FoSML: Lecture 20

- Given $\{x_i\}_{i=1}^m \subset \mathbb{R}^D$, one needs to evaluate $K(x_i, x_j)$ for all (x_i, x_j) pairs, in order to use, e.g., kernel SVM.
- Letting C_K = cost of evaluating $K(x_i, x_j)$ for one (x_i, x_j) pair, this is $O(C_K \cdot m^2)$, which is quadratic in m. This is not acceptable when m is large.
- Moreover, given the kernel matrix / fast construction of $K \in \mathbb{R}^{m \times m}$, an evaluation of a single test point is also slow. Indeed, we classify using

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

so to classify a single x is $O(C_K \cdot m)$.

- The interesting case here is for data s.t. $\tilde{D} < m$, where $\dim(\mathcal{H}_K) = \tilde{D}$, where $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}_K}$, $\Phi : \mathbb{R}^D \to \mathcal{H}_K$. Indeed, otherwise we can handle the primal problem and incur complexity $O(\tilde{D}m)$. But, typically $\tilde{D} > m$ and it is easy to find cases where $\tilde{D} = +\infty$ (i.e. Gaussian kernels).
- Our goal is to discuss some ways around this. One interesting way is to approximate Φ with Ψ, s.t. (1) $\Psi : \mathbb{R}^D \to \mathbb{R}^{D'}$, $D' \ll \tilde{D}$

(2) $\langle \Psi(x_i), \Psi(x_j) \rangle \approx \langle \Phi(x_i), \Phi(x_j) \rangle$.

In this sense, we drop the dimension of the embedding space ((1)), while preserving in an approximate sense the desirable aspects of Φ related to our kernel K, namely $\langle \Psi(x_i) \Psi(x_j) \rangle \approx K(x_i, x_j)$.

• To do so, we will need a little harmonic analysis:

Theorem (Bochner): Let $K(x,x') = G(x-x')$ be a continuous kernel defined over a locally compact set $\mathcal{X}$. Then the following are equivalent:

(a.) K is positive definite

(b.) $\exists \mu$, a non-negative measure, s.t. $G(x) = \int_{\mathcal{X}} \exp(-i\langle w, x\rangle) d\mu(w)$.

• Condition (b) says G is the Fourier transform of a non-negative measure μ.

• Note that the condition $K(x,x') = G(x-x')$ gives such kernels the ~~marked~~ marked shift-invariant, since $K(x+z, x'+z) = K(x,x') \quad \forall x, x', z \in \mathcal{X}$.

Proposition: Let $K(x,x') = G(x,x')$ be a shift-invariant, continuous kernel with associated measure $\mu \geq 0$. Suppose $K(x,x) = 1 \ \forall x$, i.e. μ is a probability measure. Then

$$K(x,x') = \mathop{\mathbb{E}}_{w \sim \mu}\left(\left[\cos(w\cdot x), \sin(w\cdot x)\right]^T \left[\cos(w\cdot x'), \sin(w\cdot x')\right] \right).$$

Proof: Note that since K, μ are $\mathbb{R}$-valued, it suffices to consider the real part of $w \mapsto \exp(-i\langle w, x\rangle)$, since the imaginary part necessary averages out to 0. So, noting $\mathrm{Re}(e^{i\langle w, x\rangle}) = \cos(\langle w, x\rangle)$, we get

$$K(x,x') = \int_{\mathcal{X}} \cos(\langle w, x - x'\rangle) \, d\mu(w)$$

Noting that $\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, we get

$$\int_{\mathcal{X}} \cos(\langle w, x - x' \rangle)\, d\mu(w)$$

$$= \int_{\mathcal{X}} \cos(\langle w, x \rangle - \langle w, x' \rangle)\, d\mu(w)$$

$$= \int_{\mathcal{X}} \left(\cos(\langle w, x \rangle) \cos(\langle w, x' \rangle) + \sin(\langle w, x \rangle) \sin(\langle w, x' \rangle) \right) d\mu(w)$$

$$= \mathop{\mathbb{E}}_{w \sim \mu} \left(\left[\cos(w \cdot x), \sin(w \cdot x) \right]^T \cdot \left[\cos(w \cdot x'), \sin(w \cdot x') \right] \right). \blacksquare$$

- This suggests a scheme for estimating an approximate kernel mapping.
- Given $D \geq 1$, let $\Psi : \mathcal{X} \to \mathbb{R}^{2D}$

$$x \mapsto \sqrt{\frac{1}{D}} \left(\cos(w_1 \cdot x), \sin(w_1 \cdot x), \ldots, \cos(w_D \cdot x), \sin(w_D \cdot x) \right)^T$$

where $\{w_i\}_{i=1}^{D} \overset{iid}{\sim} \mu$, where the FT of μ is K.

- Why do this? Well, $\langle \Psi(x), \Psi(x') \rangle$

$$= \frac{1}{D} \sum_{i=1}^{D} \left[\cos(w_i \cdot x), \sin(w_i \cdot x) \right]^T \left[\cos(w_i \cdot x'), \sin(w_i \cdot x') \right],$$

which is the empirical version of

$$\mathop{\mathbb{E}}_{w \sim \mu} \left(\left[\cos(w \cdot x), \sin(w \cdot x) \right]^T \cdot \left[\cos(w \cdot x'), \sin(w \cdot x') \right] \right).$$

· In fact, as we might expect, this empirical expectation improves as $D \to \infty$.

Lemma: Let K be a continuously differentiable kernel function that satisfies the above conditions, i.e. K is shift-invariant and has associated non-negative measure μ. Assume moreover that $\mathcal{X}$ is compact and let $D = \dim(\mathcal{X})$. Let R be s.t. $\mathcal{X} \subset B_R(0)$ and $\sigma_p^2 = \mathbb{E}_{w \sim \mu}(\|w\|^2) < \infty$. Then for

$\Psi : \mathcal{X} \to \mathbb{R}^{2D}$ defined as $x \mapsto \frac{1}{\sqrt{D}}\left(\cos(w_1 \cdot x), \sin(w_1 \cdot x), \ldots, \cos(w_D \cdot x), \sin(w_D \cdot x)\right)$,

let $r \in (0, 2R]$ and let $\varepsilon > 0$. Then

$$\mathbb{P}_{w \sim \mu}\left(\sup_{x, x' \in \mathcal{X}} \left| \langle \Psi(x), \Psi(x') \rangle - K(x, x') \right| \geq \varepsilon \right)$$

$$\leq 2 \cdot \mathcal{N}(2R, r) \exp\left(-D\varepsilon^2/8\right) + \frac{4r\sigma_p}{\varepsilon},$$

where $\mathcal{N}(R, r)$ is the minimum number of balls of radius r needed to cover a ball of radius R.

Proof: Let $\mathcal{Z} = \{x - x' \mid x, x' \in \mathcal{X}\}$. Clearly $\mathcal{X} \subset B_R(0) \Rightarrow \mathcal{Z} \subset B_{2R}(0)$. Since $\mathcal{X}$ is closed, $\mathcal{Z}$ is also closed and hence compact. Let $B := \mathcal{N}(2R, r)$, and let $\{z_j\}_{j=1}^{B}$ be the centers of the minimal cover. In particular, $\forall z \in \mathcal{Z}$, $\exists j$ s.t. $z = z_j + \delta$, $\|\delta\|_2 \leq r$.

Define $S: \mathcal{Z} \to \mathbb{R}$

$$z \mapsto \langle \Psi(x), \Psi(x') \rangle - k(x, x'),$$

where $z = x - x'$; this map is well-defined because k is shift invariant, same with $\langle \Psi(x), \Psi(x') \rangle$ by construction. Since S is C^1 over $\mathcal{Z}$, S is L-Lipschitz, with $L = \sup_{z \in \mathcal{Z}} \|\nabla S(z)\|_2$. Moreover, if $L < \frac{\varepsilon}{2r}$ and for all $j \in \{1, \dots, B\}$, we have $|S(z_j)| < \frac{\varepsilon}{2}$, then it follows that

$$\begin{aligned} |S(z)| &= |S(z_j + \delta)| \\ &\leq L \cdot \|z_j - (z_j + \delta)\|_2 + |S(z_j)| \\ &\leq rL + \frac{\varepsilon}{2} \\ &< \varepsilon. \end{aligned}$$

So, it suffices to bound the probability that $L \geq \frac{\varepsilon}{2r}$ and $|S(z_j)| \geq \frac{\varepsilon}{2}$.

(a.) Bounding probability of $L \geq \frac{\varepsilon}{2r}$: By linearity of $\mathbb{E}$,

$$\begin{aligned} \mathbb{E}\left(\nabla \langle \Psi(x), \Psi(x') \rangle\right) &= \nabla \mathbb{E}\left(\langle \Psi(x), \Psi(x') \rangle\right) \\ &= \nabla k(x, x'). \end{aligned}$$

So, $\mathbb{E}(L^2) = \mathbb{E}\left(\sup_{z \in \mathcal{Z}} \|\nabla S(z)\|_2^2\right)$

$$= \mathbb{E}\left(\sup_{x, x' \in \mathcal{X}} \|\nabla(\langle \Psi(x), \Psi(x') \rangle) - \nabla k(x, x')\|_2^2\right)$$

$$\leq 2\mathbb{E}\left(\sup_{x,x'\sim\chi}\|\nabla\langle\Psi(x),\Psi(x')\rangle\|_2^2\right)+2\sup_{x,x'\sim\chi}\|\nabla k(x,x')\|_2^2$$

$$=2\mathbb{E}\left(\sup_{x,x'\sim\chi}\|\nabla\langle\Psi(x),\Psi(x')\rangle\|_2^2\right)+2\sup_{x,x'\in\chi}\|\mathbb{E}(\nabla\langle\Psi(x),\Psi(x')\rangle\|_2$$

$$\leq 4\mathbb{E}\left(\sup_{x,x'\sim\chi}\|\nabla\langle\Psi(x),\Psi(x')\rangle\|_2^2\right) \quad (\star)$$

We now analyze $\nabla\langle\Psi(x),\Psi(x')\rangle$ where ∇ is taken w.r.t. $z=x^*-x'$.

Indeed, $\nabla\langle\Psi(x),\Psi(x')\rangle$

$$=\nabla\frac{1}{D}\sum_{i=1}^{D}\cos(w_i\cdot x)\cos(w_i\cdot x')+\sin(w_i\cdot x)\sin(w_i\cdot x')$$

$$=\nabla\frac{1}{D}\sum_{i=1}^{D}\cos(w_i(x-x'))$$

$$=\frac{1}{D}\sum_{i=1}^{D}\sin(w_i(x-x'))\cdot w_i. \quad (\star\star)$$

Combining $(\star)$, $(\star\star)$ yield

$$\mathbb{E}(L^2)\leq 4\mathbb{E}\left(\sup_{x,x'\sim\chi}\|\nabla\langle\Psi(x),\Psi(x')\rangle\|_2^2\right)$$

$$\leq 4\mathbb{E}\left(\sup_{x,x'\sim\chi}\left\|\frac{1}{D}\sum_{i=1}^{D}\sin(w_i(x-x'))\cdot w_i\right\|_2^2\right)$$

$$\leq 4\mathbb{E}\left(\sup_{x,x'\sim\chi}\left\|\frac{1}{D}\sum_{i=1}^{D}\|w_i\|_2^2\right)\right.$$

$$=4\mathbb{E}(\|w_1\|_2^2)$$

$$=4\sigma_p^2.$$

So, by Markov's inequality,

$$\mathbb{P}\left(L \geq \frac{\varepsilon}{2r}\right) \leq \left(\frac{4rz_{\max}}{\varepsilon}\right)^2.$$

(b) Bounding probability that $|S(z_j)| \geq \frac{\varepsilon}{2}$: Note that S is the sum of D i.i.d. r.v.'s, each bounded by $\frac{2}{D}$. So, by Hoeffding and a union bound,

$$\mathbb{P}\left(\exists j \in \{1,\dots,B\} \mid |S(z_j)| \geq \frac{\varepsilon}{2}\right) \leq \sum_{i=1}^{B} \mathbb{P}\left(|S(z_j)| \geq \frac{\varepsilon}{2}\right)$$

$$\leq 2B\,\mathbb{P}\left(|S(z_1)| \geq \frac{\varepsilon}{2}\right)$$

$$\leq 2B \exp\left(-D\varepsilon^2/8\right).$$

Putting (a.), (b.) together gives the desired result:

$$\mathbb{P}\left(\sup_{z \in \mathbb{Z}} |S(z)| \geq \varepsilon\right) \leq 2\underbrace{N(2R,r)}_{=B} \exp\left(-D\varepsilon^2/8\right) + \left(\frac{4rz_{\max}}{\varepsilon}\right)^2. \blacksquare$$