Introduction to Statistical Learning Theory

Lecture 6

We will study a new criteria for learnability - stability.

Intuitively, a stable algorithm is one that a small change to the input results in a small change to the output.

There are a few ways to formalize this idea, we will go with the following:

Consider a training set $S = \{z_1, ..., z_m\}$ and an additional example z'. Define $S^{(i)} = S \cup z'/z_i$ an alternative training set where z' replaces z_i.

If an algorithm is stable, we would expect $\ell(A(S^{(i)}), z_i)$ to be close to $\ell(A(S), z_i)$.

Definition 1.1 (Replace-One-Stable - ROS)

Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is Replace-One-Stable with rate $\epsilon(m)$ if for all S and z'

$$\ell(A(S^{(i)}), z_i) - \ell\left(A(S), z_i\right) \leq \epsilon(m)$$

Definition 1.2 (On-Average-Replace-One-Stable - OAROS)

We say that a learning algorithm A is On-Average-Replace-One-Stable with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$ we have

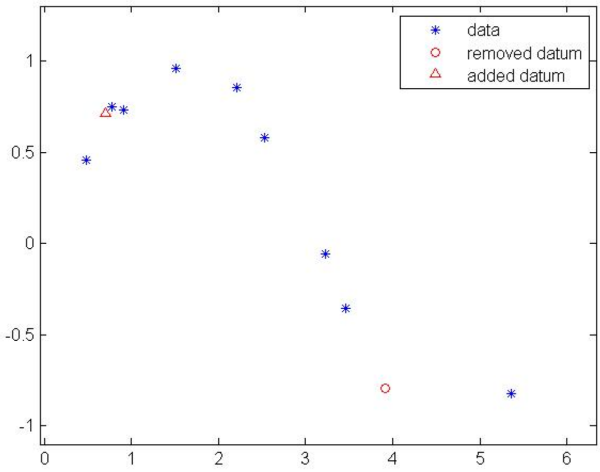
$$\mathop{\mathbb{E}}_{S,z'} \mathop{\mathbb{E}}_{i \sim U(m)} \left[\ell(A(S^{(i)}), z_i) - \ell\left(A(S), z_i\right)\right] \leq \epsilon(m)$$

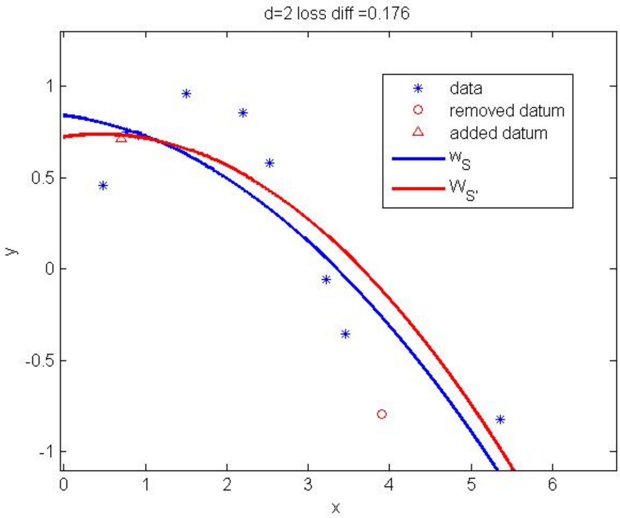
Where $U(m)$ is the uniform distribution on $1, ..., m$.

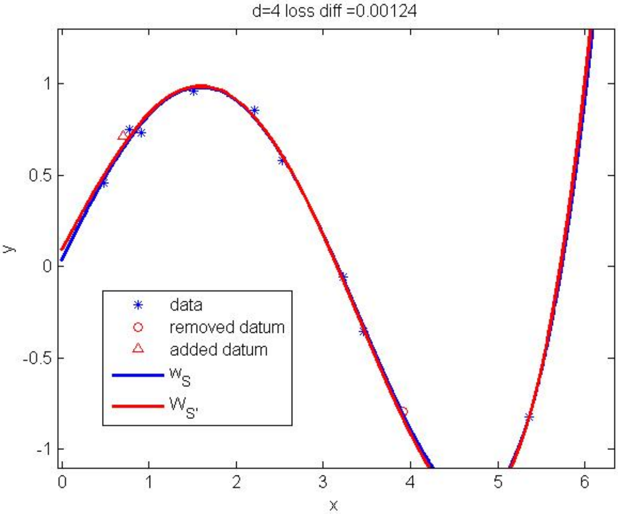
We will see some examples that will give some intuition as to why this leads to genralization.

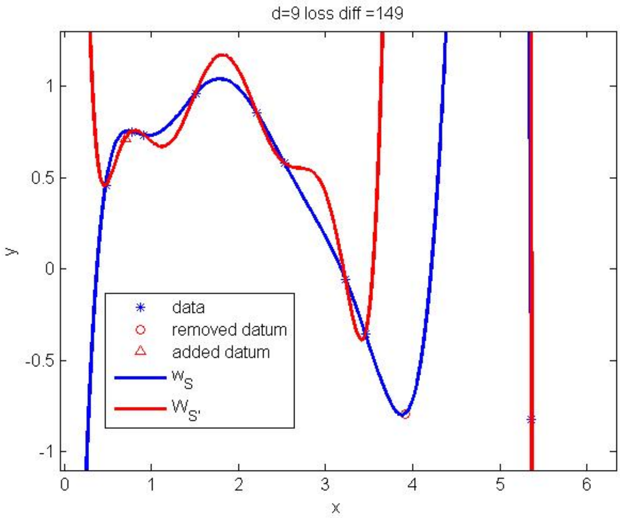
$\mathcal{X} = [0, 2\pi]$ with uniform distribution, $\mathcal{Y} = \mathbb{R}$ and let ℓ be the square loss $\ell(y_1, y_2) = (y_1 - y_2)^2$. We define the probability on y (give x) as $y = sin(x) + \mathcal{N}(0, 0.05)$, and we are given $m = 10$ data points.

Our hypothesis spaces are polynomials with degree d, and we use the ERM algorithm.


data
removed datum
added datum
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6


d=2 loss diff =0.176
data
removed datum
added datum
w_S
$w_{S'}$
y
x
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6


d=4 loss diff =0.00124
data
removed datum
added datum
w_S
w_S'
x
y
0
1
2
3
4
5
6
1
0.5
0
-0.5
-1


d=9 loss diff =149
y
x
data
removed datum
added datum
w_S
$w_{S'}$
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6

We will show that stable algorithms do not overfit, then show how regularization can produce stability. As ROS implies OAROS it is enough to prove for OAROS

Theorem 1.3

Let A be a learning algorithm with OAROS stability rate $\epsilon(m)$, then

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] \leq \epsilon(m) \tag{1}$$

Proof - We will show that
$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] = \mathop{\mathbb{E}}_{S,z'}\mathop{\mathbb{E}}_{i\sim U(m)}\left[\ell(A(S^{(i)}), z_i) - \ell\left(A(S), z_i\right)\right]$,
then we are done by definition.

Since S and z' are drawn i.i.d from $\mathcal{D}$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_S[L_{\mathcal{D}}(A(S))] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S), z')] &= \mathop{\mathbb{E}}_{S,z'}[\ell(A(S^{(i)}), z_i)] \\ &= \mathop{\mathbb{E}}_{S,z'} \mathop{\mathbb{E}}_{i\sim U(m)}[\ell(A(S^{(i)}), z_i)]\end{aligned}$$

On the other hand,

$$\mathop{\mathbb{E}}_S[L_S(A(S))] = \mathop{\mathbb{E}}_S \mathop{\mathbb{E}}_{i\sim U(m)}[\ell(A(S), z_i)] = \mathop{\mathbb{E}}_{S,z'} \mathop{\mathbb{E}}_{i\sim U(m)}[\ell(A(S), z_i)]$$

And this finishes the proof. □

Stability itself is not a sufficient condition of learnability. Take for example the constant learning algorithm which returns the same hypothesis h for all S.

Definition 1.4 (Approximately-ERM)

Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is an approximately-ERM (or AERM) with rate $\epsilon(m)$ if for all datasets S of size m we have

$$L_S(A(S)) \leq L_S(h_{ERM}) + \epsilon(m)$$

Theorem 1.5 (Learnability of stable AERM)

If algorithm A is OAROS stable with rate $\epsilon_{stable}(m)$ and AERM with rate $\epsilon_{ERM}(m)$ then

$$\mathbb{E}_{S}\left[L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}(h^*)\right] \leq \epsilon_{ERM} + \epsilon_{stable} \tag{2}$$

where $h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Proof:

$$\mathbb{E}_{S}\left[L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}(h^*)\right] = \mathbb{E}_{S}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] + \mathbb{E}_{S}\left[L_S(A(S)) - L_S(h^*)\right] + \mathbb{E}_{S}\left[L_S(h^*) - L_{\mathcal{D}}(h^*)\right] \leq \epsilon_{stable} + \epsilon_{ERM} + 0$$

The last theorem did not exactly prove PAC learnability - we gave a bound on the expectation while we need a high probability bound. This can be fixed - see assignment 1, question 3.

We have shown that $AERM + stability \Rightarrow learnable$. If is possible to prove the converse - that if a problem is learnable, it is learnable by a stable AERM algorithm.

We will now show a how a standard ML practice, ℓ_2-regularization, stabilizes learning.

We will first need to quick introduction to strong convexity.

Definition 2.1 (Strong convexity)

A function f is λ-strongly convex for $\lambda > 0$ if for all x, y in its domain and $\alpha \in [0, 1]$

$$f\left(\alpha x + (1-\alpha)y\right) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\lambda\alpha(1-\alpha)}{2}||x-y||_2^2$$

This gives some intuition - a smooth function is convex iff $\nabla^2 f \succeq 0$. A smooth function is λ strongly convex iff $\nabla^2 f \succeq \lambda I$.

Many of the properties of strongly arise from the simple fact that $f(x)$ is λ strongly convex iff $g(x) = f(x) - \frac{\lambda}{2}||x||^2$ is convex.

Lemma 2.2

1. *The function* $f_S(x) = \frac{\lambda}{2}||x||^2$ *is* λ *strongly convex.*
2. *If* f *is* λ_1 *strongly convex and* g *is* λ_2 *strongly convex then* $f+g$ *is* $\lambda_1+\lambda_2$ *strongly convex.*
3. *If* f *is convex and* g *is* λ *strongly convex then* $f+g$ *is* λ *strongly convex.*
4. *If* f *is* λ *strongly convex and* x^* *is the minimizer of* f *then for any* x, $f(x) - f(x^*) \geq \frac{\lambda}{2}||x - x_0||^2$.

Proof - 1+2 follow from definition. 3 follows from 2 using the fact that convex is 0-strongly convex. We prove 4 for twice differential function: From Tylor theorem

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{1}{2}(x - x^*)^T \nabla^2 f(z)(x - x^*) \geq \frac{\lambda}{2}||x - x^*||^2$$

We will now prove that l_2 regularization is stable for Lipschitz loss.

Theorem 2.3

Define the l_2 regularized ERM algorithm as $A(S) = \arg\min_w \left(L_S(w) + \lambda||w||^2\right)$. *If ℓ be a ρ-Lipschitz convex loss function, $A(S)$ is Replace-One-Stable with rate* $\epsilon(m) = \frac{2\rho^2}{\lambda m}$

Proof: Define $f_S(v) = L_S(v) + \lambda||v||^2$. From Lemma 2.2 if is 2λ strongly convex and $f_S(v) - f_S(A(S)) \geq \lambda||v - A(S)||^2$. On the other side:

$$f_S(v) - f_S(u) = L_S(v) - L_S(u) + \lambda(||v|| - ||u||) = L_{S^{(i)}}(v) - L_{S^{(i)}}(u) + \lambda(||v|| - ||u||) + \frac{\ell(v, z_i) - \ell(u, z_i)}{m} + \frac{\ell(u, z') - \ell(v, z')}{m}.$$

$$f_S(v) - f_S(u) = L_S(v) - L_S(u) + \lambda(||v|| - ||u||) = L_{S^{(i)}}(v) - L_{S^{(i)}}(u) +$$
$$\lambda(||v|| - ||u||) + \frac{\ell(v, z_i) - \ell(u, z_i)}{m} + \frac{\ell(u, z') - \ell(v, z')}{m}$$

If we set $v = A(S^{(i)})$, $u = A(S)$ and remember that v minimizes $L_{S^{(i)}}(w) + \lambda||w||^2$ we can conclude that

$$\lambda||A(S^{(i)}) - A(S)||^2 \leq f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} +$$
$$\frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m} \leq \frac{2\rho}{m}||A(S^{(i)}) - A(S)||.$$

So $||A(S^{(i)}) - A(S)|| \leq \frac{2\rho}{\lambda m}$ and $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{2\rho^2}{\lambda m}$ □

As we have seen $AERM + stability \Rightarrow learnability$. We have shown that l_2 regularized ERM is stable, we now need AERM.

Theorem 2.4

Let $A(S) = \arg\min_w (L_S(w) + \lambda ||w||^2)$, then $A(S)$ is AERM with rate $\epsilon(m) = \lambda ||w_{ERM}||^2$

As $L_S(A(S)) \leq L_S(A(S)) + \lambda ||A(S)||^2 \leq L_S(w_{ERM}) + \lambda ||w_{ERM}||^2$ □

Corollary 2.5

Let ℓ be a convex ρ-Lipschitz loss function and assume $\forall w \in \mathcal{H} : ||w|| \leq B$ then for $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ the regularized ERM satisfies

$$\mathop{\mathbb{E}}_S [L_{\mathcal{D}}(A(S))] \leq \min_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \rho B \sqrt{\frac{8}{m}}$$

Proof - We have $\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq \min_{w\in\mathcal{H}} L_{\mathcal{D}}(w) + \epsilon_{stable}(m) + \epsilon_{ERM}(m)$.
We proved that $\epsilon_{ERM} \leq \lambda B^2$ and $\epsilon(m) = \frac{2\rho^2}{\lambda m}$. Setting $\lambda = B\rho\sqrt{\frac{8}{m}}$ finishes the proof. □

The problem with this proof is that we added the boundness assumption. Even without it we can prove

Theorem 2.6

Let ℓ be a convex ρ-Lipschitz loss function. The regularized ERM satisfies

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(w^*) + \lambda||w^*||^2 + \frac{2\rho^2}{\lambda m}$$

where $w^* = \arg\min_{w\in\mathcal{H}} L_{\mathcal{D}}(w)$.

Proof: We have

$$\mathbb{E}\left[L_S(A(S))\right] \leq \mathbb{E}\left[L_S(A(S)) + \lambda||A(S)||_2^2\right] \leq \mathbb{E}\left[L_S(w^*) + \lambda||w^*||_2^2\right] =$$
$$= L_{\mathcal{D}}(w^*) + \lambda||w^*||_2^2.$$

On the other hand we have

$$\begin{aligned}\mathbb{E}\left[L_{\mathcal{D}}(A(S))\right] =& \mathbb{E}\left[L_S(A(S))\right] + \mathbb{E}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] \\ &\leq L_{\mathcal{D}}(w^*) + \lambda||w^*||_2^2 + \epsilon_{stable}(m) \\ &\leq L_{\mathcal{D}}(w^*) + \lambda||w^*||_2^2 + \frac{2\rho^2}{\lambda m}\end{aligned}$$

□

Theorem 2.6 proves that regularized ERM can learn if the right λ is chosen. We however cannot chose the right one without knowing $||w^*||$.

Nevertheless there are many practical methods of finding the right parameter such as validation set, cross validation etc.

An important example of such a problem is the SVM we discussed previously.

We will show an example of a learning problem that is learnable (via RLM) but without uniform convergence.

As almost all ”standard” learnable problems have uniform convergence, this is an infinite dimensional example.

Define $\mathcal{H} = \mathcal{B} = \{(x_1, ..., x_n, ...) | \sum_{i=1}^{\infty} x_i^2 \leq 1\}$. $Z = \mathcal{B} \times \{0, 1\}^\infty$

The loss function is defined as $\ell(h, (x, \alpha)) = \sum \alpha_i \times (x_i - h_i)^2$.

Intuition - h^* is the center of mass, but at each example α picks dimensions to ignore.

Lemma 3.1

The loss function is convex and Lipschitz, and therefore the problem is learnable with regularized loss minimization (note that $\mathcal{H}$ is bounded in norm).

To prove Lipschitz, it is enough to prove bounded gradient norm. As $||\nabla \ell(h,(x,\alpha))||^2 \leq 4||x-h||^2$. □

Lemma 3.2

The problem is not learnable via ERM, and therefore does not have uniform convergence.

Define the distribution $\mathcal{D}$ such that $x \equiv \mathbf{0}$ and α_i are i.i.d with probability 1/2.
For each finite sample S^m there exists with probability 1 a dimension k such that $\alpha_k = 0$ for all $(x, \alpha) \in S^m$. Then $h = \mathbf{e_k}$ has $L_S(h) = 0$ but $L_{\mathcal{D}}(h) = 1/2$. □

It is possible to modify the example such that a unique ERM exists.