Kernel methods : $f_\theta(x) = \theta^T \varphi(x)$ $\varphi(x) \in \mathbb{R}^d$

↳ feature map

Linear models

⇒ estimation error was bounded as $\frac{1}{\sqrt{n}}$ independent of dimension

⇒ Convex optimization

⇒ $d = +\infty$ ⇒ $f_\theta(x) = \langle \theta, \varphi(x) \rangle$

$\theta \in \mathcal{H}$ Hilbert space

$\varphi(x) \in \mathcal{H}$

⚠ kernels ≠ kernels

conditionally positive-definite

$$\hat{f}(x) = \frac{\sum_i k(x, x_i)\, y_i}{\sum_i k(x, x_i)}$$

Representer theorem

$x_1, \dots, x_n \in \mathcal{X}$
$y_1, \dots, y_n \in \mathcal{Y}$

$f_\theta(x) = \langle \theta, \varphi(x) \rangle$

where $\varphi : \mathcal{X} \to \mathcal{H}$ Hilbert space

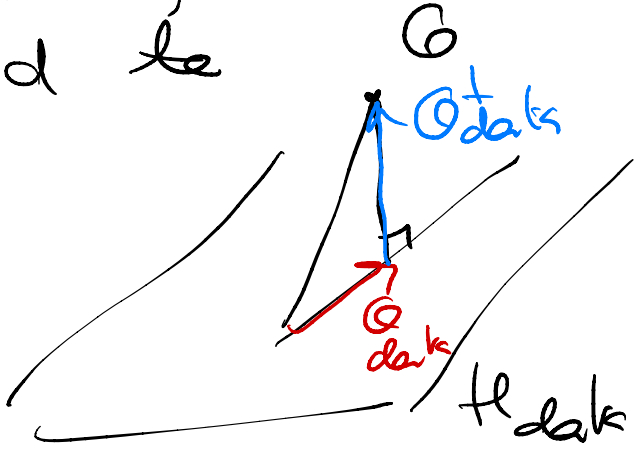
Regularized ERM

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2_{\mathcal{H}}$$

NO CONVEXITY

Theorem (Kimeldorf, Wahba, 1971) If $\lambda > 0$, the minimization problem can be reduced to

$$\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i), \quad \alpha \in \mathbb{R}^n$$



Proof: $\mathcal{H}_{data} = \left\{ \sum_{i=1}^{n} \alpha_i \varphi(x_i), \alpha \in \mathbb{R}^n \right\}$

Pythagore theorem: $\theta = \theta_{data} + \theta^{\perp}_{data}$

$\in \mathcal{H}_{data}$ $\quad$ $\in \mathcal{H}^{\perp}_{data}$

$\|\theta\|^2 = \|\theta_{data}\|^2 + \|\theta^{\perp}_{data}\|^2$

$\forall i$.

$\langle \varphi(x_i), \theta \rangle = \langle \varphi(x_i), \theta_{data} \rangle$

Any minimizer is such that $\theta^{\perp}_{data} = 0$

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2_{\mathcal{H}}$$

Q: how to solve the problem?

$$\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

$$\langle \theta, \varphi(x_i) \rangle = \left\langle \sum_{j=1}^{n} \alpha_j \varphi(x_j), \varphi(x_i) \right\rangle = \sum_{j=1}^{n} \alpha_j \underbrace{\langle \varphi(x_j), \varphi(x_i) \rangle}_{k(x_i, x_j)}$$

kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$

kernel matrix $K \in \mathbb{R}^{n \times n}$, $K_{ij} = k(x_i, x_j)$

$$= (K\alpha)_i$$

$$\|\theta\|^2 = \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j \underbrace{\langle \varphi(x_i), \varphi(x_j) \rangle}_{k(x_i, x_j) = K_{ij}}$$

$$= \alpha^T K \alpha$$

Kernel trick

Equivalent problem =

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^T K \alpha$$

Definition (positive-definite kernel = p.d. kernel)

$h: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a p.d. kernel

$\Leftrightarrow$ all matrices of pairwise evaluations are symmetric (PSD) positive semi-definite

$\Leftrightarrow \forall x_1, \dots, x_n \in \mathcal{X}$, $K \in \mathbb{R}^{n \times n}$ s.t. $K_{ij} = h(x_i, x_j)$

K is PSD

- all eigenvalues are non-negative
- $\Leftrightarrow \alpha^T K \alpha \geq 0 \quad \forall \alpha \in \mathbb{R}^n$

Theorem (Aronszajn, 1950): $h: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is p.d.

$\Leftrightarrow$ there exists $\varphi: \mathcal{X} \to \mathcal{H}$, s.t. $\forall x, x', h(x, x') = \langle \varphi(x), \varphi(x') \rangle$

proof of $\Leftarrow$: $\alpha^T K \alpha = \sum_{i,j}^{n} \alpha_i \alpha_j K_{ij} = \sum_{i,j} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle$

$$= \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|^2 \geq 0$$

p.d kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ $\Rightarrow$ there exists $\varphi : \mathcal{X} \to \mathcal{H}$

s.t $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$

φ: feature map; $\mathcal{H}$: feature space

Two views for learning:

① $f_\theta(x) = \langle \theta, \varphi(x) \rangle$ linear model $\theta \in \mathcal{H}$ with penalty $\|\theta\|^2$

② $f : \mathcal{X} \to \mathbb{R}$ predictor functions

$\mathcal{F} = \{ f, f = f_\theta \text{ for some } \theta \in \mathcal{H} \}$

$\|f\|_{\mathcal{F}}^2 = \inf \|\theta\|^2$ s.t $f = f_\theta$

$$\inf_{\theta \in \mathcal{H}} \hat{R}(f_\theta) + \frac{\lambda}{2} \|\theta\|_{\mathcal{H}}^2 = \inf_{f \in \mathcal{F}} \hat{R}(f) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$$

Kernel calculus : $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ p.d

$k_1 + k_2$ and $k_1 k_2$ are also p.d

$\sum k_n$ p.d if ≥ 0

Linear kernel : $\mathcal{X} = \mathbb{R}^d$ $k(x,y) = x^T y$

$\varphi(x) = x$ explicit feature map

space of functions = linear functions $\theta^T x$ with penalty $\|\theta\|_2^2$ ℓ_2-norm

if $d > n$, representer theorem is useful if $k(x,x')$ easy to compute.

naive running time to compute kernel matrix $O(n^2 d)$

polynomial kernel: $x \in \mathbb{R}^d$ $\quad k(x,x') = (x^\top x')^s$ $\quad s \in \mathbb{N}$, $s \geq 1$

k is p.d as a product of p.d kernels.

$$k(x,x') = \left(\sum_{j=1}^{d} x_j x'_j\right)^s = \sum_{\substack{\alpha_1 + \dots + \alpha_d = s \\ \alpha_j \geq 0}} \binom{s}{\alpha_1 \dots \alpha_d} (x_1 x'_1)^{\alpha_1} \dots (x_d x'_d)^{\alpha_d}$$

$$x_1^{\alpha_1} \dots x_d^{\alpha_d} \cdot x_1'^{\alpha_1} \dots x_d'^{\alpha_d}$$

binomial theorem

$$\frac{s!}{\alpha_1! \dots \alpha_d!}$$

$$= \varphi(x)^\top \varphi(x')$$

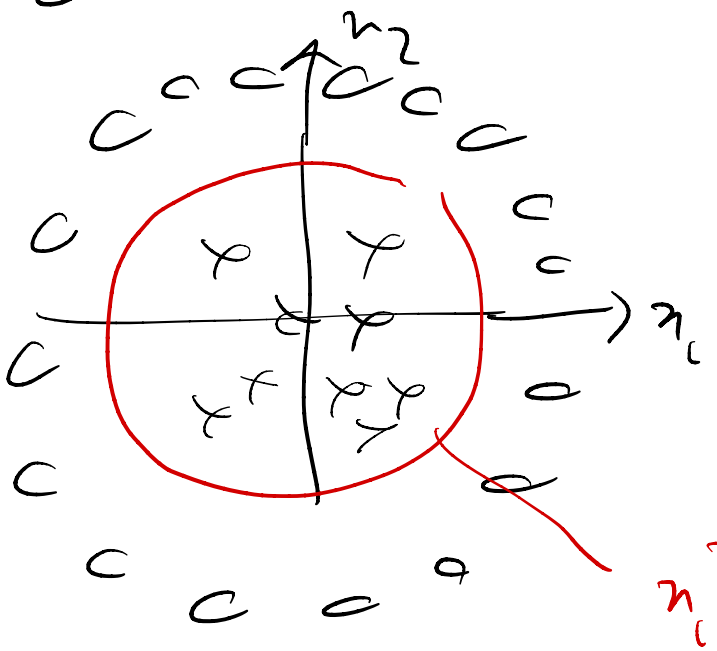
where $\varphi(x)_{\alpha_1 \dots \alpha_d} = \sqrt{\binom{s}{\alpha_1 \dots \alpha_d}}\, x_1^{\alpha_1} \dots x_d^{\alpha_d}$

The dimension of $\varphi(x)$ is $\binom{d+s-1}{s} \sim d^s$

space of functions is homogeneous polynomials of degree s

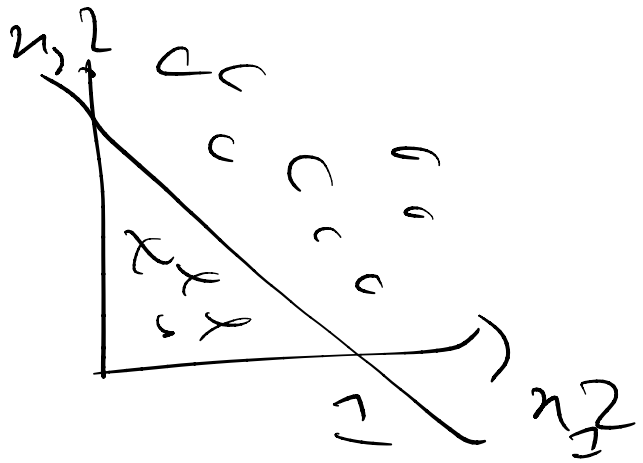
$d=2$

$s=2$, $\varphi(x) = (x_1^2, x_2^2, x_1 x_2)$



$x_1^2 + x_2^2 = 1$

$\Rightarrow$



Translation-invariant kernels : $\mathcal{X} = [0,1]$.

$k(x,x') = q(x-x')$ $\quad$ q 1-periodic $\quad$ $q(x+1) = q(x) \ \forall x$

Fourier series : $f : \mathbb{R} \to \mathbb{R}$ is 1-periodic and $\int_0^1 f(x)^2$ finite

or the natural basis : $\int_0^1 e^{2i\pi m x} \left(e^{2i\pi m' x}\right)^* dx = 0$

$e_m : x \mapsto e^{2i\pi m x}$

$m \in \mathbb{Z}$

$$f = \sum_{m \in \mathbb{Z}} \underbrace{\langle f, e_m \rangle}_{L^2([0,1])} e_m$$

$$f(x) = \sum_{m \in \mathbb{Z}} \hat{f}_m e^{2i\pi m x}$$ Fourier series

with $\hat{f}_m = \langle f, e_m \rangle = \int_0^1 f(x) e^{-2i\pi m x} dx$

Parseval theorem

$$\int_0^1 |f|^2 = \sum_{m \in \mathbb{Z}} |\hat{f}_m|^2$$

If f is differentiable

$$f'(x) = \sum_m \hat{f}_m (2i\pi m) e^{2i\pi m x}$$

Fourier coeff.

c^{425}

c^{425}

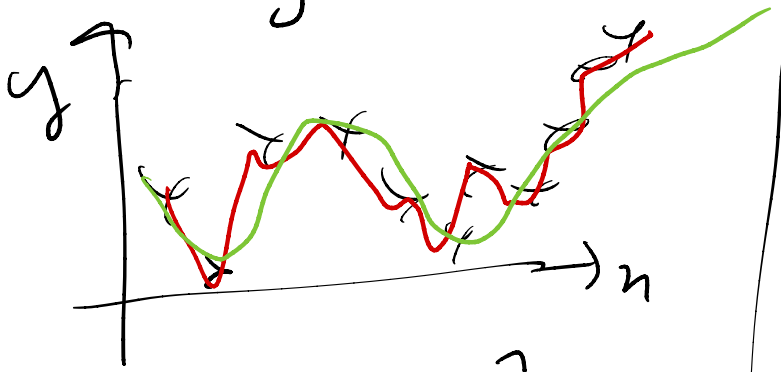
Consequence : $\int_c^1 |G'(x)|^2 = \sum_m |\text{Fourier coeff of } G'|^2$

$= \sum_m |2m\pi|^2 |\hat{G}_m|^2$

Sobolev norm $= \int |G|^2 + \int |G'|^2 = \sum_m (1 + 4\pi^2 m^2) |\hat{G}_m|^2$

$SU(G) = \|G\|^2 + \int |G'|^2$

Learning problem: $\min R(\hat{G}) + \frac{\lambda}{2} \|G\|^2$



$\varphi(x)_m = \frac{e^{+2i m \pi x}}{\sqrt{1 + 4\pi^2 m^2}}$

$f_G(x) = \langle G, \varphi(x) \rangle$

Fourier series $\hat{f}_G(x) = \sum_m \underbrace{(\hat{f}_G)_m \sqrt{1 + 4\pi^2 m^2}}_{\alpha_m} \frac{e^{2i m \pi x}}{\sqrt{1 + 4\pi^2 m^2}}$

$\lambda = 0$ $\quad \lambda >> c$ small

$\|G\|^2 = \sum_m |\alpha_m|^2 = SU(f_G)^2$

kernel: $k(x,x') = \langle \varphi(x), \varphi(x') \rangle = \sum_m \frac{1}{1 + 4\pi^2 m^2} e^{2i m \pi (x - x')} = \varphi(x - x')$

Translation-invariant kernels in $\mathbb{R}^d$

examples
$k(x,x') = e^{-\alpha \|x-x'\|_2^2}$

$\mathcal{X} = \mathbb{R}^d$, $k(x,x') = q(x-x')$.

question: when is this a kernel?

Fourier transform $= \hat{q}(\omega) = \int_{\mathbb{R}^d} q(x) e^{-i\omega^T x} dx$

inversion formula $= q(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i\omega^T x} d\omega$

Bochner theorem: k is p.d $\iff$ $\hat{q}(\omega) \geq 0$ $\forall \omega \in \mathbb{R}^d$

$k(x,x') = q(x-x')$

partial proof $\Leftarrow$

$$\alpha^T K \alpha = \sum_{s,j} \alpha_s \alpha_j q(x_s - x_j) = \frac{1}{(2\pi)^d} \sum_{s,j} \alpha_s \alpha_j \int \hat{q}(\omega) e^{i\omega^T (x_s - x_j)} d\omega$$

using FT

$$= \frac{1}{(2\pi)^d} \sum_{s,j} \alpha_s \alpha_j \int \hat{q}(\omega) e^{i\omega^T x_s - i\omega^T x_j} d\omega$$

$$= \frac{1}{(2\pi)^d} \int \hat{q}(\omega) \sum_{s,j} \alpha_s \alpha_j e^{i\omega^T x_s} e^{-i\omega^T x_s} d\omega$$

$$= \frac{1}{(2\pi)^d} \int \hat{q}(\omega) \left| \sum_j \alpha_j e^{i\omega^T x_j} \right|^2 d\omega \geq 0$$

$\sum_{s,j} \alpha_j \alpha_s e^{i\omega^T x_j} \left(e^{i\omega^T x_s}\right)^*$

Summary = ① if q has non negative Fourier transform
$k(x,x') = q(x-x')$ is a p.d kernel

② the associated norm of functions $\hat{}$

$$\|f\|^2 = \int \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} d\omega$$

Examples = gaussian kernel $q(x) = e^{-\|x\|_2^2}$

Exponential kernel — $e^{-\|x\|_2}$

---

kernels beyond $\mathcal{X} = \mathbb{R}^d$

sequence on alphabet $\mathcal{A}$

$x_1 x_2 \dots x_n$, where $x_i \in \mathcal{A}$

$\varphi(x)$ indexed by all sequences of length ℓ.

$\varphi(x)_y$ = # ℓ y is contained in x

$\varphi(x)^T \varphi(x')$ counts # common subsequences

2

Algorithm + Analysis

goal: $\min_f \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_H^2$

① **Representer theorem:**

$f(x) = \langle \Theta, \varphi(x)\rangle$ and $\Theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$

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

α is obtained as the minimizer of $\frac{1}{n}\sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^T K \alpha$

$(y_i - (K\alpha)_i)^2$

ex: least-squares (ridge regression)

$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T K\alpha = F(\alpha)$

Ⓐ $F'(\alpha) = \frac{1}{n}K(K\alpha - y) + \lambda K\alpha = 0 \Leftrightarrow (K^2 + n\lambda K)\alpha = Ky$

one solution is $\alpha = (K + n\lambda I)^{-1} y$ (may not be unique)

Ⓑ GD: Hessian $F''(\alpha) = K^2 + n\lambda K = H$

speed $\sim \frac{\lambda_{max}(H)}{\lambda_{min}(H)}$ = condition #

Alternative: $K = \Phi\Phi^T$

⇒ see book 7.4.1

Analysis $\hat{g} = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \; \hat{R}(g) + \cancel{\lambda \|g\|_{\mathcal{H}}^2}$ such that $\|g\|_{\mathcal{H}} \leq D$

↓ empirical risk

Q: performance on unseen data

$\mathbb{E}(R(\hat{g}) - R_*)$

↳ expected risk = "test error".

$\forall g : \mathcal{X} \to \mathbb{R}$, $R(g) = \mathbb{E}\, \ell(y, g(x))$

$R(\hat{g}) - R_* = R(\hat{g}) - \inf_{\|g\| \leq D} R(g) + \inf_{\|g\| \leq D} R(g) - R_*$

$\|g\| \leq D$ — estimation error

$\|g\| \leq D$ — approximation error

Chapter 4. G Lipschitz (ℓ), $\|\varphi(x)\|^2 \leq R^2$ ($\varphi(x) = h(x, \cdot)$)

$\mathbb{E}(\,) \leq 4GRD/\sqrt{n}$

Approximation error

$\|\theta\| \le D$

From chapter 6: if $\theta_0 = \arg\min R(\theta)$ Bayes predictor

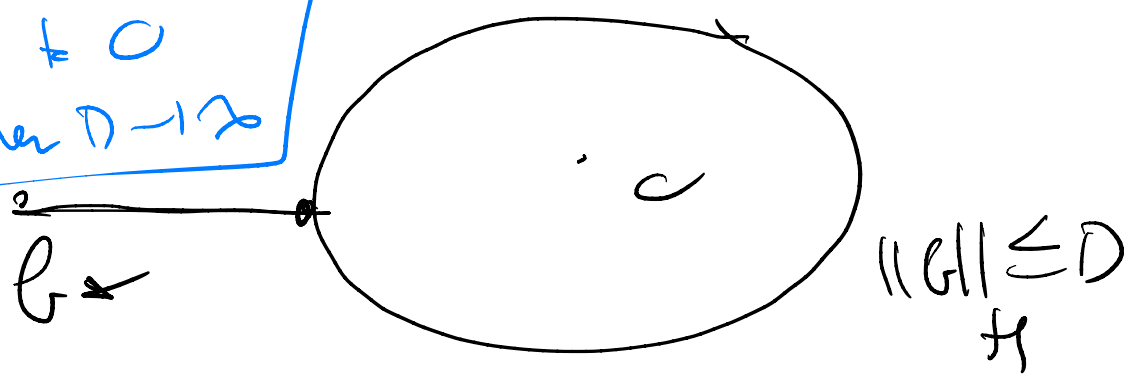
$$R(\theta) - R(\theta_0) \le G \cdot \mathbb{E}\,\|\theta - \theta_0\|_{L^2(P)}$$

$$\left(\mathbb{E}\left(\ell(y, \theta(x)) - \ell(y, \theta_0(x))\right) \le \mathbb{E}\left[G\,|\theta(x) - \theta^*(x)| \right] \right.$$

by say Lipschitz continuity of the loss

$$\text{approx. error} \le G \inf_{\|\theta\| \le D} \|\theta - \theta_0\|_{L^2(P)}$$

Fact: approx. error goes to 0 when $D \to \infty$



In chapter 4: $x \in \mathbb{R}^d$

$\theta(x) = \theta_0^T \varphi(x)$

$\theta^*(x) = \theta^{*T} \varphi(x)$

$\|\theta - \theta_0\|_{L^2} \le R\,\|\theta - \theta_0\|_2$

Note: if $\theta_0 \in \mathcal{H}$, $D \ge \|\theta_0\| \Rightarrow$ approx. error $= 0$

For local averaging: f^* has bounded derivative

excess risk $R(\hat{f}) - R_* \sim \frac{1}{n^{1/d+2}}$ curse of dimensionality

For kernel methods: f^* has s bounded derivatives

excess risk $R(\hat{f}) - R_* \sim \frac{1}{n^{s/d+2s}}$

⚠ if regularization parameter is well-chosen

⇒ adaptive!

For neural nets: even more adaptivity

$$\|f\|^2_{L^1(\rho)} = \mathbb{E}_{\rho(x)} |f(x)|^2$$