Kernel Methods

Reading: Ch. 4 of kc-notes

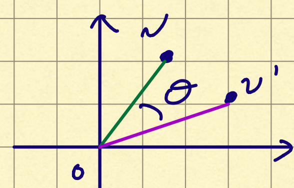
Background:

- $\mathcal{V}$: vector space over $\mathbb{R}$

  $v, v' \in \mathcal{V}$, $\alpha, \beta \in \mathbb{R}$ $\qquad \alpha v + \beta v' \in \mathcal{V}$

- inner product on $\mathcal{V}$: $\langle \cdot, \cdot \rangle_{\mathcal{V}}$
  - linear in each argument:
    $\langle \alpha v + \beta v', w \rangle_{\mathcal{V}} = \alpha \langle v, w \rangle_{\mathcal{V}} + \beta \langle v', w \rangle_{\mathcal{V}}$
  - symmetric: $\langle v, w \rangle_{\mathcal{V}} = \langle w, v \rangle_{\mathcal{V}}$
  - positive definite: $\langle v, v \rangle_{\mathcal{V}} \geq 0$, $\|v\|_{\mathcal{V}} = 0$ iff $v = 0$

inner product – "angle"



$$\langle v, v' \rangle = v^T v' = \|v\| \cdot \|v'\| \cdot \cos\theta$$

$$\cos\theta = \frac{\langle v, v' \rangle}{\|v\| \cdot \|v'\|}$$

- inner product space:
  $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$

  $\mathbb{R}^d$, $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ usual dot product

  $\ell^2 := \{ c = (c_1, c_2, \dots) \in \mathbb{R}^{\mathbb{N}} : \sum_i c_i^2 < \infty \}$

  $\langle c, c' \rangle_{\ell^2} = \sum_i c_i c_i'$

  $L^2([0,1])$ – $f : [0,1] \to \mathbb{R}$

  $\langle f, g \rangle_{L^2} = \int_0^1 f(t) g(t)\, dt$

inner product $\longrightarrow$ norm ("length")

$\langle \cdot, \cdot \rangle_V \longrightarrow \|\cdot\|_V$

$\|v\|_V := \sqrt{\langle v, v \rangle_V}$

$\mathbb{R}^d$: $\|v\|$ .. Euclidean norm, $\sqrt{\sum_{i=1}^d v_i^2}$

ℓ^2: $\|c\|_{\ell^2} = \sqrt{\sum_i c_i^2}$

$L^2([0,1])$: $\|f\|_{L^2} = \sqrt{\int_0^1 f(t)^2 \, dt}$ ...

__Properties of norms:__ $\|\cdot\|_V : V \to \mathbb{R}_+$

- $\|\alpha v\|_V = |\alpha| \cdot \|v\|_V$ (homogeneity) $\quad v \in V$, $\alpha \in \mathbb{R}$
- $\|v\|_V \geq 0$, $\|v\|_V = 0$ iff $v = 0$ (nondegeneracy)
- $\|v + v'\|_V \leq \|v\|_V + \|v'\|_V$ (triangle inequality)

$\|\cdot\|_V := \sqrt{\langle \cdot, \cdot \rangle_V}$ is indeed a norm.

__Cauchy-Schwarz inequality:__

$(V, \langle \cdot, \cdot \rangle_V) \longrightarrow (V, \|\cdot\|_V)$

$$\boxed{|\langle v, v' \rangle_V| \leq \|v\|_V \|v'\|_V}$$

__Proof idea:__ $\langle v + \lambda v', v + \lambda v' \rangle_V \geq 0 \quad \forall \lambda \in \mathbb{R}$

$$\langle v+\lambda v', v+\lambda v'\rangle_{\mathcal{V}} = \langle v,v\rangle_{\mathcal{V}} + 2\lambda\langle v,v'\rangle_{\mathcal{V}} + \lambda^2\langle v',v'\rangle_{\mathcal{V}}$$
$$= \|v\|_{\mathcal{V}}^2 + 2\lambda\langle v,v'\rangle_{\mathcal{V}} + \lambda^2\|v'\|_{\mathcal{V}}^2$$

– quadratic fcn of λ, always ≥ 0.

$F(\lambda) = a\lambda^2 + b\lambda + c \qquad b^2 - 4ac \geq 0$

$a = \|v'\|_{\mathcal{V}}^2, \quad b = 2\langle v,v'\rangle_{\mathcal{V}}, \quad c = \|v\|_{\mathcal{V}}^2$

<u>Recommendation:</u> M. Steele "The Cauchy-Schwarz Master Class"

<u>Note:</u> not all norms are induced by $\langle \cdot,\cdot\rangle$

Ex. $\mathbb{R}^d$

$\|v\|_\infty := \max_{1\leq i\leq d} |v_i|$

$\|v\|_p := \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$ – only $p=2$ comes from $\langle\cdot,\cdot\rangle$

$(p \geq 1)$

<u>Hilbert Space</u> $\quad (\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$

$\|h\|_{\mathcal{H}} = \sqrt{\langle h,h\rangle_{\mathcal{H}}}$

– complete: every Cauchy sequence $(h_i)_{i\geq 1}$ in $\mathcal{H}$ has a limit in $\mathcal{H}$

$(h_i)_{i\geq 1}$

$h_i \to h \iff \lim_{i\to\infty} \|h_i - h\|_{\mathcal{H}} = 0$

$\|h_i - h_k\|_{\mathcal{H}} \leq \|h_i - h\|_{\mathcal{H}} + \|h - h_k\|_{\mathcal{H}} \xrightarrow[i,k\to\infty]{} 0$

$$\lim_{\min(i,k)\to\infty} \|h_i - h_k\|_{\mathcal{H}} = 0 \quad (\text{Cauchy seq.})$$

Hilbert spaces

- $(\mathbb{R}^d, \|\cdot\|_2)$
- $(\mathbb{R}^d, \|\cdot\|_A)$ $\qquad A = A^T, \; A > 0$

$$\|v\|_A = \sqrt{\langle v, Av\rangle} = \sqrt{v^T A v}$$

- $(\ell^2, \|\cdot\|_{\ell^2})$
- $(L^2([0,1]), \|\cdot\|_{L^2})$ $\qquad$ (complete: Riesz-Fischer thm.)
- $(\Omega, \mathcal{F}, \mathbb{P})$ - prob. space

$X : \Omega \to \mathbb{R}$

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) := \left\{X : \mathbb{E}X^2 = \int_\Omega X^2(\omega)\,\mathbb{P}(d\omega) < \infty\right\}$$

$$\|X\|_2 := \sqrt{\mathbb{E}|X|^2}$$

Kernels

- motivation (sketch)

$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ iid in $\mathbb{R}^d \times \{\pm 1\}$

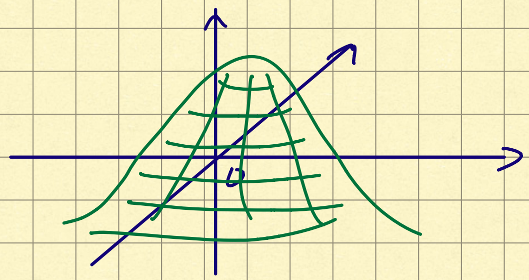
$x \mapsto \operatorname{sgn} f(x)$

Aizerman - Braverman - Rozonoer (1960s)

– 'potential' fcn $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$

$K(x, x') = K(x', x)$ $\qquad$ (symmetric)

if $\|x - x'\| \to \infty$, then $K(x, x') \to 0$



fix x^*

$f(x) := K(x, x^*)$

as x moves away from x^*,

$K(x, x^*)$ decreases

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

s.t. $c_i > 0$ if $y_i = +1$

$c_i < 0$ if $y_i = -1$

— $f(x) = \sum_{i: y_i = +1} |c_i| K(x, x_i) - \sum_{i: y_i = -1} |c_i| K(x, x_i)$

$x \mapsto \operatorname{sgn} f(x)$ — classifier

$c_1, \dots, c_n$: learned from data

— general idea: classifiers of the form

$$\operatorname{sgn} f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} c_i K(x, x_i)\right)$$

$c_1, \dots, c_n \in \mathbb{R}$, tuned to minimize classification error

Hilbert spaces?

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

for some fcns $\psi_1, \psi_2, \dots : \mathbb{R}^d \to \mathbb{R}$

$$K(x, x_i) = \sum_j \underbrace{\psi_j(x_i)}_{\in \mathbb{R}} \psi_j(x) \in \operatorname{span}\{\psi_1, \psi_2, \dots\}$$

so $\psi_1(x), \psi_2(x), \ldots$ are "features" of x!

$$\left.\begin{aligned} f(x) &= \sum_i c_i \psi_i(x) \\ g(x) &= \sum_i c_i' \psi_i(x) \end{aligned}\right\} \quad \langle f, g \rangle_\psi = \sum_i c_i c_i'$$

- we'd like to work w/ functions in $\mathrm{Span}(\psi_1, \psi_2, \ldots)$
- how do we choose ψ_i's?
- how expressive are such fcn spaces?

Kernel trick: design algos that use only the values $K(x_i, x_j)$ $\quad i,j = 1, \ldots, n$

Reproducing Kernel Hilbert spaces (RKHS)

Def (Mercer kernel)

Let $\mathcal{X}$ be a closed subset of $\mathbb{R}^d$.

A fcn $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel if the following holds:

- symmetry: $K(x, x') = K(x', x)$
- continuity: if $x_n \to x$ in $\mathcal{X}$, then
$$\lim_{n \to \infty} K(x_n, x') = K(x, x') \qquad \forall x' \in \mathcal{X}$$
- positive semidefiniteness: for any $n \in \mathbb{N}$ and $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the matrix
$$[K(x_i, x_j)]_{i,j=1,\ldots,n}$$
is positive semidefinite.

$$G_K(x_1,\dots,x_n) := [K(x_i,x_j)]_{i,j=1,\dots,n}$$

— Gram matrix

for any $\alpha_1,\dots,\alpha_n \in \mathbb{R}$, $\sum_{i,j} \alpha_i \alpha_j K(x_i,x_j) \geq 0$.

<u>Examples</u>

- $\mathcal{X} = \mathbb{R}^d$

  $K(x,x') = \langle x,x' \rangle$

- $\mathcal{X} = \mathbb{R}^d$

  $K(x,x') = (1+\langle x,x' \rangle)^k$ $\quad (k \geq 1)$

- $\mathcal{X} = \mathbb{R}^d$

  $K(x,x') = k(x-x')$, where

  $k : \mathbb{R}^d \to \mathbb{R}$ is reflection-sym.

  $k(-x) = k(x)$

$K(x,x') = k(x-x')$ is a Mercer kernel when

$$\hat{k}(\xi) := \int_{\mathbb{R}^d} e^{-i\langle \xi, x\rangle} k(x)\,dx \geq 0$$

(Fourier transform of k)

$k(x) = e^{-a\|x\|^2}$ $\quad a > 0$ (Gauss)

$k(x) = \dfrac{1}{1+\gamma\|x\|^2}$ $\quad (\gamma > 0)$ (Cauchy)

<u>RKHS:</u> let K be given

$\forall x \in \mathcal{X}$, $K_x(\cdot) := K(x,\cdot)$ — cont. on $\mathcal{X}$

$$\mathcal{L}_K(\mathcal{X}) := \text{span}\{K_x : x \in \mathcal{X}\}$$

$$f \in \mathcal{L}_K(\mathcal{X}) \iff f(x) = \sum_{j=1}^{n} c_j K(x, x_j)$$

for some $n \in \mathbb{N}$, $c_1, \dots, c_n \in \mathbb{R}$, $x_1, \dots, x_n \in \mathcal{X}$

Observations:

- $\mathcal{L}_K(\mathcal{X})$ is a vector space
- $K_x \in \mathcal{L}_K(\mathcal{X})$ for all $x \in \mathcal{X}$
- inner product: $f, g \in \mathcal{L}_K(\mathcal{X})$

$$f(x) = \sum_{j=1}^{m} c_j K(x, x_j)$$

$$g(x) = \sum_{l=1}^{n} c_l' K(x, x_l')$$

$$\langle f, g \rangle_K := \sum_{j=1}^{m} \sum_{l=1}^{n} c_j c_l' K(x_j, x_l')$$

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

extend to $\mathcal{L}_K(\mathcal{X})$ by linearity

$\mathcal{H}_K$: Hilbert space obtained by "completing" $\mathcal{L}_K(\mathcal{X})$ [include limits of all Cauchy seqs]

→ Hilbert space of cont. $f : \mathcal{X} \to \mathbb{R}$, s.t.
- $K_x \in \mathcal{H}_K$ $\forall x$
- $g \in \mathcal{H}_K$ can be approx. by some $\hat{g} \in \mathcal{L}_K(\mathcal{X})$

- $g \in \mathcal{H}_K \Rightarrow g(x) = \langle g, K_x \rangle_K$

(reproducing kernel property)

<u>Note:</u>

- $\mathcal{H}_K \subseteq C(\mathcal{X}; \mathbb{R})$ (inclusion may be strict)
- $\mathcal{H}_K$ depends on the choice of K
- build $K(x,x') = \sum_i \varphi_i(x)\varphi_i(x')$ by choosing $\varphi_1, \varphi_2, \ldots$

Ex.: $K(x,x') = \exp\left(-\frac{1}{2}(x-x')^2\right)$ on $\mathbb{R}$

$$K(x,x') = \sum_{i=0}^{\infty} \varphi_i(x)\varphi_i(x'),$$

$$\varphi_i(x) := e^{-x^2/2} \frac{x^i}{\sqrt{i!}}.$$