18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Ali Makhdoumi

Lecture 6
Sep. 28, 2015

5. LEARNING WITH A GENERAL LOSS FUNCTION

In the previous lectures we have focused on binary losses for the classification problem and developed VC theory for it. In particular, the risk for a classification function $h : \mathcal{X} \to \{0, 1\}$ and binary loss function the risk was

$$R(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbb{1}(h(X) \neq Y)].$$

In this lecture we will consider a general loss function and a general regression model where Y is not necessarily a binary variable. For the binary classification problem, we then used the followings:

- Hoeffding's inequality: it requires boundedness of the loss functions.
- Bounded difference inequality: again it requires boundedness of the loss functions.
- VC theory: it requires binary nature of the loss function.

Limitations of the VC theory:

- Hard to find the optimal classification: the empirical risk minimization optimization, i.e.,
$$\min_h \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i)$$
is a difficult optimization. Even though it is a hard optimization, there are some algorithms that try to optimize this function such as Perceptron and Adaboost.
- This is not suited for regression. We indeed know that classification problem is a subset of Regression problem as in regression the goal is to find $\mathbb{E}[Y|X]$ for a general Y (not necessarily binary).

In this section, we assume that $Y \in [-1, 1]$ (this is not a limiting assumption as all the results can be derived for any bounded Y) and we have a regression problem where $(X, Y) \in \mathcal{X} \times [-1, 1]$. Most of the results that we preset here are the analogous to the results we had in binary classification. This would be a good place to review those materials and we will refer to the techniques we have used in classification when needed.

5.1 Empirical Risk Minimization

5.1.1 Notations

Loss function: In binary classification the loss function was $\mathbb{1}(h(X) \neq Y)$. Here, we replace this loss function by $\ell(Y, f(X))$ which we assume is symmetric, where $f \in \mathcal{F}$, $f : \mathcal{X} \to [-1, 1]$ is the regression functions. Examples of loss function include

- $\ell(a,b) = \mathbb{I}(a \neq b)$ (this is the classification loss function).
- $\ell(a,b) = |a-b|$.
- $\ell(a,b) = (a-b)^2$.
- $\ell(a,b) = |a-b|^p$, $p \geq 1$.

We further assume that $0 \leq \ell(a,b) \leq 1$.
Risk: risk is the expectation of the loss function, i.e.,

$$R(f) = \mathbb{E}_{X,Y}[\ell(Y, f(X))],$$

where the joint distribution is typically unknown and it must be learned from data.
Data: we observe a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent draws from a joint distribution $P_{X,Y}$, where $(X,Y) \in \mathcal{X} \times [-1,1]$. We denote the data points by $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
Empirical Risk: the empirical risk is defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)),$$

and the empirical risk minimizer denoted by $\hat{f}^{\text{erm}}$ (or $\hat{f}$) is defined as the minimizer of empirical risk, i.e.,

$$\operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_n(f).$$

In order to control the risk of $\hat{f}$ we shall compare its performance with the following oracle:

$$\bar{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f).$$

Note that this is an oracle as in order to find it one need to have access to P_{XY} and then optimize $R(f)$ (we only observe the data D_n). Since $\hat{f}$ is the minimizer of the empirical risk minimizer, we have that $\hat{R}_n(\hat{f}) \leq \hat{R}_n(\bar{f})$, which leads to

$$\begin{aligned} R(\hat{f}) &\leq R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\hat{f}) - \hat{R}_n(\bar{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) + R(\bar{f}) \\ &\leq R(\bar{f}) + R(\hat{f}) - \hat{R}_n(\hat{f}) + \hat{R}_n(\bar{f}) - R(\bar{f}) \leq R(\bar{f}) + 2\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|. \end{aligned}$$

Therefore, the quantity of interest that we need to bound is

$$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|.$$

Moreover, from the bounded difference inequality, we know that since the loss function $\ell(\cdot,\cdot)$ is bounded by 1, $\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$ has the bounded difference property with $c_i = \frac{1}{n}$ for $i = 1, \ldots, n$, and the bounded difference inequality establishes

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)\,| - \mathbb{E}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)\,|\right] \geq t\right] \leq \exp\left(\frac{-2t^2}{\sum_i c_i^2}\right) = \exp\left(-2nt^2\right),$$

which in turn yields

$$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|\right] + \sqrt{\frac{\log(1/delta)}{2n}}, \text{ w.p. } 1-\delta.$$

As a result we only need to bound the expectation $\mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|]$.

5.1.2 Symmetrization and Rademacher Complexity

Similar to the binary loss case we first use symmetrization technique and then introduce Rademacher random variables. Let $D_n = \{(X_1, Y_1), \dots (X_n, Y_n)\}$ be the sample set and define an independent sample (ghost sample) with the same distribution denoted by $D'_n = \{(X'_1, Y'_1), \dots (X'_n, Y'_n)\}$(for each i, (X'_i, Y'_i) is independent from D_n with the same distribution as of (X_i, Y_i)). Also, let $\sigma_i \in \{-1, +1\}$ be i.i.d. $\mathsf{Rad}(\frac{1}{2})$ random variables independent of D_n and D'_n. We have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}\left[\ell(Y_i, f(X_i))\right]|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|D_n\right]|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|D_n\right]|\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}\left[|\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|\ |D_n\right]\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \frac{1}{n}\sum_{i=1}^{n} \ell(Y'_i, f(X'_i))|\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i\left(\ell(Y_i, f(X_i)) - \ell(Y'_i, f(X'_i))\right)|\right] \\
&\overset{(c)}{\leq} 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Y_i, f(X_i))|\right] \\
&\leq 2\sup_{D_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i))|\right].
\end{aligned}
$$

where (a) follows from Jensen's inequality with convex function $f(x) = |x|$, (b) follows from the fact that (X_i, Y_i) and (X'_i, Y'_i) has the same distributions, and (c) follows from triangle inequality.

Rademacher complexity: of a class $\mathcal{F}$ of functions for a given loss function $\ell(\cdot, \cdot)$ and samples D_n is defined as

$$
\mathcal{R}_n(\ell \circ \mathcal{F}) = \sup_{D_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(y_i, f(x_i))|\right].
$$

Therefore, we have

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}[\ell(Y_i, f(X_i))]|\right] \leq 2\mathcal{R}_n(\ell \circ \mathcal{F})
$$

and we only require to bound the Rademacher complexity.

5.1.3 Finite Class of functions

Suppose that the class of functions $\mathcal{F}$ is finite. We have the following bound.

Theorem: Assume that $\mathcal{F}$ is finite and that ℓ takes values in $[0,1]$. We have

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

Proof. From the previous lecture, for $B \subseteq \mathbb{R}^n$, we have that

$$\mathcal{R}_n(B) = \mathbb{E}\left[\max_{b\in B} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i b_i|\right] \leq \max_{b\in B} |b|_2 \frac{\sqrt{2\log(2|B|)}}{n}.$$

Here, we have

$$B = \left\{ \begin{pmatrix} \ell(y_1, f(x_1)) \\ \vdots \\ \ell(y_n, f(x_n)) \end{pmatrix}, f \in \mathcal{F} \right\}.$$

Since ℓ takes values in $[0,1]$, this implies $B \subseteq \{b : |b|_2 \leq \sqrt{n}\}$. Plugging this bound in the previous inequality completes the proof. □

5.2 The General Case

Recall that for the classification problem, we had $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$. We have seen that the cardinality of the set $\{(f(x_1), \ldots, f(x_n)), f \in \mathcal{F}\}$ plays an important role in bounding the risk of $\hat{f}^{\text{erm}}$ (this is not exactly what we used but the XOR argument of the previous lecture allows us to show that the cardinality of this set is the same as the cardinality of the set that interests us). In this lecture, this set might be uncountable. Therefore, we need to introduce a metric on this set so that we can treat the close points in the same manner. To this end we will define covering numbers (which basically plays the role of VC dimension in the classification).

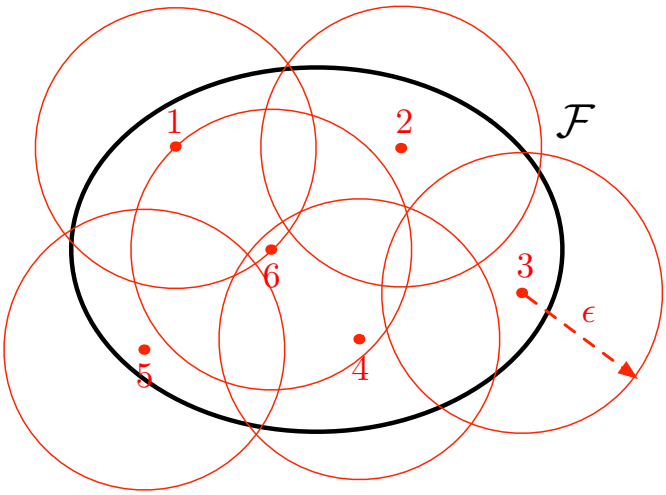
5.2.1 Covering Numbers

Definition: Given a set of functions $\mathcal{F}$ and a pseudo metric d on $\mathcal{F}$ ($(\mathcal{F}, d)$ is a metric space) and $\varepsilon > 0$. An *ε-net* of $(\mathcal{F}, d)$ is a set V such that for any $f \in \mathcal{F}$, there exists $g \in V$ such that $d(f,g) \leq \varepsilon$. Moreover, the *covering numbers* of $(\mathcal{F}, d)$ are defined by

$$N(\mathcal{F}, d, \varepsilon) = \inf\{|V| \,:\, V \text{ is an } \varepsilon\text{-net}\}.$$

For instance, for the $\mathcal{F}$ shown in the Figure 5.2.1 the set of points $\{1,2,3,4,5,6\}$ is a covering. However, the covering number is 5 as point 6 can be removed from V and the resulting points are still a covering.

Definition: Given $x = (x_1, \ldots, x_n)$, the *conditional Rademacher average* of a class of



functions $\mathcal{F}$ is defined as

$$\hat{\mathcal{R}}_n^x = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right|\right].$$

Note that in what follows we consider a general class of functions $\mathcal{F}$. However, for applying the results in order to bound empirical risk minimization, we take x_i to be (x_i, y_i) and $\mathcal{F}$ to be $\ell \circ \mathcal{F}$. We define the empirical l_1 distance as

$$d_1^x(f,g) = \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - g(x_i)|.$$

Theorem: If $0 \le f \le 1$ for all $f \in \mathcal{F}$, then for any $x = (x_1, \ldots, x_n)$, we have

$$\hat{R}_n^x(\mathcal{F}) \le \inf_{\varepsilon \ge 0} \left\{\varepsilon + \sqrt{\frac{2\log\left(2N(\mathcal{F}, d_1^x, \varepsilon)\right)}{n}}\right\}.$$

Proof. Fix $x = (x_1, \ldots, x_n)$ and $\varepsilon > 0$. Let V be a minimal ε-net of $(\mathcal{F}, d_1^x)$. Thus, by definition we have that $|V| = N(\mathcal{F}, d_1^x, \varepsilon)$. For any $f \in \mathcal{F}$, define $f^\circ \in V$ such that

$d_1^x(f, f^\circ) \leq \varepsilon$. We have that

$$\begin{aligned}
R_n^x(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i (f(x_i) - f^\circ(x_i))|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f^\circ(x_i)|\right] \\
&\leq \varepsilon + \mathbb{E}\left[\max_{f\in V} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)|\right] \\
&\leq \varepsilon + \sqrt{\frac{2\log(2|V|)}{n}} \\
&= \varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}.
\end{aligned}$$

Since the previous bound holds for any ε, we can take the infimum over all $\varepsilon \geq 0$ to obtain

$$R_n^x(\mathcal{F}) \leq \inf_{\varepsilon\geq 0} \{\varepsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, d_1^x, \varepsilon))}{n}}\}.$$

□

The previous bound clearly establishes a trade-off because as ε decreases $N(\mathcal{F}, d_1^x, \varepsilon)$ increases.

5.2.2 Computing Covering Numbers

As a warm-up, we will compute the covering number of the ℓ_2 ball of radius 1 in $\mathbb{R}^d$ denoted by B_2. We will show that the covering is at most $(\frac{3}{\varepsilon})^d$. There are several techniques to prove this result: one is based on a probabilistic method argument and one is based on greedily finding an ε-net. We will describe the later approach here. We select points in V one after another so that at step k, we have $u_k \in B_2 \setminus \cup_{j=1}^k B(u_j, \varepsilon)$. We will continue this procedure until we run out of points. Let it be step N. This means that $V = \{u_1, \ldots, u_N\}$ is an ε-net. We claim that the balls $B(u_i, \frac{\varepsilon}{2})$ and $B(u_j, \frac{\varepsilon}{2})$ for any $i, j \in \{1, \ldots, N\}$ are disjoint. The reason is that if $v \in B(u_i, \frac{\varepsilon}{2}) \cap B(u_j, \frac{\varepsilon}{2})$, then we would have

$$\|u_i - u_j\|_2 \leq \|u_i - v\|_2 + \|v - u_j\|_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which contradicts the way we have chosen the points. On the other hand, we have that $\cup_{j=1}^N B(u_j, \frac{\varepsilon}{2}) \subseteq (1 + \frac{\varepsilon}{2})B_2$. Comparing the volume of these two sets leads to

$$|V|(\frac{\varepsilon}{2})^d \mathsf{vol}(B_2) \leq (1 + \frac{\varepsilon}{2})^d \mathsf{vol}(B_2)\,,$$

where $\mathsf{vol}(B_2)$ denotes the volume of the unit Euclidean ball in d dimensions. It yields,

$$|V| \leq \frac{\left(1 + \frac{\varepsilon}{2}\right)^d}{\left(\frac{\varepsilon}{2}\right)^d} = \left(\frac{2}{\varepsilon} + 1\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d.$$

For any $p \geq 1$, define

$$d_p^x(f,g) = \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|^p \right)^{\frac{1}{p}},$$

and for $p = \infty$, define

$$d_\infty^x(f,g) = \max_i |f(x_i) - g(x_i)|.$$

Using the previous theorem, in order to bound $\hat{\mathcal{R}}_n^x$ we need to bound the covering number with d_1^x norm. We claim that it is sufficient to bound the covering number for the infinity-norm. In order to show this, we will compare the covering number of the norms $d_p^x(f,g) = \left(\frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|^p\right)^{\frac{1}{p}}$ for $p \geq 1$ and conclude that a bound on $N(\mathcal{F}, d_\infty^x, \varepsilon)$ implies a bound on $N(\mathcal{F}, d_p^x, \varepsilon)$ for any $p \geq 1$.

Proposition: For any $1 \leq p \leq q$ and $\varepsilon > 0$, we have that

$$N(\mathcal{F}, d_p^x, \varepsilon) \leq N(\mathcal{F}, d_q^x, \varepsilon).$$

Proof. First note that if $q = \infty$, then the inequality evidently holds. Because, we have

$$(\frac{1}{n} \sum_{i=1}^{n} |z_i|^p)^{\frac{1}{p}} \leq \max_i |z_i|,$$

which leads to $B(f, d_\infty^x, \varepsilon) \subseteq B(f, d_p^x, \varepsilon)$ and $N(f, d_\infty, \varepsilon) \geq N(f, d_p, \varepsilon)$. Now suppose that $1 \leq p \leq q < \infty$. Using Hölder's inequality with $r = \frac{q}{p} \geq 1$ we obtain

$$\left(\frac{1}{n} \sum_{i=1}^{n} |z_i|^p \right)^{\frac{1}{p}} \leq n^{-\frac{1}{p}} \left(\sum_{i=1}^{n} 1 \right)^{(1-\frac{1}{r})\frac{1}{p}} \left(\sum_{i=1}^{n} |z_i|^{pr} \right)^{\frac{1}{pr}} = \left(\frac{1}{n} \sum_{i=1}^{n} |z_i|^q \right)^{\frac{1}{q}}.$$

This inequality yeilds

$$B(f, d_q^x, \varepsilon) = \{g \,:\, d_q^x(f,g) \leq \varepsilon\} \subseteq B(f, d_p^x, \varepsilon),$$

which leads to $N(f, d_q, \varepsilon) \geq N(f, d_p, \varepsilon)$. □

Using this propositions we only need to bound $N(\mathcal{F}, d_\infty^x, \varepsilon)$.

Let the function class be $\mathcal{F} = \{f(x) = \langle f, x \rangle, f \in B_p^d, x \in B_q^d\}$, where $\frac{1}{p} + \frac{1}{q} = 1$. This leads to $|f| \leq 1$.

Claim: $N(\mathcal{F}, d_\infty^x, \varepsilon) \leq (\frac{2}{\varepsilon})^d$.

This leads to

$$\hat{R}_n^x(\mathcal{F}) \leq \inf_{\varepsilon > 0} \{\varepsilon + \sqrt{\frac{2d \log(4/\varepsilon)}{n}}\}.$$

Taking $\varepsilon = O(\sqrt{\frac{d \log n}{n}})$, we obtain

$$\hat{R}_n^x(\mathcal{F}) \leq O(\sqrt{\frac{d \log n}{n}}).$$

18.657 Mathematics of Machine Learning
Fall 2015