathbb{E}_S[I_S[f_S]]] - \mathbb{E}_S[I[f_S]] = \mathbb{E}_z[\mathbb{E}_S[\frac{1}{n}\sum_{i=1}^{n} V(f, z_i)]] -$

$\mathbb{E}_S[\mathbb{E}_z[V(f_S, z)]]$ Then by linearity of expectation and some simple algebra we have: $\mathbb{E}_z[\mathbb{E}_S[\frac{1}{n}\sum_{i=1}^{n} V(f, z_i) - V(f_S, z)]] = \mathbb{E}_z[\mathbb{E}_S[\frac{1}{n}\sum_{i=1}^{n}(V(f_{S^{i,z}}, z) - V(f_S, z))]]$ Then by stability we know that each term $V(f_{S^{i,z}}, z) - V(f_S, z) \leq \beta$. Thus: $\mathbb{E}_S[I_S[f_S]] - \mathbb{E}_S[I[f_S]] \leq \beta$ Which concludes the proof of the lemma and shows that with confidence $1-\delta$ we have: $\mathbb{E}_S[I[f_S]] - I_S[f_S] \leq \beta + (n\beta + M)\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$ □

Acknowledgments

This work was done as a final project the MIT class "Statistical Learning Theory and Applications" (9.520) [2] at MIT taught by professor Tomaso Poggio, professor Lorenzo Rosasco et al. and we would like to thank the staff for the fantastic class and supporting material. In particular we'd like to thank for the supporting slides [4, 5] and text book [3] (in progress at the time of this writing 2014-2015). We also thank Quora for providing the platform to share the first version of this tutorial [1].

References

[1] *(60) Uniform Stability and Generalization in Learning theory - Brando Miranda's Posts*. URL: https://gzmvtkignlhlqibs.quora.com/Uniform-Stability-and-Generalization-in-Learning-theory.

[2] *9.520, Fall 2014*. URL: http://www.mit.edu/~9.520/fall14/.

[3] *9.520, Fall 2014*. URL: http://www.mit.edu/~9.520/fall14/.

[4] Lorenzo Rosasco et al. *Generalization Bounds and Stability*. URL: http://www.mit.edu/~9.520/fall14/slides/class15/class15_stability.pdf.

[5] Lorenzo Rosasco et al. *Stability of Tikhonov Regularization*. URL: http://www.mit.edu/~9.520/fall14/slides/class16/class16_stability.pdf.

[6] *Stability (learning theory) - Wikipedia*. URL: https://en.wikipedia.org/wiki/Stability_(learning_theory).