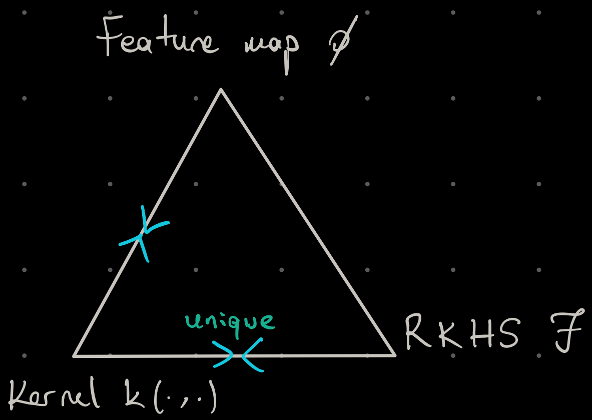
8 - Kernel Methods: Properties and Applications



Recall: 1- Reproducing property:

$f \in \mathcal{F}$, $x \in \mathcal{X}$ $\quad f(x) = \langle k(x,\cdot), f \rangle$

2 - Moore - Aronszajn

$f, g \in \mathcal{F}$ RKHS

$$f(x) = \sum_i \alpha_i k(x, x_i) \qquad g(x) = \sum_j \beta_j k(x, x_j)$$

w/ inner prod $\langle f, g \rangle = \sum_{ij} \alpha_i \beta_j k(x_i, x_j)$.

- Basic properties and examples

Ex (linear kernel): $k(x,x') = \langle x, x' \rangle$ where $x, x' \in \mathbb{R}^d = \mathcal{X}$.

RKHS for k: $\mathcal{F} = \left\{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) : \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\}$

$$= \left\{ f(x) = \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle : \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\}$$

$$= \left\{ f(x) = \langle x, \underbrace{\sum_{i=1}^{n} \alpha_i x_i}_{\in \mathbb{R}^d} \rangle : \forall n \in \mathbb{N}, \forall x_i \in \mathbb{R}^d, \forall \alpha_i \in \mathbb{R} \right\}$$

$$= \left\{ f(x) = \langle x, \theta \rangle : \theta \in \mathbb{R}^d \right\}.$$

next lecture!

Inner prod. for $\mathcal{F}$: $\quad f(x) = \langle x, \theta_1 \rangle \qquad g(x) = \langle x, \theta_2 \rangle$

$= 1 \cdot k(x, \theta_1) \qquad = \beta_1 k(x, \theta_2)$

$= \alpha_1 k(x, x_1)$

$$\langle f, g \rangle = \langle \theta_1, \theta_2 \rangle.$$

Ex (Common kernels):

1. Identity kernel: $k(x,x') = 1$. Kernel since $k_{ij} = 1$ is PSD.
2. Indicator fnc: $k(x,x') = \mathbb{1}\{\|x - x'\| \le 0\}$. Kernel since $K = I$.
3. Polynomial kernel: $k(x,x') = (1 + \langle x, x' \rangle)^m$ is kernel.
4. Gaussian kernel: $k(x,x') = \exp\left\{ -\frac{1}{2\sigma^2} \|x - x'\|^2 \right\}$.

– Properties

1. Inner prod.: A fnc of the form $k(x,x') = \langle \phi(x), \phi(x') \rangle$ is a kernel (see prev. lecture).

2. Summation: For k_1 and k_2 kernels, $k = k_1 + k_2$ is a kernel.

$$K_1 \succeq 0 \, , \, K_2 \succeq 0 \quad \Rightarrow \quad K = K_1 + K_2 \text{ is PSD.}$$

3. Hadamard product: For k_1 and k_2 kernels, $k = k_1 \cdot k_2$ is a kernel.

$$K_1 \succeq 0, \; K_2 \succeq 0 : \qquad K_1 = \sum_k d_k u_k u_k^T \qquad K_2 = \sum_k b_k v_k v_k^T$$

$$(K = K_1 \circ K_2)_{ij} = (K_1)_{ij} (K_2)_{ij} = \left(\sum_k d_k u_{ki} u_{kj} \right) \left(\sum_l b_l v_{li} v_{lj} \right)$$

$$= \sum_{kl} d_k b_l (u_{ki} v_{li}) \cdot (u_{kj} v_{lj})$$

$$\Rightarrow K = \sum_{kl} \underbrace{d_k \cdot b_l}_{\geq 0} (u_k \circ v_l)(u_k \circ v_l)^T \succeq 0 .$$

– **Polynomial kernel**: $k(x,x') = \underbrace{(1 + \underbrace{\langle x, x' \rangle}_{\text{inner prod.}})}_{\text{sum of two kernels}}{}^{m} \rightarrow$ product rule.

– **Gaussian kernel**: $k(x,x') = \exp\left\{ -\frac{1}{2\sigma^2} \|x - x'\|^2 \right\}$

$$k(x,x') = \underbrace{\exp\left\{ -\frac{\|x\|^2}{2\sigma^2} \right\} \exp\left\{ -\frac{\|x'\|^2}{2\sigma^2} \right\}}_{k_1} \underbrace{\exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\}}_{k_2}$$

k_1 is a kernel since it is an inner prod.

k_2 is a kernel,

$$k_2(x,x') = \exp\left\{ \frac{\langle x, x' \rangle}{\sigma^2} \right\} = \underbrace{\sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\langle x, x' \rangle}{\sigma^2} \right)^i}_{\text{summation of poly kernels}}$$

$\Rightarrow k_1 \cdot k_2$ is a kernel.

– Learning w/ kernels

* Observe a dataset $\mathcal{D} = \{(x_i, y_i) : i = 1 \dots n\}$.
* We have an RKHS $\mathcal{F}$
* Consider $\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$

Theorem (Representer thm): For a dataset $\mathcal{D} = \{(x_i, y_i) : i = 1 \dots n\}$ and a kernel $k(\cdot, \cdot)$, let $V_{\mathcal{D}} = \{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) : \alpha_i \in \mathbb{R} \}$.

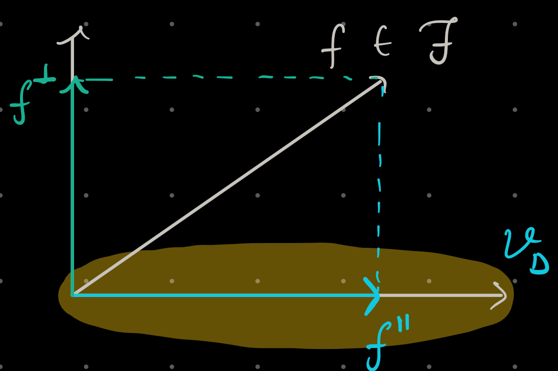
Then, $\hat{f} \in V_{\mathcal{D}}$.

Remark: Algorithmic consequence: Minimizing over $\mathcal{F}$ = Minimizing over $V_{\mathcal{D}}$.

proof:
- $V_{\mathcal{D}}$ is a subspace of $\mathcal{F}$.
- Define orthogonal complement of $V_{\mathcal{D}}$

$$V_{\mathcal{D}}^{\perp} = \{ f' \in \mathcal{F} : \langle f, f' \rangle = 0 \quad \forall f \in V_{\mathcal{D}} \}.$$

(A vector space is the <u>sum</u> of a subspace and its orthogonal complement)



$$f(x) = f^{\parallel}(x) + f^{\perp}(x) \quad \text{where}$$

$$\underset{\in \mathcal{F}}{f^{\parallel} \in V_{\mathcal{D}}} \quad \text{and} \quad \underset{\in \mathcal{F}}{f^{\perp} \in V_{\mathcal{D}}^{\perp}}.$$

– Note that $(x_i, y_i) \in \mathcal{D}$, by the reproducing property of $\mathcal{F}$

$$f^{\perp}(x_i) = \langle \underset{\in V_{\mathcal{D}}^{\perp}}{f^{\perp}}, \underset{\in V_{\mathcal{D}}}{k(x_i, \cdot)} \rangle = 0.$$

$$\Rightarrow f \in \mathcal{F}, \quad f(x_i) = f^{\parallel}(x_i) + f^{\perp}(x_i) = f^{\parallel}(x_i).$$

$$\Rightarrow \ell(y_i, f(x_i)) = \ell(y_i, f^{\parallel}(x_i))$$

- For the regularizer: $\|f\|_{\mathcal{F}}^2 = \|f^{\parallel} + f^{\perp}\|_{\mathcal{F}}^2 = \|f^{\parallel}\|_{\mathcal{F}}^2 + \|f^{\perp}\|_{\mathcal{F}}^2$.

- Thus, $$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2$$

$$= \underset{\substack{f^{\parallel} \in V_D \\ f^{\perp} \in V_D^{\perp}}}{\arg\min}\ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f^{\parallel}(x_i)) + \frac{\lambda}{2}\|f^{\parallel}\|_{\mathcal{F}}^2}_{\substack{-\ f^{\perp} \text{ has no effect.} \\ -\ \text{might as well choose } f^{\perp} = 0.}} + \underbrace{\frac{\lambda}{2}\|f^{\perp}\|_{\mathcal{F}}^2}_{\geq 0.}$$

$$\Rightarrow \hat{f} \in V_D.$$

Ex (Squared error loss): For $\ell(y_i, f(x_i)) = \frac{1}{2}(y_i - f(x_i))^2$

$$\hat{f} = \underset{\mathcal{F}}{\arg\min}\ \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2.$$

By the representer theorem, $\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ for some $\alpha_i \in \mathbb{R}$.

Finding $\hat{f}$ is equal to finding α_i's.

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min}\ \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{n}\alpha_j k(x_i, x_j)\Big)^2}_{ii} + \underbrace{\frac{\lambda}{2}\|f\|_{\mathcal{F}}^2}_{i}$$

i) $\|f\|_{\mathcal{F}}^2 = \langle f, f \rangle = \sum_{ij} \alpha_i k(x_i, x_j)\alpha_j = \alpha^T K \alpha$

ii) $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle k_i, \alpha \rangle)^2 = \frac{1}{2n}\|y - K\alpha\|_2^2$

$K_{ij} = k(x_i, x_j)$

$$k_i = \begin{bmatrix} k(x_i, x_1) \\ \vdots \\ k(x_i, x_n) \end{bmatrix}$$

$$\Rightarrow \hat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\arg\min}\ \frac{1}{2n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T K \alpha$$

$$\Rightarrow \hat{\alpha} = \left(\frac{1}{n}K + \lambda I\right)^{-1}\frac{1}{n}y.$$

(More on this next lecture)

– Maximum Mean Discrepancy (MMD)

Goal: Measure distance between prob. distributions given samples.

Def: $f, f' : \mathcal{X} \to \mathbb{R}$, $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ and

$$\|f - f'\|_\infty = \sup_{x \in \mathcal{X}} \left| f(x) - f'(x) \right| .$$

The following is a way to measure distance between two distributions.

Def (MMD): Let p, q be prob. distributions on $\mathcal{X}$. For some $\mathcal{F} = \{ f : \mathcal{X} \to \mathbb{R} \}$, define

$$d_{\mathcal{F}}(p,q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)] \right| .$$

- How to choose $\mathcal{F}$? (Want $d_{\mathcal{F}}(p,q) = 0 \iff p = q$)
- How to compute $d_{\mathcal{F}}$?

Remark: If $\mathcal{F}$ is 1-Lipschitz fnc, $d_{\mathcal{F}}$ is $\mathcal{L}_1$-Wasserstein metric.

$$\underset{\text{By Monge-Kantorovich duality.}}{d_{\mathcal{L}_1}(p,q) \xrightarrow{\hspace{1em}} W_1(p,q)} \triangleq \inf_{\substack{\text{couplings } (x,y) \\ x \sim p,\ y \sim q}} \mathbb{E}\left[\|x - y\|_2 \right]$$

Theorem (Dudley's MMD thm): For $\mathcal{F} = C_0$ bdd cnts fncs, then $d_{C_0}(p,q) = 0 \iff p = q$.

– $\mathcal{L}_1$ and C_0 are too complex and not practical.

Def (Universal kernel): A kernel k is universal if its RKHS $\mathcal{F}$ is dense in C_0.

- $\mathcal{F}$ is dense in C_0 if for $f \in C_0$, $\forall \epsilon > 0$, $\exists f' \in \mathcal{F}$ s.t.

$$\|f - f'\|_\infty \leq \epsilon.$$

- $\mathcal{F}$ is "representative" of C_0.

Theorem (Stewart's thm): For unit ball $\mathcal{G} = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq 1\}$ where $\mathcal{F}$ is the RKHS of a universal kernel, we have

$$d_{\mathcal{G}}(p,q) = 0 \iff p = q.$$

Proof: $\Leftarrow$ is obvious. For the other side, let $d_{\mathcal{G}}(p,q) = 0$ for some $p \neq q$.

- $p \neq q \Rightarrow d_{C_0}(p,q) = \epsilon > 0$ by Dudley's MMD thm.

$$\Rightarrow \exists h \in C_0 \quad \text{s.t} \quad |\mathbb{E}_p[h(x)] - \mathbb{E}_q[h(x)]| = \epsilon.$$

If $f \notin \mathcal{G}$, rescale h, f, ϵ so $f \in \mathcal{G}$.

- $\mathcal{F}$ is dense in $C_0 \Rightarrow \exists f \in \mathcal{F}$ s.t $\|f - h\|_\infty \leq \frac{\epsilon}{4}$

$$\Rightarrow \underbrace{|\mathbb{E}_p[f(x)] - \mathbb{E}_p[h(x)]| \leq \frac{\epsilon}{4}}_{*} \text{ and } \underbrace{|\mathbb{E}_q[f(x)] - \mathbb{E}_q[h(x)]| \leq \frac{\epsilon}{4}}_{**}$$

- $$\epsilon = |\mathbb{E}_p h(x) - \mathbb{E}_q h(x)| = |\mathbb{E}_p h(x) \pm \mathbb{E}_p f(x) \pm \mathbb{E}_q f(x) - \mathbb{E}_q h(x)|$$

$$\text{(by triangle ineq.)} \leq \underbrace{|\mathbb{E}_p h(x) - \mathbb{E}_p f(x)|}_{* \ \leq \epsilon/4} + \underbrace{|\mathbb{E}_p f(x) - \mathbb{E}_q f(x)|}_{\leq d_{\mathcal{G}}(p,q) = 0} + \underbrace{|\mathbb{E}_q f(x) - \mathbb{E}_q h(x)|}_{** \ \leq \epsilon/4}$$

$$\leq \epsilon/2$$ contradiction. ∎

- We showed that the unit ball in RKHS is good enough.
- But how to compute expectations?

  * By reproducing property

$$\mathbb{E}_p f(x) = \mathbb{E}_p \langle \underset{\text{fixed}}{f}, \underbrace{k(x,\cdot)}_{\text{random}} \rangle = \langle f, \underbrace{\mathbb{E}_p k(x,\cdot)}_{:=\mu_p} \rangle = \langle f, \mu_p \rangle$$

where μ_p is the RKHS embedding of p.

  * MMD becomes:

$$d_{\mathcal{G}}(p,q) = \sup_{f \in \mathcal{G}} |\mathbb{E}_p f(x) - \mathbb{E}_q f(x)|$$

$\mathcal{G} = \{f : \|f\|_{\mathcal{F}} \leq 1\}$

$$= \sup_{f \in \mathcal{G}} |\langle f, \mu_p - \mu_q \rangle|$$

$$= \|\mu_p - \mu_q\|_{\mathcal{F}} \quad \text{(big simplification).}$$

  * $d_{\mathcal{G}}(p,q)^2 = \|\mu_p - \mu_q\|^2_{\mathcal{F}} = \underbrace{\|\mu_p\|^2_{\mathcal{F}}}_{1.} + \underbrace{\|\mu_q\|^2_{\mathcal{F}}}_{2.} - 2\underbrace{\langle \mu_p, \mu_q \rangle}_{3.}$

1. $\|\mu_p\|^2_{\mathcal{F}} = \langle \mu_p, \mu_p \rangle = \langle \mathbb{E}_p k(x,\cdot), \mathbb{E}_p k(x,\cdot) \rangle$

$$= \mathbb{E}_{p,p}[\langle k(x,\cdot), k(x',\cdot) \rangle] \quad x, x' \sim p \text{ indep.}$$

$$= \mathbb{E}_{pp}[k(x,x')].$$

2. $\|\mu_q\|^2_{\mathcal{F}} = \mathbb{E}_{qq}[k(y,y')]$ $\quad y, y' \sim q$ indep.

3. $\langle \mu_p, \mu_q \rangle = \langle \mathbb{E}_p k(x,\cdot), \mathbb{E}_q k(y,\cdot) \rangle = \mathbb{E}_{pq}[k(x,y)]$

$x \sim p$, $y \sim q$ indep.

Plug back in:

$$d_{\mathcal{G}}(p,q)^2 = \mathbb{E}_{pp}\, k(x,x') + \mathbb{E}_{qq}\, k(y,y') - 2\,\mathbb{E}_{pq}\, k(x,y).$$

$$x, x' \sim p \qquad y, y' \sim q \qquad \text{indep.}$$

- Now assume $x_1 \dots x_n \sim p$, $y_1 \dots y_n \sim q$ indep.

Def (U-statistic):

$$U_n \triangleq \frac{1}{\binom{n}{2}} \sum_{i<j} k(x_i,x_j) + k(y_i,y_j) - k(x_i,y_j) - k(x_j,y_i)$$

Remark: - U_n is an unbiased estimator of $d_{\mathcal{G}}(p,q)^2$.

- It is also consistent

$$U_n \xrightarrow{P} d_{\mathcal{G}}(p,q)^2$$

- You can use U_n to measure distance between p and q.

(Generalization: next lecture!)