Machine Learning Theory (CS 6783)

Lecture 12 : Characterizing Learnability Via Algorithmic Stability

1 Algorithmic Stability

Before we talk about stability of a learning algorithm, we need to give a notation for a learning algorithm. Specifically, we define an algorithm $\hat{\mathbf{y}}$ by a mapping of form $\hat{\mathbf{y}} : \bigcup_{t=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^t \mapsto \mathcal{Y}^{\mathcal{X}}$. That is a function that takes as input sample (in $\mathcal{X} \times \mathcal{Y}$) of arbitrary length and maps it to a model that maps input instances in $\mathcal{X}$ to outcome $\mathcal{Y}$. In other words, a learning algorithm takes a sample of any length and outputs a model (a model that predicts outcome in $\mathcal{Y}$ given an input in $\mathcal{X}$). Indeed, an algorithm like ERM takes a sample and returns as output a model that minimizes training error on given sample. Now with this definition of an algorithm, we are ready to talk about algorithmic stability.

Informally, an algorithm is said to be stable if deleting a sample from the training set does not change outcome by much. To define such stability, let us first introduce some notation. Given a sample S, let $S^{\setminus i}$ denote the sample got by deleting the i'th sample point from S. That is, $S^{\setminus i} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$.

Definition 1. *An algorithm $\hat{\mathbf{y}}$ is said to be stable w.r.t. a distribution $\mathbf{D}$ with rate ϵ_{stable} if:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\left[\left|\ell(\hat{\mathbf{y}}(S)(x_i), y_i) - \ell(\hat{\mathbf{y}}(S^{\setminus i})(x_i), y_i)\right|\right] \leq \epsilon_{\text{stable}}(n)$$

Definition 2. *An algorithm $\hat{\mathbf{y}}$ is said to be an Approximate ERM (AERM) w.r.t. a class of models $\mathcal{F}$ with rate ϵ_{ERM} if for any sample S of size n,*

$$\widehat{L}_S(\hat{\mathbf{y}}(S)) \leq \min_{f \in \mathcal{F}} \widehat{L}_S(f) + \epsilon_{\text{ERM}}(n)$$

If an algorithm is stable, its test loss and training loss are close (or in other words it generalizes well). If further, the algorithm is an approximate ERM (i.e it approximately minimizes training loss), then such an algorithm has low excess risk in expectation. The following theorem shows that a stable algorithm that is also an AERM, has a low expected excess risk.

Theorem 1. *If a learning algorithm is LOO stable with rate ϵ_{stable} and is also an AERM with rate ϵ_{ERM}, then we have the bound on expected excess risk,*

$$\mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S))\right] - \min_{f \in \mathcal{F}} \widehat{L}_{\mathbf{D}}(f) \leq \epsilon_{\text{stable}}(n+1) + \epsilon_{\text{ERM}}(n+1)$$

Proof.

$$\begin{aligned}
\epsilon_{\text{stable}}(n) &\geq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S\left[\left|\ell(\hat{\mathbf{y}}(S)(x_i),y_i)-\ell(\hat{\mathbf{y}}(S^{\setminus i})(x_i),y_i)\right|\right]\\
&\geq \mathbb{E}_S\left[\left|\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{\mathbf{y}}(S)(x_i),y_i)-\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{\mathbf{y}}(S^{\setminus i})(x_i),y_i)\right|\right]\\
&= \mathbb{E}_S\left[\left|\widehat{L}_S(\hat{\mathbf{y}}(S))-\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{\mathbf{y}}(S^{\setminus i})(x_i),y_i)\right|\right]\\
&\geq \left|\mathbb{E}_S\left[\widehat{L}_S(\hat{\mathbf{y}}(S))-\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{\mathbf{y}}(S^{\setminus i})(x_i),y_i)\right]\right|\\
&= \left|\mathbb{E}_S\left[\widehat{L}_S(\hat{\mathbf{y}}(S))-\frac{1}{n}\sum_{i=1}^{n}\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus i}))\right]\right|\\
&= \left|\mathbb{E}_S\left[\widehat{L}_S(\hat{\mathbf{y}}(S))-\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))\right]\right|\\
&\geq \mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))-\widehat{L}_S(\hat{\mathbf{y}}(S))\right]\\
&\geq \mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))-\min_{f\in\mathcal{F}}\widehat{L}_S(f)\right]-\epsilon_{\text{ERM}}(n)\\
&= \mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))\right]-\mathbb{E}_S\left[\min_{f\in\mathcal{F}}\widehat{L}_S(f)\right]-\epsilon_{\text{ERM}}(n)\\
&\geq \mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))\right]-\min_{f\in\mathcal{F}}\mathbb{E}_S\left[\widehat{L}_S(f)\right]-\epsilon_{\text{ERM}}(n)\\
&= \mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))\right]-\min_{f\in\mathcal{F}}\widehat{L}_{\mathbf{D}}(f)-\epsilon_{\text{ERM}}(n)
\end{aligned}$$

Hence we have proved that

$$\mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S^{\setminus n}))\right]-\min_{f\in\mathcal{F}}\widehat{L}_{\mathbf{D}}(f)\leq\epsilon_{\text{stable}}(n)+\epsilon_{\text{ERM}}(n)$$

However, note that the above says that if we provide the algorithm with sample of size $n-1$ (ie. the last sample deleted), then in expectation we have the excess risk bound of $\epsilon_{\text{stable}}(n)+\epsilon_{\text{ERM}}(n)$. Since n is arbitrary, we can conclude that

$$\mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S))\right]-\min_{f\in\mathcal{F}}\widehat{L}_{\mathbf{D}}(f)\leq\epsilon_{\text{stable}}(n+1)+\epsilon_{\text{ERM}}(n+1)$$

□

Remark 1.1. *Note that a simple Markov inequality can convert an expected statement to one that holds with say probability* $1/2$*. From this, using the style of analysis you guys did in Assignment 1, Question 1, you can convert the statement into a high probability one.*

The below theorem shows that the converse is also true. That is, if a problem is learnable with some rate against all distributions, then there always exists a stable AERM.

Theorem 2. *If a learning algorithm* $\hat{\mathbf{y}}$ *has the following expected excess risk guarantee for all distributions* $\mathbf{D}$

$$\mathbb{E}_S\left[\widehat{L}_{\mathbf{D}}(\hat{\mathbf{y}}(S))\right] - \min_{f\in\mathcal{F}} \widehat{L}_{\mathbf{D}}(f) \leq \epsilon_{\text{rate}}(n),$$

then there always exists a randomized learning algorithm that is both stable and an AERM

Proof. Given access to a learning algorithm $\hat{\mathbf{y}}$ with rate ϵ_{rate}, we will show that one can obtain a stable AERM using this algorithm as a routine. To this end, consider the following algorithm $\tilde{\hat{\mathbf{y}}}$ that does the following. Given any sample of any size n, we first draw a new sample S' by drawing $n' = \text{round}(\sqrt{n})$ samples uniformly with replacement from the sample set S. Then we run the algorithm $\hat{\mathbf{y}}$ on S' and so, $\tilde{\hat{\mathbf{y}}}(S) = \hat{\mathbf{y}}(S')$. Now let us consider the properties if this randomized algorithm.

Stability : Note that this algorithm only depends on at most n' samples of the n samples in S and so deleting any of the remaining $n - n'$ samples does not even alter the outcome of this algorithm. Hence,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\left[\left|\ell(\tilde{\hat{\mathbf{y}}}(S)(x_i), y_i) - \ell(\tilde{\hat{\mathbf{y}}}(S^{\setminus i})(x_i), y_i)\right|\right] \leq O\left(\frac{n'}{n}\right) \leq O\left(\sqrt{\frac{1}{n}}\right)$$

Almost ERM: Now let us use the fact that the Algorithm $\hat{\mathbf{y}}$ has a guaranteed rate of ϵ_{rate} to show that $\tilde{\hat{\mathbf{y}}}$ is an AERM. To this end, note that the algorithm $\tilde{\hat{\mathbf{y}}}$ simply runs algorithm $\hat{\mathbf{y}}$ on sample S' and sample S' is an iid draw from empirical sample S. Hence, the population loss according to the uniform distribution is the training loss $\widehat{L}_S$. However the learning guarantee of $\hat{\mathbf{y}}$ tells us that

$$\mathbb{E}\left[\widehat{L}_S(\hat{\mathbf{y}}(S'))\right] - \inf_{f\in\mathcal{F}} \widehat{L}_S(f) \leq \epsilon_{\text{rate}}(n') = \epsilon_{\text{rate}}(\sqrt{n})$$

This automatically shows that the randomized algorithm is an AERM with rate $\epsilon_{\text{rate}}(\sqrt{n})$ in expectation. □

Combining the two theorems above we can conclude the following corollary.

Corollary 3. *A statistical learning problem is learnable if an only if there exists a stable, approximate ERM for the problem.*