COS-511: Learning Theory **Spring 2017**

Lecture 10: Surrogate Loss functions & Rademacher Complexity

Lecturer: Roi Livni

10.1 Surrogate Loss Functions

Definition 10.1. *A loss function ℓ is called a surrogate loss function (speicificially surrogate loss for the $0-1$ loss) if*

1. $\ell \geq \ell_{0,1}$.
2. ℓ *is convex.*

Fact 10.1. *Let ℓ be a convex loss function then for any target function h and any distribution D:*

$$err(h) \leq \mathbf{E}\left[\ell(h(\mathbf{x}), y)\right]$$

It is not hard to see that fact 10.1 is true. The implication of this fact is that if we can minimize a surrogate loss function and achieve small loss – then we obtain a target function with small zero one error. Of course, if we fail to find a target function with small loss, there is no guarantee there is no solution with small zero one error.

Example 10.1 (SVM)**.** *The support vector machine algorithm is concerened with finding the linear classifier with largest margin, as we will see this can be formulized as minimizing a surrogate loss function.*

For the realizable case, the objective of Hard-SVM may be written as

$$\begin{array}{ll} \text{minimize} & \|\mathbf{w}\|^2 \\ \textit{subject to} & y_i \mathbf{w} \cdot \mathbf{x}_i \geq 1. \forall i = 1, \dots m. \end{array}$$

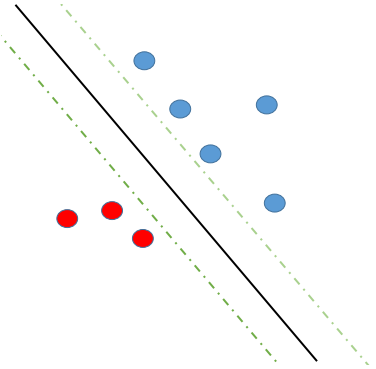
By choosing sufficiently small λ this can be written as

$$\begin{array}{ll} \text{minimize} & \lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum \xi_i \\ \text{subject to} & \xi_i \geq 0 \\ & \xi_i \geq 1 - y_i \mathbf{w} \cdot \mathbf{x}^{(i)} \end{array}$$

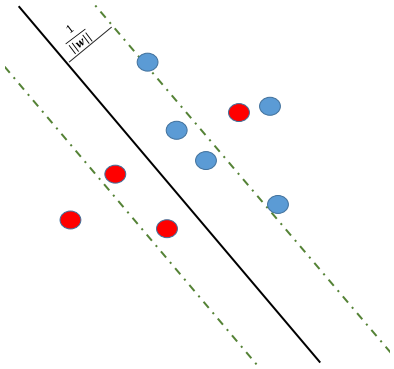
Which we can re-write as

$$\text{minimize} \qquad \lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum \max(0, 1 - y_i \mathbf{w} \cdot \mathbf{x}^{(i)})$$

Let us denote $\ell_{\text{hinge}}(a, y) = \max(0, 1 - y \cdot a)$ Then we obtain the regulerized SVM formulation



(a) The Hard-SVM setup when points are sepearable. The Hard SVM algorithm aims at finding a classifier that separates the norm with at least $\mathbf{w} \cdot \mathbf{x} \geq 1$. The dashed line represent that domain where $\mathbf{w} \cdot \mathbf{x} = 1$



(b) The Soft-SVM aims at finding a solution that minimizes the hinge loss with additional norm regulerization. The solution may include points inside the margin (where the hinge loss penalizes even correctly classified points) for minimizing the loss on misclassified points.

$$\text{minimize} \qquad \lambda\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell_{\text{hinge}}(\mathbf{w} \cdot \mathbf{x}^{(i)}, y_i)$$

Note that the above problem, is well defined even for large λ (even though it will not necessarily converge to a realizable solution, even if one exists). Also note that ℓ_{hinge} is a surrogate loss function. We may also want to consider the constrained version.

10.2 Rademacher Complexity

So far we have developed a tool to analyze the sample complexity of binary classifiers (VC–dimension). This however does not fit our model of learning convex problems. We thus, need to develop a new mechanism to analyse the sample complexity of convex learning problems. The Rademacher complexity is a tool to analyse how well we can approximate the mean of a class of target functions through an empirical sample. Namely, given a distribution D and a sample $S = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ denote:

$$\mathcal{L}_D(f) = \mathbf{E}_{\mathbf{z}\sim D}\left[f(\mathbf{z})\right], \quad \mathcal{L}_S = \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{z}^{(i)})$$

Given a class of target functions $\mathcal{F}$ we want to have a bound w.h.p w.r.t S over

$$\sup_{f\in\mathcal{F}} |\mathcal{L}_D(f) - \mathcal{L}_S(f)|$$

Such a bound will entail the uniform convergence property. We will then apply this bound to class of target functions of the form $\mathcal{F} = \{f : f = \ell(h(\mathbf{x}_i), y_i), h \in \mathcal{C}\}$, to achieve guarantees for learning problems.

Our key tool for obtaining the uniform convergence property is through the Rademacher complexity which we next define. The idea behind the Rademacher complexity is as follows: Suppose we are given a sample S and we divide it randomly to a train set of points S_{train}, and test set S_{test}: What could go wrong, if we

train an algorithm by observing S_{train} and then testing it on S_{test}? Roughly, thing might go wrong, if for some function, the discrepency:

$$\sum_{\mathbf{z}\in S_{\text{train}}} f(\mathbf{z}) - \sum_{\mathbf{z}\in S_{\text{test}}} f(\mathbf{z})$$

is large.

This is exactly what the Rademacher complexity measures: We consider a random partition of a set into two distinct sets and we evaluate the discrepency w.r.t to the worst case function in $\mathcal{F}$. To define the Rademacher Complexity we begin with a simple definition of a Rademacher random variable:

Definition 10.2 (Rademacher random variables)**.** *Let σ be a vector whose elements are chosen independently and uniformly from $\{-1, +1\}$. That is, with probability $1/2$ a given element is either -1 or 1.*

Definition 10.3 (Empirical Rademacher complexity)**.** *Given a sample $S = \{x_1, \ldots, x_m\}$ chosen from $\mathcal{D}^m$, define the empirical Rademacher complexity $\hat{\mathfrak{R}}_S(\mathcal{F})$ as*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\mathbf{z}^{(i)})\right|\right]$$

Definition 10.4 (Rademacher complexity)**.** *For some $m \geq 1$, let the Rademacher complexity of $\mathcal{F}$ be the expectation of the empirical Rademacher over all samples S of size m drawn from some distribution $\mathcal{D}$.*

$$\mathfrak{R}_m(\mathcal{F}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(\mathcal{F})\right]$$

Theorem 10.5. *Let $\mathcal{F}$ be a class of functions bounded by c and $S = \{\mathbf{z}^{(i)}\}_{i=1}^m$ a sample drawn IID then with probability $1-\delta$*

$$\sup_{f\in F} \mathcal{L}_D(f) - L_S(f) \leq 2\mathfrak{R}_m(\mathcal{F}) + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right)$$

Proof. Let us write $\Phi(S) = \sup_{f\in\mathcal{F}} L_D(f) - L_S(f)$. We will first bound the expectation of $\Phi(S)$ then using a concentration inequality we will bound $\Phi(S)$ w.h.p:

$$\begin{aligned}
\mathbb{E}[\Phi(S)] &= \mathop{\mathbb{E}}_{S}[\sup_{f\in\mathcal{F}}\{\mathcal{L}_D(f) - \mathcal{L}_S(h)\}] && \text{by definition}\\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{h\in\mathcal{H}}\left\{\left[\mathop{\mathbb{E}}_{S'}\mathcal{L}_{S'}(f) - \mathcal{L}_S(h)\right]\right\}\right] && \text{expectation of i.i.d. sample error}\\
&\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{h\in\mathcal{H}}\{\mathcal{L}_{S'}(f) - \mathcal{L}_S(f)\}\right] && \text{Jensen's and convexity of supremum}\\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_{i=1}^{m}\left[f(\mathbf{z}'^{(i)}) - f(\mathbf{z}^{(i)})\right]\right\}\right] && \text{by definition}
\end{aligned}$$

Now suppose we generate the sample S, S' as follows, after generating S, S' for every i with probaiblity $1/2$ we set $\mathbf{z}'^{(i)} \in S'$ and $\mathbf{z}^{(i)} \in S$, but with probability $1/2$ we alternate and let $\mathbf{z}^{(i)} \in S'$ then clearly the sample

S, S' are still generated by the same distribution. Therefore we have:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[f(\mathbf{z}'^{(i)})-f(\mathbf{z}^{(i)})\right]\right\}\right] && \sigma \text{ doesn't change } \mathbb{E}\\
&\leq \mathop{\mathbb{E}}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[f(\mathbf{z}'^{(i)})\right]\right|\right\}\right] + \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[f(\mathbf{z}^{(i)}\right]\right|\right\}\right] && \text{sub-additivity of supremum}\\
&= 2\mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[f(\mathbf{z}^{(i)}\right]\right|\right\}\right] && \sigma_i \text{ and } -\sigma_i \text{ distributed same way}\\
&= 2\mathfrak{R}_m(\mathcal{F})\ \square
\end{aligned}
$$

So far we've bounded the generalization error in expectation, to derive high probability bounds we rely on the following inequality

Lemma 10.6 (**Mcdiarmid Inequality**). *: Let X be some set and $f : X^m \to \mathbb{R}$. be a function of m variables such that for some $a > 0$, for all $i \in [m]$ and for all $x_1, \ldots, x_m, x_i' \in X$ we have*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq a.$$

Let $X_1, \ldots, X_m$ be IID random varaibles taking value at X then w.p at least $1 - \delta$:

$$|f(X_1, \ldots, X_m) - \mathbb{E}\left[f(X_1, \ldots, X_m)\right]| \leq a\sqrt{\ln(2/\delta)m/2}$$

□

To apply Mcdiarmind's inequality note that $\Phi(S) - \Phi(S') \leq \frac{2c}{m}$, hence we obtain the w.p at least $1 - \delta$

$$\sup_{f\in F}\mathcal{L}_D(f) - L_S(f) \leq \mathbb{E}\left[\Phi(S)\right] + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right) \leq 2\mathfrak{R}_m(\mathcal{F}) + O\left(c\sqrt{\frac{\ln 1/\delta}{m}}\right)$$