18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 10
Scribe: Aden Forrow Oct. 13, 2015

Recall the following definitions from last time:

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. $\forall n \in \mathbb{Z}_+, \forall x_1, x_2, \dots, x_n$, the $n \times n$ matrix with entries $K(x_i, x_j)$ is positive definite. Equivalently, $\forall a_1, a_2, \dots, a_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$$

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

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

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

As before, $\langle \cdot, \cdot \rangle_W$ and $\| \cdot \|_W$ respectively denote the inner product and norm of W. The subscript W will occasionally be omitted. We can think of the elements of W as infinite linear combinations of functions of the form $K(x, \cdot)$. Also note that

$$\langle K(x, \cdot), K(y, \cdot)\rangle_W = K(x, y)$$

Since so many of our tools rely on functions being bounded, we'd like to be able to bound the functions in W. We can do this uniformly over $x \in \mathcal{X}$ if the diagonal $K(x, x)$ is bounded.

Proposition: Let W be a RKHS with PSD K such that $\sup_{x \in \mathcal{X}} K(x, x) = k_{\max}$ is finite. Then $\forall f \in W$,

$$\sup_{x \in \mathcal{X}} |f(x)| \leq \|f\|_W \sqrt{k_{\max}}$$

.

Proof. We rewrite $f(x)$ as an inner product and apply Cauchy-Schwartz.

$$f(x) = \langle f, K(x, \cdot)\rangle_W \leq \|f\|_W \|K(x, \cdot)\|_W$$

Now $\|K(x, \cdot)\|_W^2 = \langle K(x, \cdot), K(x, \cdot)\rangle_W = K(x, x) \leq k_{\max}$. The result follows immediately. □

1.5.2 Risk Bounds for SVM

We now analyze support vector machines (SVM) the same way we analyzed boosting. The general idea is to choose a linear classifier that maximizes the margin (distance to classifiers) while minimizing empirical risk. Classes that are not linearly separable can be embedded in a higher dimensional space so that they are linearly separable. We won't go into that, however; we'll just consider the abstract optimization over a RKHS W.

Explicitly, we minimize the empirical φ-risk over a ball in W with radius λ:

$$\hat{f} = \min_{f \in W, \|f\|_W \leq \lambda} \hat{R}_{n,\varphi}(f)$$

The soft classifier $\hat{f}$ is then turned into a hard classifier $\hat{h} = \text{sign}(\hat{f})$. Typically in SVM φ is the hinge loss, though all our convex surrogates behave similarly. To choose W (the only other free parameter), we choose a PSD $K(x_1, x_2)$ that measures the similarity between two points x_1 and x_2.

As written, this is an intractable minimum over an infinite dimensional ball $\{f, \|f\|_W \leq \lambda\}$. The minimizers, however, will all be contained in a finite dimensional subset.

Theorem: Representer Theorem. Let W be a RKHS with PSD K and let $G : \mathbb{R}^n \mapsto \mathbb{R}$ be any function. Then

$$\begin{aligned}\min_{f \in W, \|f\| \leq \lambda} G(f(x_1), \ldots, f(x_n)) &= \min_{f \in \bar{W}_n, \|f\| \leq \lambda} G(f(x_1), \ldots, f(x_n)) \\ &= \min_{\alpha \in \mathbb{R}^n, \alpha^\top \mathbb{K} \alpha \leq \lambda^2} G(g_\alpha(x_1), \ldots, g_\alpha(x_n)),\end{aligned}$$

where

$$\bar{W}_n = \{f \in W | f(\cdot) = g_\alpha(\cdot) = \sum_{i=1}^n \alpha_i K(x_i, \cdot)\}$$

and $\mathbb{K}_{ij} = K(x_i, x_j)$.

Proof. Since $\bar{W}_n$ is a linear subspace of W, we can decompose any $f \in W$ uniquely as $f = \bar{f} + f^\perp$ with $\bar{f} \in \bar{W}_n$ and $f^\perp \in \bar{W}_n^\perp$. The Pythagorean theorem then gives

$$\|f\|_W^2 = \|\bar{f}\|_W^2 + \|f^\perp\|_W^2$$

Moreover, since $K(x_i, \cdot) \in \bar{W}_n$,

$$f^\perp(x_i) = \langle f^\perp, K(x_i, \cdot) \rangle_W = 0$$

So $f(x_i) = \bar{f}(x_i)$ and

$$G(f(x_1), \ldots, f(x_n)) = G(\bar{f}(x_1), \ldots, \bar{f}(x_n)).$$

Because $f^\perp$ does not contribute to G, we can remove it from the constraint:

$$\min_{f \in W, \|\bar{f}\|^2 + \|f^\perp\|^2 \leq \lambda^2} G(f(x_1), \ldots, f(x_n)) = \min_{f \in W, \|\bar{f}\|^2 \leq \lambda^2} G(\bar{f}(x_1), \ldots, \bar{f}(x_n)).$$

Restricting to $f \in \bar{W}_n$ now does not change the minimum, which gives us the first equality. For the second, we need to show that $\|g_\alpha\|_W \leq \lambda$ is equivalent to $\alpha^\top \mathbb{K}\alpha \leq \lambda^2$.

$$\begin{aligned}
\|g_\alpha\|^2 &= \langle g_\alpha, g_\alpha \rangle \\
&= \langle \sum_{i=1}^n \alpha_i K(x_i, \cdot), \sum_{j=1}^n \alpha_j K(x_j, \cdot) \rangle \\
&= \sum_{i,j=1}^n \alpha_i \alpha_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle \\
&= \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \\
&= \alpha^\top \mathbb{K} \alpha
\end{aligned}$$

□

We've reduced the infinite dimensional problem to a minimization over $\alpha \in \mathbb{R}^n$. This works because we're only interested in G evaluated at a finite set of points. The matrix $\mathbb{K}$ here is a Gram matrix, though we will not not use that. $\mathbb{K}$ should be a measure of the similarity of the points x_i. For example, we could have $W = \{\langle x, \cdot \rangle_{\mathbb{R}^d}, x \in \mathbb{R}^d\}$ with $K(x, y$ the usual inner product $K(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$.

We've shown that $\hat{f}$ only depends on K through $\mathbb{K}$, but does $\hat{R}_{n,\varphi}$ depend on $K(x, y)$ for $x, y \notin \{x_i\}$? It turns out not to:

$$\hat{R}_{n,\varphi} = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i g_\alpha(x_i)) = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i \sum_{j=1}^n \alpha_j K(x_j, x_i)).$$

The last expression only involves $\mathbb{K}$. This makes it easy to encode all the knowledge about our problem that we need. The hard classifier is

$$\hat{h}(x) = \text{sign}(\hat{f}(x)) = \text{sign}(g_{\hat{\alpha}}(x)) = \text{sign}(\sum_{j=1}^n \hat{\alpha}_j K(x_j, x))$$

If we are given a new point x_{n+1}, we need to compute a new column for $\mathbb{K}$. Note that x_{n+1} must be in some way comparable or similar to the previous $\{x_i\}$ for the whole idea of extrapolating from data to make sense.

The expensive part of SVMs is calculating the $n \times n$ matrix $\mathbb{K}$. In some applications, $\mathbb{K}$ may be sparse; this is faster, but still not as fast as deep learning. The minimization over the ellipsoid $\alpha^\top \mathbb{K} \alpha$ requires quadratic programming, which is also relatively slow. In practice, it's easier to solve the Lagrangian form of the problem

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i g_\alpha(x_i)) + \lambda' \alpha^\top \mathbb{K} \alpha$$

This formulation is equivalent to the constrained one. Note that λ and λ' are different.

SVMs have few tuning parameters and so have less flexibility than other methods.

We now turn to analyzing the performance of SVM.

Theorem: Excess Risk for SVM. Let φ be an L-Lipschitz convex surrogate and W a RKHS with PSD K such that $\max_x |K(x,x)| = k_{\max} < \infty$. Let $\hat{h}_{n,\varphi} = \text{sign}\, \hat{f}_{n,\varphi}$, where $\hat{f}_{n,\varphi}$ is the empirical φ-risk minimizer over $\mathcal{F} = \{f \in W . \|f\|_W \leq \lambda\}$ (that is, $\hat{R}_{n,\varphi}(\hat{f}_{n,\varphi}) \leq \hat{R}_{n,\varphi}(f) \forall f \in \mathcal{F}$). Suppose $\lambda\sqrt{k_{\max}} \leq 1$. Then

$$R(\hat{h}_{n,\varphi}) - R^* \leq 2c\left(\inf_{f\in\mathcal{F}}(R_\varphi(f) - R^*_\varphi)\right)^\gamma + 2c\left(8L\lambda\sqrt{\frac{k_{\max}}{n}}\right)^\gamma + 2c\left(2L\sqrt{\frac{2\log(2/\delta)}{n}}\right)^\gamma$$

with probability $1-\delta$. The constants c and γ are those from Zhang's lemma. For the hinge loss, $c = \frac{1}{2}$ and $\gamma = 1$.

Proof. The first term comes from optimizing over a restricted set $\mathcal{F}$ instead of all classifiers. The third term comes from applying the bounded difference inequality. These arise in exactly the same way as they do for boosting, so we will omit the proof for those parts. For the middle term, we need to show that $R_{n,\varphi}(\mathcal{F}) \leq \lambda\sqrt{\frac{k_{\max}}{n}}$.

First, $|f(x)| \leq \|f\|_W\sqrt{k_{\max}} \leq \lambda\sqrt{k_{\max}} \leq 1$ for all $f \in \mathcal{F}$, so we can use the contraction inequality to replace $R_{n,\varphi}(\mathcal{F})$ with $R_n(\mathcal{F})$. Next we'll expand $f(x_i)$ inside the Rademacher complexity and bound inner products using Cauchy-Schwartz.

$$\begin{aligned}
R_n(\mathcal{F}) &= \sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right] \\
&= \frac{1}{n}\sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma_i\langle K(x_i,\cdot), f\rangle\right|\right] \\
&= \frac{1}{n}\sup_{x_1,\ldots,x_n} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\langle\sum_{i=1}^n \sigma_i K(x_i,\cdot), f\rangle\right|\right] \\
&\leq \frac{\lambda}{n}\sup_{x_1,\ldots,x_n} \sqrt{\mathbb{E}\left[\|\sum_{i=1}^n \sigma_i K(x_i,\cdot)\|_W^2\right]}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\|\sum_{i=1}^n \sigma_i K(x_i,\cdot)\|_W^2\right] &= \mathbb{E}\left[\langle\sum_{i=1}^n \sigma_i K(x_i,\cdot), \sum_{j=1}^n \sigma_j K(x_j,\cdot)\rangle_W\right] \\
&= \sum_{i,j=1}^n \langle K(x_i,\cdot), K(x_j,\cdot)\rangle \mathbb{E}[\sigma_i\sigma_j] \\
&= \sum_{i,j=1}^n K(x_i,x_j)\delta_{ij} \\
&\leq nk_{\max}
\end{aligned}$$

So $R_n(\mathcal{F}) \leq \lambda\sqrt{\frac{k_{\max}}{n}}$ and we are done with the new parts of the proof. The remainder follows as with boosting, using symmetrization, contraction, the bounded difference inequality, and Zhang's lemma. □