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

Lecture 9: Deriving stability of a classification

Lecturer: Abir De *Scribe: Group 9*

9.1 Defining Stability

We define an algorithm $A : \mathcal{S} \to \mathbb{R}^d$ to be stable if

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

The condition for this stability is for sets S and S' such that they differ only in element hence $S' = e' \cup (S \backslash e)$ where $e \neq e'$.

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

We want to see that whether it is true that changing one element from the set changes its stability or not.

Proof. We see that $\|A(S) - A(S\backslash e)\| = \mathcal{O}(1/|S|)$ and $\|A(S') - A(S'\backslash e')\| = \mathcal{O}(1/|S'|)$.

$$\begin{aligned}\|A(S) - A(S')\| &\leq \|A(S) - A(S\backslash e)\| + \|A(S'\backslash e') - A(S')\| \\ &= \mathcal{O}(1/|S|) + \mathcal{O}(1/|S'|) \\ &= \mathcal{O}\left(\frac{1}{|S|}\right)\end{aligned}$$

Hence the statement is true. □

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 9.1. *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.

9.3 Group Details and Individual Contribution

Classification section done by Mahadevan Subramanian, Roll no. 190260027. First section by Shivang Tiwari (190040112). □