CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Mohamed Ibrahim Ahmed

Lecture # 10
October 21, 2016

1 Recap

In the last lecture, we covered linear models using three algorithms: linear program, Perceptron and SVM. We assumed for these algorithms that the given data is linearly separable. However, what happens if the data is not linearly separable? Also, we want to prove in this lecture better error bounds for SVM, bounds that do not depend on dimensionality (d) of the data as VC theory shows. In this lecture, we want to prove algorithm specific bounds for the error of the SVM algorithm. First let's see how to deal with non-separable data.

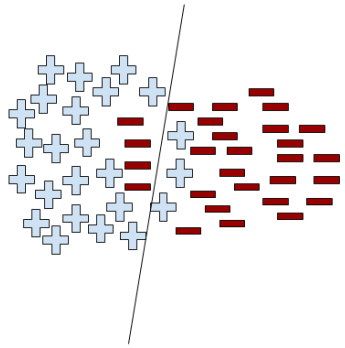

Figure 1: *Non linearly separable example.*

2 Soft-Margin SVM

In soft-margin SVM, we put penalties for violating the constraints. Formally we optimize

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \frac{1}{m}\sum_{1=1}^{m}\epsilon_i + \lambda \left\|w\right\|^2 \\ \text{where} \quad & \epsilon_i = max(0, 1 - y_i(w.x_i))\ i = 1, \ldots, m. \end{aligned}$$

Which is equivalent to:

$$\underset{x}{\text{minimize}} \quad \frac{1}{m}\sum_{1=1}^{m}(1 - y_i(w.x_i))_+ + \lambda \left\|w\right\|^2$$

where $(x)_+ = max(0, x)$.

Theorem 1. . *Let w_S be the output of soft margin SVM on $S = [(x_1, y_1), ..., (x_m, y_m)]$ generated from Distribution D. Then, $\forall \delta > 0$, $w.p. \leq 1 - \delta$,*

$$err(w_S) \leq margin_{err}(w_S) + 2\beta + (4\beta m + 2)\sqrt{\frac{log(\frac{1}{\delta})}{2m}}$$

, where $\beta = \frac{1}{\lambda m}$.

Here $margin_{error}(w_S) = \frac{1}{m}\sum_{i=1}^{m} \epsilon_i$, which is an upper bound on the empirical error.

Proof Sketch: We will use the notion of *Algorithmic Stability* in order to prove the theorem. The proof will involve three main steps:

1. We will first show that Soft-margin SVM is β-stable. Intuitively this means that small changes in the training set do not change the classifier output by the algorithm too much.
2. We will look at $F(S) = err(w_S) - margin_{err}(w_S)$[1] and use the fact that the algorithm is stable to show that
   - $E[F(S)] \leq 2\beta$
   - $F(S)$ is $(4\beta + \frac{2}{m})$-Lipschitz and then use McDiarmid's Inequality.

On to the formal proof now. First we need to formally define *algorithmic stability*.

Definition 1 (Stability)**.** *Consider a learning algorithm A that takes as input a set $S = [(x_1, y_1), ..., (x_m, y_m)]$, and outputs a function $w_S : X \mapsto \Re$. Let $S^{\setminus i} = [(x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ...]$, i.e., the set S with the ith example removed. Let $w_{S^{\setminus i}}$ be the output of the algorithm on $S^{\setminus i}$. We say that A is β-stable if $\forall x \in X, \forall i \in [m], \forall S \subset X^m$, then $|w_S(x) - w_{S^{\setminus i}}(x)| \leq \beta$.*

Next we will prove a general theorem about the stability of regularized minimization procedures.

Theorem 2. *Suppose A works as follows: Output $w_S = argmin_w \frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i, y_i) + \lambda \|w\|^2$ Furthermore assume that*

- *l is convex in first argument.*
- *l is σ-lipschitz: $|l(a, y) - l(b, y)| \leq \sigma |a - b|, \forall a, b \in \Re$.*

Then A is β-stable for $\beta = \frac{\sigma R^2}{\lambda m}$, where $R = max_{i \in X} \|x_i\|$.

In other words, strong convexity of the objective function implies the stability of the minimization algorithm. Soft-margin SVM satisfies both the conditions:

- $l(w \cdot x_i, y_i) = max(0, 1 - y_i(w \cdot x_i))$

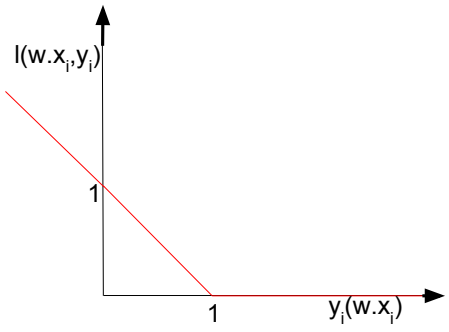

Figure 2: *Loss function of the Soft-margin SVM is convex in first argument.*

- $l(a, y) = max(0, 1 - ay)$ and $l(a, y) = max(0, 1 - by)$, then $|l(a, y) - l(b, y)| \leq y |a - b|$ (l is 1-lipschitz)

[1] Actually we will look at a clipped version since $F(S)$ is unbounded.

Proof. Let $R(w) = \frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i, y_i)$, $R^{\setminus i}(w) = \frac{1}{m}\sum_{j=1,j\neq i}^{m} l(w \cdot x_j, y_j)$. As l is convex, then $l(\frac{(w_1+w_2)}{2} \cdot x, y) \leq \frac{1}{2}l(w_1 \cdot x, y) + \frac{1}{2}l(w_2 \cdot x, y)$. As a sum of convex functions is convex function, we get

$$R(\frac{w_1+w_2}{2}) \leq \frac{1}{2}R(w_1) + \frac{1}{2}R(w_2) \tag{1}$$

$$R^{\setminus i}(\frac{w_1+w_2}{2}) \leq \frac{1}{2}R^{\setminus i}(w_1) + \frac{1}{2}R^{\setminus i}(w_2) \tag{2}$$

From, $w_1 = argmin_w \frac{1}{m}\sum_{i=1}^{m} l(w \cdot x_i, y_i) + \lambda \|w\|^2$ and $w_2 = argmin_w \frac{1}{m}\sum_{j=1,j\neq i}^{m} l(w \cdot x_j, y_j) + \lambda \|w\|^2$, then:

$$R(\frac{w_1+w_2}{2}) + \lambda \left\|\frac{w_1+w_2}{2}\right\|^2 \geq R(w_1) + \lambda \|w_1\|^2 \tag{3}$$

$$R^{\setminus i}(\frac{w_1+w_2}{2}) + \lambda \left\|\frac{w_1+w_2}{2}\right\|^2 \geq R^{\setminus i}(w_2) + \lambda \|w_2\|^2 \tag{4}$$

If we sum Eq.(3),Eq.(4),-Eq.(1),-Eq.(2), then we will have the following:

$$2.\frac{\lambda}{4}\|w_1+w_2\|^2 \geq \frac{R(w_1)}{2} - \frac{R(w_2)}{2} - (\frac{R^{\setminus i}(w_1)}{2} - \frac{R^{\setminus i}(w_2)}{2})$$

$$\begin{aligned}\frac{R^{\setminus i}(w_1)}{2} - \frac{R^{\setminus i}(w_2)}{2} - (\frac{R(w_1)}{2} - \frac{R(w_2)}{2}) &\geq \lambda\|w_1\|^2 + \lambda\|w_2\|^2 - \frac{\lambda}{2}(\|w_1\|^2 + \|w_2\|^2 + 2w_1 \cdot w_2) \\ &\geq \frac{\lambda}{2}\|w_1 - w_2\|^2\end{aligned}$$

If we rearrange the left hand side:

$$\begin{aligned}\frac{R^{\setminus i}(w_1)}{2} - \frac{R(w_1)}{2} - \frac{R^{\setminus i}(w_2)}{2} - \frac{R(w_2)}{2} &\geq -\frac{1}{2m}l(w_1 \cdot x_i, y_i) + \frac{1}{2m}l(w_2 \cdot x_i, y_i) \\ \frac{\lambda}{2}\|w_1 - w_2\|^2 &\leq \frac{1}{2m}|l(w_1 \cdot x_i, y_i) - l(w_2 \cdot x_i, y_i)|\end{aligned}$$

Given l is σ-lipschitz, then:

$$\begin{aligned}\frac{\lambda}{2}\|w_1 - w_2\|^2 &\leq \frac{\sigma}{2m}|w_2 \cdot x_i - w_1 \cdot x_i| \\ &\leq \frac{\sigma R}{2m}\|w_1 - w_2\| \\ &\downarrow \\ \|w_1 - w_2\| &\leq \frac{\sigma R}{\lambda m} \\ &\downarrow \\ |w_1 \cdot x - w_2 \cdot x| \leq \|w_1 - w_2\|\|x\| &\leq \frac{\sigma R^2}{\lambda m}\end{aligned}$$

□

We proved so far that given $A(S) = argmin_w \frac{1}{m}\sum_{i=1}^{m}(1 - y_i(wx_i)) + \lambda \|w\|^2$, then A is β-stable where $\beta = \frac{1}{\lambda m}$[2]. Now, we want to prove Theorem 2.

Instead of working with the original loss function, we will work with a clipped version as our loss can be unbounded.

Define

$$\begin{aligned} l_T(w \cdot x, y) &= 1 \ \text{ if } y(w \cdot x) < 0 \\ &= 1 - y(w.x) \ \text{ if } 1 \geq y(w \cdot x) \geq 0 \\ &= 0 \ \text{ otherwise} \end{aligned}$$

It is easy to check that $E_{x,y}[l_T(w_s \cdot x, y)] \geq err(w_s)$ and $\frac{1}{m}\sum_{i=1}^{m} l_T(w_s \cdot x, y) \geq margin_{error}(w_s)$, it is enough to prove $w.p. \geq 1 - \delta$ that

$$E_{x,y}[l_T(w_s \cdot x, y)] \leq \frac{1}{m}\sum_{i=1}^{m} l_T(w_s \cdot x, y) + 2\beta + (4m\beta + 2)\sqrt{\frac{log(\frac{1}{\delta})}{2m}}$$

Let $F((x_1, y_1), (x_2, y_2), ..., (x_m, y_m)) = E_{x,y}[l_T(w_s \cdot x, y)] - \frac{1}{m}\sum_{i=1}^{m} l_T(w_s \cdot x_i, y_i)$, we want to prove:

1. $E_S(F) \leq 2\beta$
2. F is $(4\beta + \frac{2}{m}) - lipschitz$

First, we prove $E_S(F) \leq 2\beta$:

Proof.

$$\begin{aligned} E_S[F] &= E_S[E_{x,y}[l_T(w_S \cdot x, y)] - \frac{1}{m}\sum_{i=1}^{m} l_T(w_S \cdot x_i, y_i)] \\ &= E_S[E'_S[\frac{1}{m}\sum_{i=1}^{m} l_T(w_S \cdot x_i^{'}, y_i^{'}) - \frac{1}{m}\sum_{i=1}^{m} l_T(w_S \cdot x_i, y_i)]] \\ &= E_{S,S'}[\sum_{i=1}^{m}\frac{1}{m}(l_T(w_S \cdot x_i^{'}, y_i^{'}) - l_T(w_S \cdot x_i, y_i))] \\ &= E_{S,S'}[\sum_{i=1}^{m}\frac{1}{m}(\underbrace{l_T(w_S \cdot x_i^{'}, y_i^{'}) - l_T(w_{S^{\setminus i}} \cdot x_i^{'}, y_i^{'})}_{\leq \beta} + l_T(w_{S^{\setminus i}} \cdot x_i^{'}, y_i^{'}) \\ &- l_T(w_{S^{\setminus i}} \cdot x_i, y_i) + \underbrace{l_T(w_{S^{\setminus i}} \cdot x_i, y_i) - l_T(w_S \cdot x_i, y_i))}_{\leq \beta}] \\ &\leq 2\beta + \underbrace{E_{S,S'}[\sum_{i=1}^{m}\frac{1}{m}(l_T(w_{S^{\setminus i}} \cdot x_i^{'}, y_i^{'}) - l_T(w_{S^{\setminus i}} \cdot x_i, y_i))]}_{0} \\ &\leq 2\beta \end{aligned}$$

□

[2]From now on we assume that $\|x_i\| \leq 1$

Next we prove that $F()$ is $(4\beta + \frac{2}{m})$-Lipschitz.

Proof. Let's look at the first term of $F(S)$, i.e. $E_{x,y}[l_T(w_S \cdot x, y)]$. We will show that the first term is β-Lipschitz. Define $S^i = \{(x_1, y_1), \ldots (x_{i-1}, y_{i-1}, (x'_i, y'_i), (x_{i+1}, y_{i+1}, \ldots (x_m, y_m)))\}$. Then,

$$E[l_T(w_S \cdot x, y) - l_T(w_{S^i} \cdot x, y)] = E[l_T(w_S \cdot x, y) - l_T(w_{S^{\setminus i}} \cdot x, y)] + E[l_T(w_{S^{\setminus i}} \cdot x, y) - l_T(w_{S^i} \cdot x, y)]$$

By Stability, both terms are bounded by β.

Next we show that the second term is $(2\beta + \frac{2}{m})$-Lipschitz. As above, the difference in the second term computed over S and S^i can be written as

$$\frac{1}{m}\sum_{i=1}^{m}[l_T(w_S \cdot x_i, y_i) - l_T(w_{S^{\setminus i}} \cdot x_i, y_i)] + \frac{1}{m}[l_T(w_{S^{\setminus i}} \cdot x'_i, y'_i) - l_T(w_{S^i} \cdot x'_i, y'_i)]$$

$$+\frac{1}{m}[l_T(w_{S^i} \cdot x'_i, y_i) - l_T(w_{S^i} \cdot x_i, y_i)] + \frac{1}{m}[l_T(w_{S^{\setminus i}} \cdot x_i, y_i) - l_T(w_{S^{\setminus i}} \cdot x'_i, y'_i)]$$

By stability, the first two terms are bounded by 2β. By boundedness of l_T the second two terms are bounded by $\frac{1}{m}$. Hence, overall the function is $(4\beta + \frac{2}{m})$-Lipschitz. □

Now we finish the proof by using McDiarmid's inequality. We know that w.p. $\geq 1 - \delta$

$$|E[F(S)] - F(S)| \leq c\sqrt{\frac{m \log \frac{1}{\delta}}{2}}$$

where c is the Lipschizness constant of the function. Substituting $c = (4\beta + \frac{2}{m})$ gives the result.

3 Additional Readings

- `http://www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf`.
- Chapter 13 in the book: `http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf`.