CSC 588: Machine learning theory **Spring 2021**

Lecture 15: l_2 Norm-based Margin Bounds; Regularized Loss Minimization Formulations

Lecturer: Chicheng Zhang *Scribe: Shahriar Golchin*

1 l_2-SVM and Its Statistical Properties

In this lecture, we are going to talk about margin-based generalization error bounds for l_2-SVMs. From the last session, we recall that the l_2-SVM is formulated as followings:

$$\min_{w} \quad ||w||_2$$
$$\text{s.t. } \forall i \quad y_i \langle w, x_i \rangle \geq 1$$

We can rewrite the aforesaid formula in a different form which is more intuitive. To do so, first, we need to change objective $||w||_2$. We can do this by changing $||w||_2$ to any monotonic function with respect to l_2 norm like $\frac{1}{2}||w||_2^2$. It is worth noting that it doesn't change the optimization problem as it is still minimizing the l_2 norm. Second, we need to introduce two new parameters; α and $\hat{w}$.

Actually, α captures the magnitude of w, and $\hat{w}$ captures the direction of w. So, we can rewrite w as $w = \alpha\hat{w}$ where $\alpha > 0$ and $||\hat{w}|| = 1$.

Therefore, now, we can rewrite the formula as below:

$$\min_{\alpha, \hat{w} \,:\, \alpha > 0, \, ||\hat{w}|| = 1} \alpha$$
$$\text{s.t. } \forall i \quad y_i \alpha \langle \hat{w}, x_i \rangle \geq 1$$

As the next step, we're trying to change the minimization problem to maximization problem. So, we can write the equivalent format of above formula:

$$\max_{\alpha, \hat{w} \,:\, \alpha > 0, \, ||\hat{w}|| = 1} \frac{1}{\alpha}$$
$$\text{s.t. } \forall i \quad y_i \alpha \langle \hat{w}, x_i \rangle \geq \frac{1}{\alpha}$$

For any fixed $\hat{w}$, the optimal choice of α is such that

$$\frac{1}{\alpha} = \min_{i} \quad y_i \langle \hat{w}, x_i \rangle$$

So, based on this observation, we can eliminate α by replacing its optimal choice:

$$\max_{\hat{w} : ||\hat{w}|| = 1} \quad \min_{i} \quad y_i \langle \hat{w}, x_i \rangle$$

If we can solve this optimization problem, then we can also solve the original optimization problem.

The intuition of this optimization problem is very similar to the Figure 1 which we saw previously. Given Figure 1, we want to identify a direction of $\hat{w}$ such that the minimization margin be as large as possible. Based on the optimization problem, first, we want all examples be classified correctly, and on the other hand, we also want all the projections' values be as large as possible. Therefore, purple $\hat{w}$ is more preferable than green $\hat{w}$.

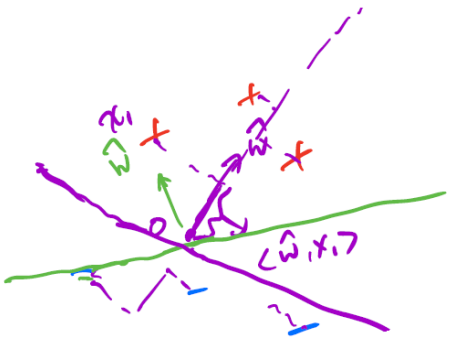

Figure 1: An intuition of the optimization problem

2 Generalization Error Bounds for l_2-SVMs (l_2-bounded linear predictors)

Theorem 1. *Introducing the set up: fix $B_2R_2 > 0$, $S = (x_1, y_1)...(x_n, y_n) \sim D$, which D is supported on $\{x \in \mathbb{R}^d : ||x||_2 \leq R_2\} \times \{\pm 1\}$, $\theta \in (0, B_1R_2]$. Then, with probability $1 - \delta$ for all $w : ||w_2|| \leq B_2$:*

$$\mathbb{P}_D(y\langle w, x\rangle \leq 0) \;\leq\; \mathbb{P}_S(y\langle w, x\rangle \leq \theta) + \mathcal{O}\left(\frac{B_2R_2}{\theta}\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right)$$

Proof. We use the same strategy as the l_1/l_∞ margin bound:

1. introduce ramp loss
2. uniform concentration of ramp losses
3. Contraction inequality of Rademacher complexity

Proof comes down to show that given $S = (x_1, y_1)...(x_n, y_n)$ such that $\forall i\ ||x||_2 \leq B_2$, and $\mathcal{G} = \{m_w : ||w||_2 \leq B_2\}$, where $m_w(x, y) = y\langle w, x\rangle$.

We want to bound the following formula:

$$\text{Rad}_S(\mathcal{G}) = \frac{1}{m}\,\mathbb{E}_{\sigma\sim U(\pm 1)^m} \sup_{w:||w||_2\leq B_2} \sum_{i=1}^{m} \sigma_i y_i \langle w, x_i\rangle$$

As examples are considered as fixed, so $\sigma_1 y_1$ ... $\sigma_m y_m$ all have the same distribution as σ_1 ... σ_n, so

$$\text{Rad}_S(\mathcal{G}) = \frac{1}{m}\,\mathbb{E}_{\sigma\sim U(\pm 1)^m} \sup_{w:||w||_2\leq B_2} \sum_{i=1}^{m} \sigma_i \langle w, x_i\rangle$$

Next, we try to remove sup. To do so, we use linearity property and move summation into the inner product and try to upper bound it.

$$\sum_{i=1}^{m} \sigma_i \langle w, x_i\rangle = \langle w, \sum_{i=1}^{m} \sigma_i x_i\rangle$$

If w is l_2-bounded, then we can use Cauchy-Schwarz inequality:

$$\begin{aligned}\langle w, \sum_{i=1}^{m} \sigma_i x_i\rangle &\leq ||w||_2\, ||\sum_{i=1}^{m} \sigma_i x_i||_2 \\ &\leq B_2 ||\sum_{i=1}^{m} \sigma_i x_i||_2\end{aligned}$$

So,

$$\text{Rad}_S(\mathcal{G}) \;\leq\; \frac{B_2}{m}\,\mathbb{E}_\sigma\,||\sum_{i=1}^{m}\sigma_i x_i||_2$$

Let's say $||\sum_{i=1}^{m}\sigma_i x_i||_2$ is like a random variable Z and using the fact that $(\mathbb{E}\,Z)^2 \leq \mathbb{E}(Z^2)$:

$$\leq \frac{B_2}{m}\sqrt{\mathbb{E}_\sigma\,||\sum_{i=1}^{m}\sigma_i x_i||_2^2}$$

$$= \frac{B_2}{m}\sqrt{\mathbb{E}_\sigma[\langle\sum_{i=1}^{m}\sigma_i x_i, \sum_{j=1}^{m}\sigma_j x_j\rangle]}$$

Using linearity of inner product:

$$= \frac{B_2}{m}\sqrt{\mathbb{E}_\sigma[\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_i\sigma_j\langle x_i, x_j\rangle]}$$

Moving $\mathbb{E}$ inside the summation:

$$= \frac{B_2}{m}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}_\sigma[\sigma_i\sigma_j]\langle x_i, x_j\rangle}$$

Given that

$$\mathbb{E}_\sigma[\sigma_i\sigma_j] = \begin{cases} 1 & i = j \\ 0 & otherwise \end{cases} \tag{1}$$

So,

$$= \frac{B_2}{m}\sqrt{\sum_{i=1}^{m}\langle x_i, x_i\rangle}$$

which $\langle x_i, x_i\rangle = ||x_i||_2^2 \leq R_2^2$, and finally,

$$\leq \frac{B_2}{m}\sqrt{mR_2^2} = \frac{B_2 R_2}{\sqrt{m}}$$

□

3 Comparison between l_1/l_∞ and l_1/l_2 Margin Bounds

Table 1 shows a comparison between l_2/l_∞ and l_2/l_2 Margin Bounds. Note: It turns out that these bounds are incomparable in general.

Table 1: A comparison between margin bounds

	Constraint on $\mathcal{X}$	**Constraint on w**	**Error Bound**
l_1/l_∞	$\|\|x\|\|_\infty \leq R_\infty$	$\|\|w\|\|_1 \leq B_1$	$\widetilde{\mathcal{O}}(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{1}{m}})$
l_2/l_2	$\|\|x\|\|_2 \leq R_2$	$\|\|w\|\|_2 \leq B_2$	$\widetilde{\mathcal{O}}(\frac{B_2 R_2}{\theta}\sqrt{\frac{1}{m}})$

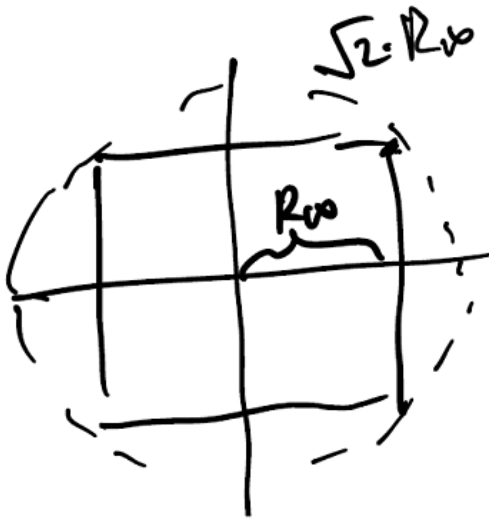

Figure 2: Illustration of 2D examples

4 Exercise

Applying l_2/l_2 generalization bound to l_1/l_∞ setting; How can we pick R_2 such that $||x||_\infty \leq R_\infty \Rightarrow ||x||_2 \leq R_2$?

Let's see Figure 2 that shows 2D examples for having more insights: It turns out that if we pick the top right corner of the box, then examples in that point have the largest l_2 norm. So, the l_2 norm of that point based on the Figure 2 is $\sqrt{2}R_\infty$. In general, in d dimensions, that corner has the larger and larger l_2 norm. Therefore, the tightest bound we can choose is $R_2 = \sqrt{d}R_\infty$.

How about B_2? For choosing B_2, let's take a look at Figure 3. As we can see in Figure 3, if we choose $B_2 = B_1$, then it would be the best choice of l_2 radius. Also, choosing $B_2 = B_1$ would be sufficient considering the fact that $||w||_2 \leq ||w||_1$.

So, with these choices and putting them into l_2/l_2 bound, we will get a generalization bound in terms of $R_2 B_2 = \sqrt{d}R_\infty B_1$, a factor of $\sqrt{d}$ worse than the original l_1/l_∞ bound.

More Exercise Applying l_1/l_∞ bound to the l_2/l_2 setting (will be $\sqrt{d}$ factor worse).

5 Coping with Linear Non-separability in SVMs

The idea is to first introduce nonlinear feature maps (or basis functions), and second, relax the SVM formulation; allowing some examples to be incorrectly classified. These two ideas will be discussed in the followings.

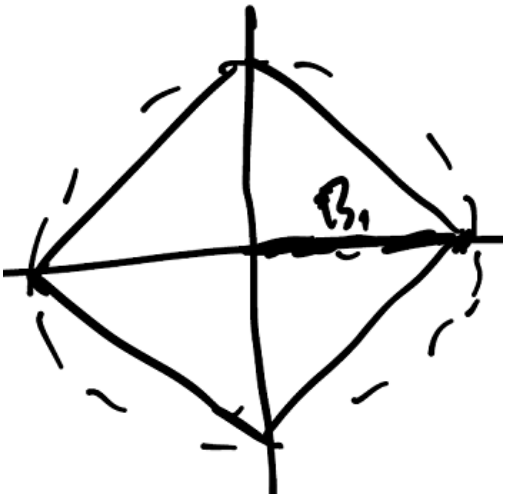

Figure 3: Illustration of 2D linear predictors

1. **Introducing Nonlinear Feature Maps (Basis Functions)**

As a motivating example, first, consider the following example shown in Figure 4. We assume that data is 2D and all positive examples are in a unit circle (left-hand side drawing), and negative examples are outside of circle. As we can see, no linear classifier can behave well. If we consider a candidate linear predictor like

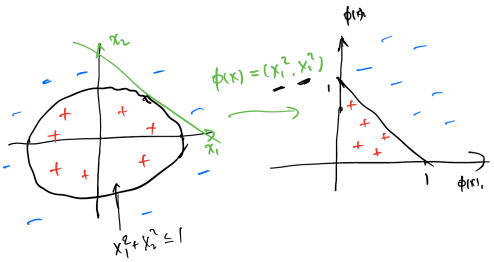

Figure 4: Example of nonlinear feature map

the green one in the drawing, then it would miss a lot of negative examples. So, we can define a feature map like $\phi(x) = (x_1^2, x_2^2)$. After applying this transformation, all positive examples will be mapped within the triangle, and negative examples outside of it. Therefore, after this transformation, data becomes more amendable for linear classification.

To summarize what we did for nonlinear feature maps, consider the following steps:

1. Define $\phi : \mathbb{R}^d \Rightarrow \mathbb{R}^m, (x_i, y_i) \Rightarrow (\phi(x_i), y_i)$
2. Solve SVM on $(\phi(x_i), y_i) \Rightarrow \hat{w} \in \mathbb{R}^m$
3. Find predictor: $sign(\langle \phi(x), \hat{w} \rangle)$

Note: there are SVM training algorithms that have time complexity independent of m. Also, if $\langle \phi(x), \phi(y) \rangle$ can be evaluated in time dependent of m, then it is called *Kernel trick*.

2. **Soft Margin SVMs**

As we know, the original SVM has the following formulas:

$$\min_{w} \quad \frac{1}{2}||w||_2^2$$
$$\text{s.t. } \forall i \quad y_i\langle w, x_i\rangle \geq 1$$

As discussed, this formula may not be feasible. One way to make it feasible is to add some slack variables ξ:

$$\min_{w,(\xi_1 \ldots \xi_m)\geq 0} \frac{\lambda}{2}||w||_2^2 + \sum_{i=1}^{m} \xi_i$$
$$\text{s.t. } \forall i \quad y_i\langle w, x_i\rangle \geq 1 - \xi_i$$

Also, we added λ as a coefficient to trade off between $||w||_2^2$ and $\sum_{i=1}^{m} \xi_i$. So, if λ is smaller, then more likely to classify more correctly, but $||w||$ will likely to be large.

If we want to eliminate ξ, for any fixed w, the optimal choices if ξ_i are $\xi_i \geq 1 - y_i\langle w, x_i\rangle$ and $\xi_i \geq 0$ which yield to $\xi_i \geq \max(0, 1 - y_i\langle w, x_i\rangle)$ and vice versa. So, the optimal choices are $\xi_i = \max(0, 1 - y_i\langle w, x_i\rangle)$.

Therefore, what is left is:

$$\min_{w} \quad \frac{\lambda}{2}||w||_2^2 + \sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle)$$

The above formula is a bit like biased complexity trade off that we have seen in the model selection lecture. We can say $\frac{\lambda}{2}||w||_2^2$ is as complexity of linear predictor and $\sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle$ as empirical risk. We can write $\sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle$ as the followings:

$$\sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle) = \phi(y_i\langle w, x_i\rangle)$$

where $\phi(Z) = \max(0, 1 - Z)$.

As Figure 5 shows, it's like a hinge, so it is called Hinge loss.

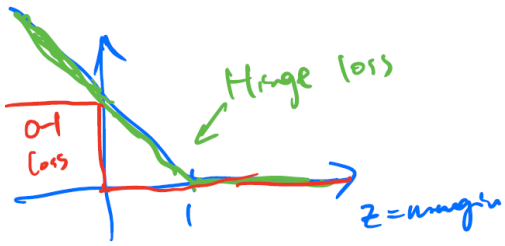

Figure 5: Hinge loss vs 0-1 loss

We can write more general form of previous formula as below which is called as regularized loss minimization:

$$\min_{w} \quad \lambda R(w) + \sum_{i=1}^{m} \phi(f_w(x_i), y_i)$$

where we can choose any $R(w)$ and $\phi(f_w(x_i), y_i)$.

6 Some Notable Examples

1. If we define $R(w) = \frac{1}{2}||w||_2^2$, $\phi(f_w(x_i), y_i) = (f_w(x) - y)^2$, $f_w(x) = \langle w, x \rangle$, then for $\lambda = 0$ it corresponds to "ordinary least squares", and for $\lambda > 0$ it corresponds to "ridge regression".
2. If we define $R(w) = ||w||_1$, then it corresponds to "LASSO".
3. If we define $R(w) = ||w||_2^2$, $\phi(f_w(x_i), y_i) = \ln(1 + \exp(-y_i f_w(x_i))$, $f_w(x) = \langle w, x \rangle$, then it corresponds to "logistic regression". If we define $y_i f_w(x_i)$ as Z, then logistic regression will have the shape like Figure 6.

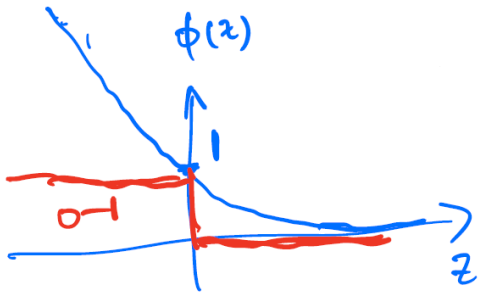

Figure 6: Logistic loss vs 0-1 loss

In the next class, we will talk about regularized loss minimization's solutions generalization performance from stability view.