CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 9: Stability of Learning Algorithms

Lecturer: Abir De *Scribe: Groups 1 & 2*

9.1 Characterization of Stability of an Algorithm

Let A be a Learning Algorithm and S be the Data set which is fed into the Learning algorithm.The outcome/output of the learning algorithm is $A(S)$.(We can think of $A(S)$ as a vector to define a norm)

Definition 9.1. *(Stability) A learning algorithm A is said to be stable iff*

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right)$$

For every S and S' such that $|S \backslash S'| = |S' \backslash S| = 1$.

The condition on S and S' means that there is only a one element mismatch between the sets.

Consider instead, what happens if we just delete one element e from the set and take the norm of the difference:(Stability towards single element deletions)

$$\|A(S) - A(S \backslash e)\|$$

We want to find the relation of the above with the previously defined notion of stability. This is dealt with in the following theorem.

Proposition 9.2. *Let A be a Learning Algorithm, $S = \{(x_i, y_i)\}$ be a data set and e be a single data point, $e = (x_r, y_r)$ for some r such that $e \in S$.The following is a sufficient condition for the Algorithm to be stable:*

$$\|A(S) - A(S \backslash e)\| = \mathcal{O}\left(\frac{1}{|S|}\right) \quad \forall e, S$$

Proof. Consider set S and S' such that $|S \backslash S'| = |S' \backslash S| = 1$. This means that there exists e and e' such that $S \backslash e = S' \backslash e'$. We shall also be using Triangle inequality. Let us start with the expression in the definition of stability:

$$\|A(S) - A(S')\| = \|A(S) - A(S \backslash e) + A(S' \backslash e') - A(S')\|$$

(We can do this since $S \backslash e = S' \backslash e'$).Now applying Triangle inequality to the right hand side:

$$\|A(S) - A(S')\| \leq \|A(S) - A(S \backslash e)\| + \|A(S' \backslash e') - A(S')\|$$

But we already have :

$$\|A(S) - A(S\backslash e)\| = \mathcal{O}\left(\frac{1}{|S|}\right)$$

$$\|A(S'\backslash e') - A(S')\| = \mathcal{O}\left(\frac{1}{|S'|}\right)$$

Using this, we have:

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right) + \mathcal{O}\left(\frac{1}{|S'|}\right)$$

Since $|S| = |S'|$:

$$\|A(S) - A(S')\| \leq \mathcal{O}\left(\frac{1}{|S|}\right)$$

□

Note: If we add noise to x_i then accuracy will decrease, but our model will become more stable.

9.2 Applying stability to classification

Let us say we have a dataset $D = \{(x_i, y_i)\}$. Let us say we have some convex loss function $l(w^T x, y)$ which is Lipschitz continuous. Let us define the following function over $S \subset D$ which has regularization

$$F_w(S) = \sum_S (l(w^T x_i, y_i) + \lambda \|w\|^2)$$

Using this function we can define the following vector which minimizes the sum of the loss as

$$w^*(S) = \text{argmin}_w F_w(S)$$

Proposition 9.3. *For the defined $F_w(S)$ with a convex and Lipschitz $l(w^T x, y)$, w^* is stable.*

Proof. Let us define the notation $l(w^*(S), e) = l(w^*(S)^T x, y)$. Now we take a close look at the value $F_{w^*(S')}(S) - F_{w^*(S)}(S)$. We must have the following hold

$$F_{w^*(S')}(S) - F_{w^*(S)}(S) = F_{w^*(S')}(S') - F_{w^*(S)}(S') + l(w^*(S'), e) - l(w^*(S), e) + l(w^*(S), e') - l(w^*(S'), e')$$

Since $w^*(S') = \text{argmin}_w F_w(S')$ we have $F_{w^*(S')}(S') - F_{w^*(S)}(S') \leq 0$ hence

$$F_{w^*(S')}(S) - F_{w^*(S)}(S) \leq l(w^*(S'), e) - l(w^*(S), e) + l(w^*(S), e') - l(w^*(S'), e') \leq 2L\|w^*(S) - w^*(S')\|$$

The last part of the inequality comes by combining the triangle inequality with the Lipschitz condition of $l(w^*(S'), e) - l(w^*(S), e) \leq L\|w^*(S) - w^*(S')\|$.
We can also expand $F_{w^*(S')}(S) - F_{w^*(S)}(S)$ as a taylor expansion about the point $w^*(S)$.

$$F_{w^*(S')}(S) - F_{w^*(S)}(S) = \left.\frac{\partial F_w(S)}{\partial w}\right|_{w=w^*(S)} (w - w^*(S)) + \frac{1}{2}(w - w^*(S))^T H (w - w^*(S)) + \ldots$$

Here $H(F_w(S))$ is the Hessian for the function $F_w(S)$ with respect to w. We know that $w^*(S)$ minimizes $F_w(S)$ hence the first term vanishes and we are left with the inequality

$$F_{w^*(S')}(S) - F_{w^*(S)}(S) \geq \frac{1}{2}(w^*(S') - w^*(S))^T H(F_{w^*(S')}(S))(w^*(S') - w^*(S))$$

We know that $l(w, e)$ is a convex function hence the Hessian $H(l(w, e))$ is positive semi-definite. Hence we can surely conclude that the Hessian of the sum of all $l(w, e)$ terms is also positive semi-definite.

Now we can look at the regularization term, this will have to add a $2\lambda|S|I$ to the Hessian by definition and so we can conclude that $H(F_w(S)) \geq 2\lambda|S|I$ since the loss terms Hessian will anyways be positive semi-definite. Hence we have

$$F_{w^*(S')}(S) - F_{w^*(S)}(S) \geq \frac{2\lambda|S|}{2}(w^*(S') - w^*(S))^T(w^*(S') - w^*(S)) \geq \lambda|S|\|w^*(S') - w^*(S)\|^2$$

By combining the two inequalities we obtain by using first the Lipschitz condition and then that of convexity we obtain

$$\lambda|S|\|w^*(S') - w^*(S)\|^2 \leq F_{w^*(S')}(S) - F_{w^*(S)}(S) \leq 2L\|w^*(S) - w^*(S')\|$$

This subsequently reduces to

$$\|w^*(S') - w^*(S)\| \leq \frac{2L}{\lambda|S|} = \mathcal{O}\left(\frac{1}{|S|}\right)$$

Hence we have proven that with a convex and Lipschitz $l(w^T x, y)$, w^* is stable. $\square$

9.3 Group Details

- 200110055 Keshav Patel Keval
- 19D070017 Bhavishya
- 200100127 Rahul
- 19D070046 Phansalkar Ishan Shrirang
- 190260027 Mahadevan Subramanian
- 190040112 Shivang Tiwari