18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Xuhong Zhang

Lecture 9
Oct. 7, 2015

Recall that last lecture we talked about convex relaxation of the original problem

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i)$$

by considering soft classifiers (i.e. whose output is in $[-1, 1]$ rather than in $\{0, 1\}$) and convex surrogates of the loss function (e.g. hinge loss, exponential loss, logistic loss):

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_{\varphi,n}(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f(X_i))$$

And $\hat{h} = \text{sign}(\hat{f})$ will be used as the 'hard' classifier.

We want to bound the quantity $R_\varphi(\hat{f}) - R_\varphi(\bar{f})$, where $\bar{f} = \operatorname{argmin}_{f \in \mathcal{F}} R_\varphi(f)$.

(1) $\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_{\varphi,n}(f)$, thus

$$\begin{aligned}
R_\varphi(\hat{f}) &= R_\varphi(\bar{f}) + \hat{R}_{\varphi,n}(\bar{f}) - \hat{R}_{\varphi,n}(\bar{f}) + \hat{R}_{\varphi,n}(\hat{f}) - \hat{R}_{\varphi,n}(\hat{f}) + R_\varphi(\hat{f}) - R_\varphi(\bar{f}) \\
&\leq R_\varphi(\bar{f}) + \hat{R}_{\varphi,n}(\bar{f}) - \hat{R}_{\varphi,n}(\hat{f}) + R_\varphi(\hat{f}) - R_\varphi(\bar{f}) \\
&\leq R_\varphi(\bar{f}) + 2 \sup_{f \in \mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|
\end{aligned}$$

(2) Let us first focus on $\mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|]$. Using the symmetrization trick as before, we know it is upper-bounded by $2\mathcal{R}_n(\varphi \circ \mathcal{F})$, where the Rademacher complexity

$$\mathcal{R}_n(\varphi \circ \mathcal{F}) = \sup_{X_1,\dots,X_n,Y_1,\dots,Y_n} \mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \sigma_i \varphi(-Y_i f(X_i))|]$$

One thing to notice is that $\varphi(0) = 1$ for the loss functions we consider (hinge loss, exponential loss and logistic loss), but in order to apply contraction inequality later, we require $\varphi(0) = 0$. Let us define $\psi(\cdot) = \varphi(\cdot) - 1$. Clearly $\psi(0) = 0$, and

$$\begin{aligned}
&\mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i f(X_i))])|] \\
&= \mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (\psi(-Y_i f(X_i)) - \mathbb{E}[\psi(-Y_i f(X_i))])|] \\
&\leq 2\mathcal{R}_n(\psi \circ \mathcal{F})
\end{aligned}$$

(3) The Rademacher complexity of $\psi \circ \mathcal{F}$ is still difficult to deal with. Let us assume that $\varphi(\cdot)$ is L-Lipschitz, (as a result, $\psi(\cdot)$ is also L-Lipschitz), apply the contraction inequality, we have

$$R_n(\psi \circ \mathcal{F}) \leq 2L R_n(\mathcal{F})$$

(4) Let $Z_i = (X_i, Y_i)$, $i = 1, 2, ..., n$ and

$$g(Z_1, Z_2, ..., Z_n) = \sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n}(\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i f(X_i))])|$$

Since $\varphi(\cdot)$ is monotonically increasing, it is not difficult to verify that $\forall Z_1, Z_2, ..., Z_n, Z_i'$

$$|g(Z_1, ..., Z_i, ..., Z_n) - g(Z_1, ..., Z_i', ..., Z_n)| \le \frac{1}{n}(\varphi(1) - \varphi(-1)) \le \frac{2L}{n}$$

The last inequality holds since g is L-Lipschitz. Apply Bounded Difference Inequality,

$$\mathbb{P}(|\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|]| > t) \le 2\exp(-\frac{2t^2}{\sum_{i=1}^{n}(\frac{2L}{n})^2})$$

Set the RHS of above equation to δ, we get:

$$\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)| \le \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{R}_{\varphi,n}(f) - R_\varphi(f)|] + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

(5) Combining (1) - (4), we have

$$R_\varphi(\hat{f}) \le R_\varphi(\bar{f}) + 8L\mathcal{R}_n(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

1.4 Boosting

In this section, we will specialize the above analysis to a particular learning model: Boosting. The basic idea of Boosting is to convert a set of weak learners (i.e. classifiers that do better than random, but have high error probability) into a strong one by using the weighted average of weak learners' opinions. More precisely, we consider the following function class

$$\mathcal{F} = \{\sum_{j=1}^{M} \theta_j h_j(\cdot) : |\theta|_1 \le 1, h_j : \mathcal{X} \mapsto [-1, 1], j \in \{1, 2, ..., M\} \text{ are classifiers}\}$$

and we want to upper bound $\mathcal{R}_n(\mathcal{F})$ for this choice of $\mathcal{F}$.

$$\mathcal{R}_n(\mathcal{F}) = \sup_{Z_1,\ldots,Z_n} \mathbb{E}[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i Y_i f(X_i)|] = \frac{1}{n}\sup_{Z_1,\ldots,Z_n} \mathbb{E}[\sup_{|\theta|_1\le 1} |\sum_{j=1}^{M}\theta_j \sum_{i=1}^{n} Y_i\sigma_i h_j(X_i)|]$$

Let $g(\theta) = |\sum_{j=1}^{M}\theta_j \sum_{i=1}^{n} Y_i\sigma_i h_j(X_i)|$. It is easy to see that $g(\theta)$ is a convex function, thus $\sup_{|\theta|_1\le 1} g(\theta)$ is achieved at a vertex of the unit ℓ_1 ball $\{\theta : \|\theta\|_1 \le 1\}$. Define the finite set

$$B_{\mathbf{X},\mathbf{Y}} \triangleq \left\{ \pm \begin{pmatrix} Y_1 h_1(X_1) \\ Y_2 h_1(X_2) \\ \vdots \\ Y_n h_1(X_n) \end{pmatrix}, \pm \begin{pmatrix} Y_1 h_2(X_1) \\ Y_2 h_2(X_2) \\ \vdots \\ Y_n h_2(X_n) \end{pmatrix}, \ldots, \pm \begin{pmatrix} Y_1 h_M(X_1) \\ Y_2 h_M(X_2) \\ \vdots \\ Y_n h_M(X_n) \end{pmatrix} \right\}$$

Then

$$\mathcal{R}_n(\mathcal{F}) = \sup_{\mathbf{X},\mathbf{Y}} R_n(B_{\mathbf{X},\mathbf{Y}})\,.$$

Notice $\max_{b\in B_{\mathbf{X},\mathbf{Y}}} |b|_2 \le \sqrt{n}$ and $|B_{\mathbf{X},\mathbf{Y}}| = 2M$. Therefore, using a lemma from Lecture 5, we get

$$\mathcal{R}_n(B_{\mathbf{X},\mathbf{Y}}) \le \big[\max_{b\in B_{\mathbf{X},\mathbf{Y}}} |b|_2\big]\frac{\sqrt{2\log(2|B_{\mathbf{X},\mathbf{Y}}|)}}{n} \le \sqrt{\frac{2\log(4M)}{n}}$$

Thus for Boosting,

$$R_\varphi(\hat{f}) \le R_\varphi(\bar{f}) + 8L\sqrt{\frac{2\log(4M)}{n}} + 2L\sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{with probability 1 - } \delta$$

To get some ideas of what values L usually takes, consider the following examples:

(1) for hinge loss, i.e. $\varphi(x) = (1+x)_+$, $L = 1$.

(2) for exponential loss, i.e. $\varphi(x) = e^x$, $L = e$.

(3) for logistic loss, i.e. $\varphi(x) = \log_2(1+e^x)$, $L = \frac{e}{1+e}\log_2(e) \approx 2.43$

Now we have bounded $R_\varphi(\hat{f}) - R_\varphi(\bar{f})$, but this is not yet the excess risk. Excess risk is defined as $R(\hat{f}) - R(f^*)$, where $f^* = \operatorname{argmin}_f R_\varphi(f)$. The following theorem provides a bound for excess risk for Boosting.

Theorem: Let $\mathcal{F} = \{\sum_{j=1}^M \theta_j h_j : \|\theta\|_1 \le 1, h_j s \text{ are weak classifiers}\}$ and φ is an L-Lipschitz convex surrogate. Define $\hat{f} = \operatorname{argmin}_{f\in\mathcal{F}} R_{\varphi,n}(f)$ and $\hat{h} = \operatorname{sign}(\hat{f})$. Then

$$R(\hat{h}) - R^* \le 2c\big(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*)\big)^\gamma + 2c\left(8L\sqrt{\frac{2\log(4M)}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma$$

with probability $1-\delta$

Proof.

$$\begin{aligned}
R(\hat{h}) - R^* &\le 2c\big(R_\varphi(f) - R_\varphi(f^*)\big)^\gamma \\
&\le 2c\left(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*) + 8L\sqrt{\frac{2\log(4M)}{n}} + 2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma \\
&\le 2c\big(\inf_{f\in\mathcal{F}} R_\varphi(f) - R_\varphi(f^*)\big)^\gamma + 2c\left(8L\sqrt{\frac{2\log(4M)}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{\log(2/\delta)}{2n}}\right)^\gamma
\end{aligned}$$

Here the first inequality uses Zhang's lemma and the last one uses the fact that for $a_i \ge 0$ and $\gamma \in [0,1]$, $(a_1+a_2+a_3)^\gamma \le a_1^\gamma + a_2^\gamma + a_3^\gamma$. □

1.5 Support Vector Machines

In this section, we will apply our analysis to another important learning model: Support Vector Machines (SVMs). We will see that hinge loss $\varphi(x) = (1+x)_+$ is used and the associated function class is $\mathcal{F} = \{f : \|f\|_W \le \lambda\}$ where W is a Hilbert space. Before analyzing SVMs, let us first introduce Reproducing Kernel Hilbert Spaces (RKHS).

1.5.1 Reproducing Kernel Hilbert Spaces (RKHS)

Definition: A function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called a *positive symmetric definite kernel* (PSD kernel) if

(1) $\forall x, x' \in \mathcal{X},\ K(x, x') = K(x', x)$

(2) $\forall n \in \mathbb{Z}_+, \forall x_1, x_2, ..., x_n$, the $n \times n$ matrix with $K(x_i, x_j)$ as its element in i^{th} row and j^{th} column is positive semi-definite. In other words, for any $a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

Let us look at a few examples of PSD kernels.

Example 1 Let $\mathcal{X} = \mathbb{R}$, $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$ is a PSD kernel, since $\forall a_1, a_2, ..., a_n \in \mathbb{R}$

$$\sum_{i,j} a_i a_j \langle x_i, x_j \rangle_{\mathbb{R}^d} = \sum_{i,j} \langle a_i x_i, a_j x_j \rangle_{\mathbb{R}^d} = \langle \sum_i a_i x_i, \sum_j a_j x_j \rangle_{\mathbb{R}^d} = \| \sum_i a_i x_i \|^2_{\mathbb{R}^d} \geq 0$$

Example 2 The Gaussian kernel $K(x, x') = \exp(-\frac{1}{2\sigma^2} \|x - x'\|^2_{\mathbb{R}^d})$ is also a PSD kernel.

Note that here and in the sequel, $\| \cdot \|_W$ and $\langle \cdot, \cdot \rangle_W$ denote the norm and inner product of Hilbert space W.

Definition: Let W be a Hilbert space of functions $\mathcal{X} \mapsto \mathbb{R}$. A symmetric kernel $K(\cdot, \cdot)$ is called **reproducing kernel** of W if

(1) $\forall x \in \mathcal{X}$, the function $K(x, \cdot) \in W$.

(2) $\forall x \in \mathcal{X},\ f \in W,\ \langle f(\cdot), K(x, \cdot) \rangle_W = f(x)$.

If such a $K(x, \cdot)$ exists, W is called a **reproducing kernel Hilbert space** (RKHS).

Claim: If $K(\cdot, \cdot)$ is a reproducing kernel for some Hilbert space W, then $K(\cdot, \cdot)$ is a PSD kernel.

Proof. $\forall a_1, a_2, ..., a_n \in \mathbb{R}$, we have

$$\begin{aligned}
\sum_{i,j} a_i a_j K(x_i, x_j) &= \sum_{i,j} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle \quad \text{(since } K(\cdot, \cdot) \text{ is reproducing)} \\
&= \langle \sum_i a_i K(x_i, \cdot), \sum_j a_j K(x_j, \cdot) \rangle_W \\
&= \| \sum_i a_i K(x_i, \cdot) \|^2_W \geq 0
\end{aligned}$$

∎

In fact, the above claim holds both directions, i.e. if a kernel $K(\cdot,\cdot)$ is PSD, it is also a reproducing kernel.

A natural question to ask is, given a PSD kernel $K(\cdot,\cdot)$, how can we build the corresponding Hilbert space (for which $K(\cdot,\cdot)$ is a reproducing kernel)? Let us look at a few examples.

Example 3 Let $\varphi_1, \varphi_2, ..., \varphi_M$ be a set of orthonormal functions in $L_2([0,1])$, i.e. for any $j, k \in \{1, 2, ..., M\}$

$$\int_x \varphi_j(x)\varphi_k(x)dx = \langle \varphi_j, \varphi_k \rangle = \delta_{jk}$$

Let $K(x, x') = \sum_{j=1}^{M} \varphi_j(x)\varphi_j(x')$. We claim that the Hilbert space

$$W = \{\sum_{j=1}^{M} a_j\varphi_j(\cdot) : a_1, a_2, ..., a_M \in \mathbb{R}\}$$

equipped with inner product $\langle \cdot,\cdot \rangle_{L_2}$ is a RKHS with reproducing kernel $K(\cdot,\cdot)$.

Proof. (1) $K(x,\cdot) = \sum_{j=1}^{M} \varphi_j(x)\varphi_j(\cdot) \in W$. (Choose $a_j = \varphi_j(x)$).

(2) If $f(\cdot) = \sum_{j=1}^{M} a_j\varphi_j(\cdot)$,

$$\langle f(\cdot), K(x,\cdot)\rangle_{L_2} = \langle \sum_{j=1}^{M} a_j\varphi_j(\cdot), \sum_{k=1}^{M} \varphi_k(x)\varphi_k(\cdot)\rangle_{L_2} = \sum_{j=1}^{M} a_j\varphi_j(x) = f(x)$$

(3) $K(x, x')$ is a PSD kernel: $\forall a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) = \sum_{i,j,k} a_i a_j \varphi_k(x_i)\varphi_k(x_j) = \sum_k (\sum_i a_i\varphi_k(x_i))^2 \geq 0$$

□

Example 4 If $\mathcal{X} = \mathbb{R}^d$, and $K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$, the corresponding Hilbert space is $W = \{\langle w, \cdot \rangle : w \in \mathbb{R}^d\}$ (i.e. all linear functions) equipped with the following inner product: if $f = \langle w, \cdot \rangle$, $g = \langle v, \cdot \rangle$, $\langle f, g \rangle \triangleq \langle w, v \rangle_{\mathbb{R}^d}$.

Proof. (1) $\forall x \in \mathbb{R}^d$, $K(x,\cdot) = \langle x, \cdot \rangle_{\mathbb{R}^d} \in W$.

(2) $\forall f = \langle w, \cdot \rangle_{\mathbb{R}^d} \in W$, $\forall x \in \mathbb{R}^d$, $\langle f, K(x,\cdot)\rangle = \langle w, x \rangle_{\mathbb{R}^d} = f(x)$

(3) $K(x, x')$ is a PSD kernel: $\forall a_1, a_2, ..., a_n \in \mathbb{R}$,

$$\sum_{i,j} a_i a_j K(x_i, x_j) = \sum_{i,j} a_i a_j \langle x_i, x_j \rangle = \langle \sum_i a_i x_i, \sum_j a_j x_j \rangle_{\mathbb{R}^d} = \| \sum_i a_i x_i \|^2_{\mathbb{R}^d} \geq 0$$

□