Kernel Machines

Basic idea: $(X,Y) \in \mathcal{X} \times \{\pm 1\}$

$$\left\{ \begin{array}{l} K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{(Mercer kernel)} \\ (\mathcal{H}_K, \langle \cdot, \cdot \rangle_K) - \text{RKHS} \end{array} \right.$$

$$\mathcal{F} \subset \mathcal{H}_K$$

classifiers: $g_f(x) = \operatorname{sgn} f(x)$

$$\mathcal{F}_\lambda := \{ f \in \mathcal{H}_K : \|f\|_K^2 \le \lambda^2 \} \qquad (\lambda > 0)$$

Rademacher complexities:

$x^n = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$

$$R_n(\mathcal{F}_\lambda(x^n)) = \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]$$

$f(x_i) = \langle f, K_{x_i} \rangle_K$

K_{x_i} — rep. of $K(x_i, \cdot)$ in $\mathcal{H}_K$

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \sum_{i=1}^n \varepsilon_i \langle f, K_{x_i} \rangle_K \right| \right]$$

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \left| \langle f, \underbrace{\sum_{i=1}^n \varepsilon_i K_{x_i}}_{\in \mathcal{H}_K} \rangle_K \right| \right]$$

Cauchy-Schwarz:

$|\langle f, g \rangle_K| \le \|f\|_K \|g\|_K$ for $f, g \in \mathcal{H}_K$

$$\le \frac{1}{n} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}_\lambda} \|f\|_K \left\| \sum_{i=1}^n \varepsilon_i K_{x_i} \right\|_K \right]$$

$$\le \frac{\lambda}{n} \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i K_{x_i} \right\|_K$$

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^{n} \varepsilon_i K_{x_i} \right\|_K = \mathbb{E}_\varepsilon \sqrt{\left\langle \sum_{i=1}^{n} \varepsilon_i K_{x_i}, \sum_{i=1}^{n} \varepsilon_i K_{x_i} \right\rangle_K}$$

$$= \mathbb{E}_\varepsilon \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j \langle K_{x_i}, K_{x_j} \rangle_K}$$

$K(x, x')$
$= \langle K_x, K_{x'} \rangle_K$

$$= \mathbb{E}_\varepsilon \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon_i \varepsilon_j K(x_i, x_j)}$$

$$\le \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) \underbrace{\mathbb{E}[\varepsilon_i \varepsilon_j]}_{\mathbb{1}_{\{i=j\}}}}$$

$$= \sqrt{\sum_{i=1}^{n} K(x_i, x_i)}$$

$$\Rightarrow \boxed{R_n(\mathcal{F}_\lambda(x^n)) \le \frac{\lambda}{n} \sqrt{\sum_{i=1}^{n} K(x_i, x_i)}}$$

Remarks:

- **Gram matrix:** $G = [K(x_i, x_j)]_{i,j=1,\dots,n}$

$$\sum_{i=1}^{n} K(x_i, x_i) = \operatorname{tr} G$$

$$R_n(\mathcal{F}_\lambda(x^n)) \le \frac{\lambda}{n} \sqrt{\operatorname{tr} G} \leftarrow \text{data-dependent}$$

- $X_1, X_2, \dots, X_n$ iid elements of $\mathcal{X}$

$$\mathbb{E} R_n(\mathcal{F}_\lambda(X^n)) \le \frac{\lambda}{n} \mathbb{E} \sqrt{\sum_{i=1}^n K(X_i, X_i)}$$

$$\le \frac{\lambda}{n} \sqrt{n \mathbb{E}[K(X, X)]}$$

$$= \frac{\lambda}{\sqrt{n}} \sqrt{\mathbb{E}[K(X, X)]}$$

or if $\sup_{x \in \mathcal{X}} K(x, x) =: C_K^2 < \infty$,

$$\mathbb{E} R_n(\mathcal{F}(X^n)) \le \frac{\lambda C_K}{\sqrt{n}}.$$

- the bound is dimension-free: $\mathcal{H}_K$ can be finite- or infinite-dim.

$K(x, x') = 1 + \langle x, x' \rangle$ on $\mathbb{R}^d$

$\mathcal{F}_\lambda = \{ x \mapsto \langle w, x \rangle + b : \quad b^2 + \|w\|^2 \le \lambda^2 \}$

$K(x, x') = e^{-a\|x - x'\|^2}$

$\mathcal{H}_K$ spanned by a countably infinite set of basis fcns $\psi_1, \psi_2, \dots$

<u>Surrogate losses:</u>

$(X_1, Y_1), \dots, (X_n, Y_n)$ iid in $\mathcal{X} \times \{\pm 1\}$

$\hat{f}_n \in \mathcal{F}_\lambda$

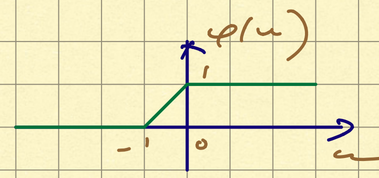
w.p. $\ge 1 - \delta$,

$$L(\operatorname{sgn} \hat{f}_n) \le A_{\varphi, n}(\hat{f}_n) + C_K M_\varphi \frac{\lambda}{\sqrt{n}} + C' \sqrt{\frac{\log(1/\delta)}{n}}$$

- assuming pen. fcn. φ is M_φ-Lipschitz.

e.g. $\varphi(u) = \min\{1, (1+u)_+\}$

or $\varphi(u) \geq \min\{1, (1+u)_+\}$



Kernel trick / Representer Thm

- <u>main idea:</u> restrict opt. to fcns of the form

$$f(x) = \sum_{i=1}^{n} c_i K(X_i, x)$$

where $(X_1, Y_1), \dots, (X_n, Y_n)$ are iid and $c_1, \dots, c_n$ will be tuned based on data.

- <u>orthogonal projection in Hilbert spaces</u>

$(\mathcal{H}, \langle \cdot, \cdot \rangle)$ — Hilbert space

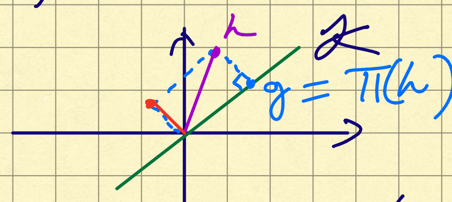
$\mathcal{K}$: closed subspace of $\mathcal{H}$

- $h, h' \in \mathcal{K} \Rightarrow \alpha h + \beta h' \in \mathcal{K}$ $\alpha, \beta \in \mathbb{R}$
  ($\mathcal{K}$ is a subspace of $\mathcal{H}$)

- (h_n) in $\mathcal{K}$ s.t. $h = \lim_{n\to\infty} h_n$ exists
  then $h \in \mathcal{K}$ (<u>closed</u> subset of $\mathcal{H}$)

<u>Then:</u> for any $h \in \mathcal{H}$, the problem

$$\min_{g \in \mathcal{K}} \|g - h\|^2$$



has a unique solution (the projection of h onto $\mathcal{K}$)

$$\Pi(h) = \underset{g \in \mathcal{K}}{\operatorname{argmin}} \|g - h\|^2$$

1) The map $\Pi : \mathcal{H} \to \mathcal{K}$ is linear:

$$\Pi(\alpha h + \beta h') = \alpha \Pi(h) + \beta \Pi(h')$$

2) $\Pi^2 = \Pi \circ \Pi = \Pi$ $\qquad \Pi(\Pi(h)) = \Pi(h)$

3) for any $h \in \mathcal{H}$, $g \in \mathcal{K}$,

$$\langle g, h \rangle = \langle \Pi(g), h \rangle = \langle g, \Pi(h) \rangle$$

4) $\mathcal{K}^\perp := \{ h \in \mathcal{H} : \langle g, h \rangle = 0 \ \forall g \in \mathcal{K} \}$ — orthogonal complement of $\mathcal{K}$ in $\mathcal{H}$ — is also a closed subspace of $\mathcal{H}$, and any $h \in \mathcal{H}$ can be uniquely represented as

$$h = g + g^\perp$$

where $g = \Pi(h)$ and $g^\perp \in \mathcal{K}^\perp$.

Back to kernels —

$(X_1, Y_1), \dots, (X_n, Y_n)$ $\qquad \mathcal{H}_K$ RKHS

$\mathcal{H}_n$: closed subspace of $\mathcal{H}_K$ spanned by $K_{X_1}, K_{X_2}, \dots, K_{X_n}$

— <u>random</u> subspace of $\mathcal{H}_K$

$\Pi_n : \mathcal{H}_K \to \mathcal{H}_n$ — orthogonal proj onto $\mathcal{H}_n$

<u>Representer thm</u> If $\mathcal{F}$ is a subset of $\mathcal{H}_K$ s.t. $\Pi_n(\mathcal{F}) \subseteq \mathcal{F}$, then

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) = \min_{\hat{f} \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \hat{f}(X_i))$$

— minimizer is of the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} c_i K_{X_i}(x)$$

where $c_1, \dots, c_n \in \mathbb{R}$ depend on $(X_1, Y_1), \dots, (X_n, Y_n)$

Proof idea

$$f(X_i) = \langle f, \underbrace{K_{X_i}}_{\in \mathcal{H}_n} \rangle_K \qquad \text{(repr. property)}$$

$$= \langle f, \Pi_n(K_{X_i}) \rangle_K$$

$$= \langle \Pi_n(f), K_{X_i} \rangle_K$$

$$= \langle \tilde{f}, K_{X_i} \rangle_K \qquad \tilde{f} := \Pi_n(f)$$

$$= \tilde{f}(X_i)$$

$$\forall f \in \mathcal{F}, \quad \ell(Y_i, f(X_i)) = \ell(Y_i, \tilde{f}(X_i))$$

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) = \min_{\tilde{f} \in \Pi_n(\mathcal{F})} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \tilde{f}(X_i))$$

— $\hat{f}_n = \underset{\tilde{f} \in \Pi_n(\mathcal{F})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \tilde{f}(X_i))$ is an element of $\mathcal{H}_n$. ∎

Optimization :

$\varphi : \mathbb{R} \to \mathbb{R}$

$$\min_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f(x_i))$$

$\mathcal{F}_\lambda = \{ f \in \mathcal{H}_K : \|f\|_K^2 \le \lambda^2 \}$

$$= \min_{c_1, \dots, c_n} \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(-Y_i \sum_{j=1}^{n} c_j K(x_i, x_j)\Big)$$

$$\text{s.t.} \quad \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j)}_{= c^T G c} \le \lambda^2$$

where $G = [K(x_i, x_j)]_{i,j}$

- by rep. thm, look for $\hat{f}_n(x) = \sum_{i=1}^{n} c_i K_{x_i}(x)$

s.t. $\|\hat{f}_n\|_K^2 \le \lambda^2$

$$\|\hat{f}_n\|_K^2 = \Big\langle \sum_{i=1}^{n} c_i K_{x_i}, \sum_{i=1}^{n} c_i K_{x_i} \Big\rangle_K$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j)$$

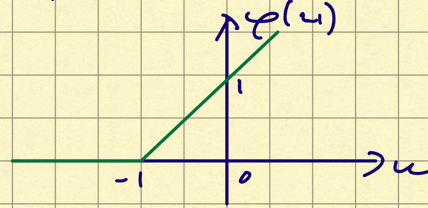
- note: if φ is convex, then

$$\min_{c \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(-Y_i \sum_{j=1}^{n} c_j K(x_i, x_j)\Big)$$

$$\text{s.t.} \quad c^T G c \le \lambda^2$$

is a convex program w/ quadratic constraints!

- take $\varphi(u) = (1+u)_+$ [hinge]



- $K(x,x') = \sum_j \psi_j(x)\psi_j(x')$

where $\psi_1, \psi_2, \ldots$ are basis fcns

$$f(x) = \operatorname{sgn}\left(\sum_j c_j \tilde{\psi}_j(x) \right)$$

let $\psi_j(x) = \sqrt{w_j}\, \tilde{\psi}_j(x)$ $\qquad w_j > 0, \sum_j w_j < \infty$

$$K(x,x') = \sum_j w_j \cdot \tilde{\psi}_j(x) \tilde{\psi}_j(x')$$

$$f(x) = \sum_j \frac{c_j}{\sqrt{w_j}} \psi_j(x)$$

$$\|f\|_K^2 = \sum_j \frac{c_j^2}{w_j}$$

$$\min \frac{1}{n} \sum_{i=1}^n \varphi\left(-Y_i \sum_j c_j \tilde{\psi}_j(x_i)\right)$$

$$\text{s.t.} \quad \sum_j c_j^2 / w_j \le \lambda^2$$

if $\mathcal{H}_K$ is fin-dim, this is more efficient
if $n \gg \dim \mathcal{H}_K$

or if $\tilde{\psi}_1, \tilde{\psi}_2, \ldots$ are arranged s.t.

$w_1 > w_2 > \ldots$, then can set all but finitely many $c_1, \ldots, c_k$ to zero.