Lecture 19: FoSML

- We will relate kernel methods to SVM.
- Since SVM's care only about hyperplane projections, which are inner products, there is intuition that Kernels should adapt naturally to this setting. This is because Kernels can be realized as inner products, admittedly in some (perhaps infinite) high-dim. Hilbert space.
- Consider the following optimization problem:

$$\text{(*)} \quad \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \text{ subject to}$$

- $0 \le \alpha_i \le C, \forall i = 1, \dots, m$
- $\sum_{i=1}^{m} \alpha_i y_i = 0,$

- Recalling some SVM results, we may write a hypothesis solution h as

$$h(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \right),$$

where $b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i)$ for any x_i with $\alpha_i \in (0, C)$.

- Rewriting (*) in vector form using the kernel matrix $K \in \mathbb{R}^{m \times m}$ associated to K and the sample $\{x_i\}_{i=1}^{m}$ yields

$$\max_{\alpha} \; 2 \underbrace{\mathbb{1}^T}_{\text{vectors}} \alpha - (\alpha \circ y)^T K (\alpha \circ y) \quad \text{subject to} \quad \cdot \; 0 \le \alpha \le C \quad \cdot \; \alpha^T y = 0$$

Here, $\alpha \circ y = (\alpha_1 y_1, \dots, \alpha_m y_m)$ is the Hadamard product of α, y.

- In vector notation, the solution is the same but with $b = y_i - (\alpha \circ y)^T K e_i$

for any x_i with $0 < \alpha_i < C$ $\qquad$ $(0, \dots, 0, 1, \dots, 0)$ ← i^{th} coordinate

- So, the solution can be written as a linear combination of kernel evaluations, at least if we ignore the offsets b.
- This is a general property of a range of optimization problems:

Theorem (Representer Theorem): Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel, with $\mathcal{H}$ its associated RKHS. For any non-decreasing $G: \mathbb{R} \to \mathbb{R}$ and any loss function $L: \mathbb{R}^m \to \mathbb{R}$,

$$\underset{h \in \mathcal{H}}{\operatorname{arg\,min}} \; F(h) = \underset{h \in \mathcal{H}}{\operatorname{arg\,min}} \; G\left(\|h\|_{\mathcal{H}}\right) + L\left(h(x_1), \dots, h(x_m)\right)$$

admits a solution of the form $h^* = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot)$. If G is in fact strictly increasing, then all solutions have this form.

Proof: Let $\mathcal{H}_1 = \operatorname{span}\left(\{K(x_i, \cdot)\}_{i=1}^m\right)$. This is a subspace of $\mathcal{H}$, and thus $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_1^\perp$, where $\mathcal{H}_1^\perp$ is the orthogonal complement of $\mathcal{H}_1$, i.e.

$$\mathcal{H}_1^\perp = \{h_0 \in \mathcal{H} \mid \langle h_0, h \rangle_{\mathcal{H}} = 0 \;\; \forall h \in \mathcal{H}_1\}.$$

So, any $h \in \mathcal{H}$ may be written as $h = h_1 + h_1^\perp$, where $h_1 \in \mathcal{H}_1$, $h \in \mathcal{H}_1^\perp$.

Since G is non-decreasing,

$$G(\|h_1\|_{\mathcal{H}}) \leq G\left(\sqrt{\|h_1\|_{\mathcal{H}}^2 + \|h_1^\perp\|_{\mathcal{H}}^2}\right) = G(\|h\|_{\mathcal{H}}).$$

By the reproducing property, and the fact that $h^\perp$ is orthogonal to all elements of H_1,

$$\begin{aligned} h(x_i) &= \langle h, K(x_i,\cdot)\rangle_H \\ &= \langle h_1 + h^\perp, K(x_i,\cdot)\rangle_H \\ &= \langle h_1, K(x_i,\cdot)\rangle_H \\ &= h_1(x_i) \end{aligned}$$

Hence, $L(h(x_1),\dots,h(x_m)) = L(h_1(x_1),\dots,h_1(x_m))$. Thus, $F(h_1) \le F(h)$. So, a minimizer of $F(h)$ necessarily lies in H_1. Moreover, if F is strictly increasing, we have that $F(h_1) < F(h)$ if $\|h^\perp\|_H > 0 \Rightarrow$ any minimizer must have $\|h^\perp\|_H = 0$, i.e. must lie in H_1. ∎

We are now in a position to prove learning bounds for KSVM:

Theorem (Rademacher complexity for kernel hypothesis classes): Let $K: X \times X \to \mathbb{R}$ be a PDS kernel and let $\Phi: X \to H$ be a feature mapping associated to K. Let $S \subseteq \{x \mid K(x,x) \le r^2\}$, $r > 0$, be a sample of size m. Let $H = \{x \mapsto \langle w, \Phi(x)\rangle_H \mid \|w\|_H \le \Lambda\}$, $\Lambda \ge 0$. Then for K the associated Gram matrix,

$$\hat{R}_S(H) \le \frac{\Lambda\sqrt{\mathrm{Tr}(K)}}{m} \le \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Proof:

$$\begin{aligned} \hat{R}_S(H) &= \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left(\sup_{\|w\|\le\Lambda}\left\langle w, \sum_{i=1}^m \sigma_i \Phi(x_i)\right\rangle_H\right) \\ &\le \frac{1}{m}\cdot\Lambda\cdot\mathop{\mathbb{E}}_{\sigma}\left(\left\|\sum_{i=1}^m \sigma_i \Phi(x_i)\right\|_H\right) \end{aligned}$$

$$\leq \frac{\Lambda}{m}\left(\mathop{\mathbb{E}}_{z}\left(\left\|\sum_{i=1}^{m} z_i \bar{\phi}(x_i)\right\|_{\mathcal{H}}^2\right)\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\mathop{\mathbb{E}}_{z}\left(\left\langle \sum_{i=1}^{m} z_i \bar{\phi}(x_i), \sum_{j=1}^{m} z_j \bar{\phi}(x_j)\right\rangle\right)\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\mathop{\mathbb{E}}_{z}\left(\sum_{i,j=1}^{m} z_i z_j \langle \bar{\phi}(x_i), \bar{\phi}(x_j)\rangle\right)\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\mathop{\mathbb{E}}_{z}\left(\sum_{i=1}^{m} z_i^2 \langle \bar{\phi}(x_i), \bar{\phi}(x_j)\rangle\right)\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\sum_{i=1}^{m} \|\bar{\phi}(x_i)\|_2^2\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\sum_{i=1}^{m} K(x_i, x_i)\right)^{1/2}$$

$$= \frac{\Lambda}{m} \cdot \sqrt{\mathrm{Tr}(K)}$$

$$\leq \frac{\Lambda}{m} \cdot \sqrt{m \cdot r^2}$$

$$= \sqrt{\frac{\Lambda^2 r^2}{m}} \,. \quad \blacksquare$$

Remark: By the Khintchine-Kahane inequality, the empirical Rademacher complexity can be lower bounded by $\frac{1}{\sqrt{2}}\sqrt{\frac{Tr(K)}{m}}\cdot\frac{\Lambda}{m}$, which suggests the argument made in the above theorem is nearly tight.

Corollary (Margin Bounds for Kernel-Based Hypothesis Classes): Let $K: X \times X \to \mathbb{R}$ be a PDS kernel with $r^2 = \sup_{x \in X} K(x,x)$. Let $\Phi: X \to H$ be a feature mapping associated to K and let $\mathcal{H} = \{x \mapsto \langle w, \Phi(x)\rangle_H \mid \|w\|_H \leq \Lambda\}$. Fix $\rho > 0$. Then $\forall \delta > 0$, each of the following holds with probability at least $1-\delta$ $\forall h \in \mathcal{H}$: (a.) $R(h) \leq \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{r^2\Lambda^2}{\rho^2 m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{m}}$

(b.) $R(h) \leq \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{Tr(K)\Lambda^2}{m\rho^2}} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$.

Proof: Apply our SVM margin bound to our Rademacher complexity result for kernel hypothesis classes.

/

- We have so far considered kernels $K: X \times X \to \mathbb{R}$ that induce SPSD Gram matrices $K \in \mathbb{R}^{m \times m}$.
- In some cases, K is constructed as $K(x,x') = f(d(x,x'))$, where $f: \mathbb{R} \to \mathbb{R}$ is increasing and $d: X \times X \to \mathbb{R}$ is a metric.
- For example, $f_\sigma(z) = \exp(z^2/\sigma^2)$, $d(x,x') = \|x - x'\|_p^2$ yields the

Gaussian Kernel.

- It makes sense to derive conditions on the metric d itself.

Defn: A kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is negative-definite symmetric (NDS) if it is symmetric and if for all $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$, with $c \in \mathbb{R}^{m \times 1}$ and $\mathbb{1}^T c = 0$, $c^T K c \leq 0$

Remark: K PSS $\Rightarrow -K$ NDS, but K NDS $\not\Rightarrow -K$ PDS.

ex: Let $D(x, x') = \|x - x'\|_2^2$. Then $D: \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is NDS. Indeed, let $c \in \mathbb{R}^{m \times 1}$ be s.t. $\mathbb{1}^T c = 0$, i.e. $\sum_{i=1}^n c_i = 0$. Let $\{x_i\}_{i=1}^m \subset \mathbb{R}^D$ be arbitrary. Then

$\langle , \rangle$ = usual $\mathbb{R}^D$ inner product

$$\sum_{i,j=1}^m c_i c_j \|x_i - x_j\|_2^2$$

$$= \sum_{i,j=1}^m c_i c_j \left(\|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j \rangle \right)$$

$$= \sum_{i,j=1}^m c_i c_j \left(\|x_i\|_2^2 + \|x_j\|_2^2 \right) - 2 \left\langle \sum_{i=1}^m c_i x_i , \sum_{j=1}^m c_j x_j \right\rangle$$

$$= \sum_{i,j=1}^m c_i c_j \left(\|x_i\|_2^2 + \|x_j\|_2^2 \right) - 2 \left\| \sum_{i=1}^m c_i x_i \right\|_2^2$$

$$\leq \sum_{i,j=1}^m c_i c_j \left(\|x_i\|_2^2 + \|x_j\|_2^2 \right)$$

$$= \left(\sum_{j=1}^m c_j \right) \cdot \left(\sum_{i=1}^m c_i \|x_i\|_2^2 \right) + \left(\sum_{i=1}^m c_i \right) \cdot \left(\sum_{j=1}^m c_j \|x_j\|_2^2 \right) = 0.$$