Kernels : Representer Theorem
Kernels
Algorithms
Analysis

$k(x,y) \geq 0$ last week
positive definite

Linear models

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

beyond least-squares

$\varphi(x)$ infinite-dimensional
θ

Representer theorem: $f(x) = \langle \theta, \varphi(x) \rangle$ "=" $\theta^T \varphi(x)$

in ML: minimize over $\theta \in \mathcal{H}$

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

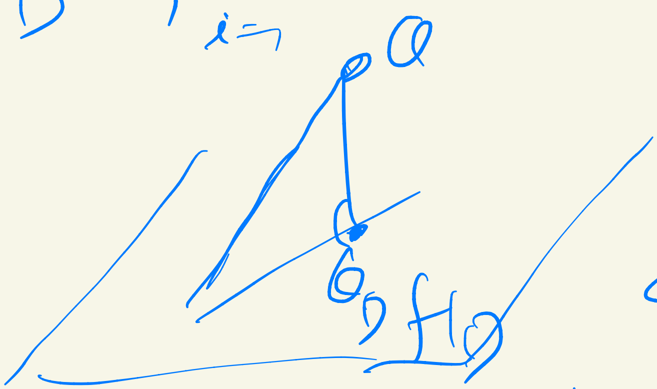
Theorem (Kimeldorf, Wahba, 1971)

The search space can be restricted to $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$ for $\alpha \in \mathbb{R}^n$

Proof: Pythagore

$$\mathcal{H}_D = \left\{ \sum_{i=1}^{n} \alpha_i \varphi(x_i), \ \alpha \in \mathbb{R}^n \right\}$$ vector space $\subset \mathcal{H}$

$$\theta = \theta_D + \theta_\perp, \quad \theta_D \in \mathcal{H}_D, \quad \theta_\perp \in \mathcal{H}_D^\perp$$



$$\langle \theta, \varphi(x_i) \rangle = \langle \theta_D, \varphi(x_i) \rangle + \underbrace{\langle \theta_\perp, \varphi(x_i) \rangle}_{0}$$

$$\|\theta\|^2 = \|\theta_D\|^2 + \|\theta_\perp\|^2 \Rightarrow$$ optimal value of $\theta_\perp = 0$

$$\min_{\theta} \frac{1}{n}\sum_{j=1}^{n} \ell(y_j, \langle \theta, \varphi(x_j)\rangle) + \frac{\lambda}{2}\|\theta\|^2 \quad \text{with } \theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

$$\underbrace{\langle \theta, \varphi(x_j)\rangle = \sum_{i=1}^{n} \alpha_i \boxed{\langle \varphi(x_i), \varphi(x_j)\rangle}}_{\sum_{i=1}^{n} \alpha_i K_{ij} = (K\alpha)_j} \quad k(x_i, x_j)$$

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

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

- kernel function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$
- kernel matrix $K \in \mathbb{R}^{n \times n}$, $K_{ij} = k(x_i, x_j)$

Equivalent problem

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

KERNEL TRICK

Definition: Let $\mathcal{X}$ be $\boxed{\text{any}}$ set. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel

iff (a) it is symmetric: $k(x, x') = k(x', x)$

(b) all kernel matrices are positive semi definite (PSD)

$$\forall x_1, \dots, x_n \in \mathcal{X}, \ \forall \alpha \in \mathbb{R}^n \quad \underbrace{\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0}_{\alpha^T K \alpha \geq 0}$$

$\iff K$ is PSD where $K_{ij} = k(x_i, x_j)$

$\iff$ all eigenvalues of K are nonnegative

Theorem (Aronszajn, 1950):

k is a positive definite kernel $\iff$ there exist $\mathcal{H}$ Hilbert space and $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$

feature map — φ; input space — $\mathcal{X}$; feature space — $\mathcal{H}$

Proof: (⇐) If $k(x,x') = \langle \varphi(x), \varphi(x') \rangle$

then $\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|^2$

(⇒) See book

Examples: $\mathcal{X}$, $\mathcal{H}$, φ, functions $f(x) = \langle \theta, \varphi(x) \rangle$
with norm $\|\theta\|$

(1) linear kernel: $\mathcal{X} = \mathbb{R}^d$, $k(x,x') = x^\top x' = \sum_j x_j x_j'$

$\varphi(x) = x$, linear functions

efficient computation, if possible $o(\dim \mathcal{H})$

if x is sparse

② polynomial: $\mathcal{X}=\mathbb{R}^d$, $\mathcal{F}$ = all polynomials of order $s \geq 1$

$$f(x) = \sum_{\alpha_1,\dots,\alpha_d} \theta_{\alpha_1,\dots,\alpha_d}\, x_1^{\alpha_1} \cdots x_d^{\alpha_d} \binom{s}{\alpha_1,\dots,\alpha_d}^{1/2} \text{ with } \alpha_1+\dots+\alpha_d = s$$

$\mathcal{H} = \mathbb{R}^m$, with $m = \binom{d+s-1}{s} \simeq d^s$

$= \langle \theta, \phi(x) \rangle$ with $\phi(x)_{\alpha_1\dots\alpha_d} = x_1^{\alpha_1} \cdots x_d^{\alpha_d} \binom{s}{\alpha_1,\dots,\alpha_d}^{1/2}$

$$k(x,x') = \langle \phi(x), \phi(x') \rangle = \sum_{\alpha_1,\dots,\alpha_d} \binom{s}{\alpha_1,\dots,\alpha_d} x_1^{\alpha_1} \cdots x_d^{\alpha_d} x_1'^{\alpha_1} \cdots x_d'^{\alpha_d}$$

$$\underbrace{\binom{s}{\alpha_1,\dots,\alpha_d}}_{\text{multinomial coeff.}} (x_1 x_1')^{\alpha_1} \cdots (x_d x_d')^{\alpha_d}$$

$$\boxed{\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n}$$

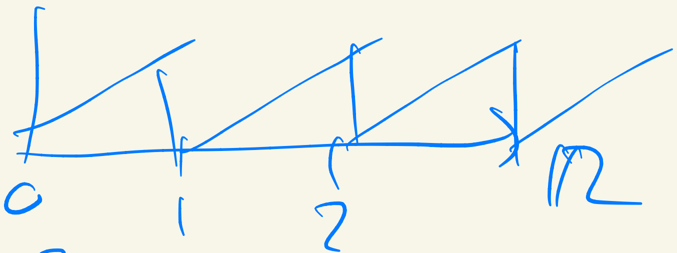
$$= (x_1 x_1' + \dots + x_d x_d')^s = (x^\top x')^s$$

③ translation invariant kernels on $[0,1)$ or $\mathbb{R}^d$

Fourier Series — Fourier transform

Reminder on Fourier Series

$f: \mathbb{R} \to \mathbb{R}$ 1-periodic



restricted to $[0,1)$

$f \in L^2$ then $f(x) = \sum_{n \in \mathbb{Z}} \hat{f}_n \underbrace{e^{2in\pi x}}$

Fourier orthonormal basis of L^2

$\|a\|^2 = \sum_{i=1}^{d} a_i^2$

$$\hat{f}_n = \int_0^1 f(x) e^{-2in\pi x} \, dx$$

Parseval $\Rightarrow$ $\int_0^1 |f(x)|^2 dx = \sum_{n \in \mathbb{Z}} |\hat{f}_n|^2$

$\int |f'(x)|^2 = \sum_{n \in \mathbb{Z}} |\hat{f}_n|^2 (2\pi n)^2$

Derivative $\Rightarrow$ $f'(x) = \sum_{n \in \mathbb{Z}} \hat{f}_n 2in\pi \, e^{2in\pi x}$ $\Rightarrow$ Fourier coefficient of f'

Translation invariant kernels

$k(x,x') = q(x-x')$ where q is 1-periodic

Assume $\sum_{m\in\mathbb{Z}} |\hat{q}_m| < +\infty$.

Prop: k is a positive definite kernel $\iff \forall m,\ \hat{q}_m \geq 0$

Proof:

$$\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) = \sum_{l,j} \alpha_l \alpha_j q(x_l - x_j) = \sum_{l,j} \alpha_l \alpha_j \underbrace{\sum_{m\in\mathbb{Z}} \hat{q}_m e^{2i\pi m(x_l - x_j)}}_{= q(x_l - x_j)}$$

$$= \sum_{m\in\mathbb{Z}} \hat{q}_m \underbrace{\sum_{l,j} \alpha_l \alpha_j e^{2i\pi m x_l} e^{-2i\pi m x_j}}_{\left|\sum_j \alpha_j e^{-2im\pi x_j}\right|^2}$$

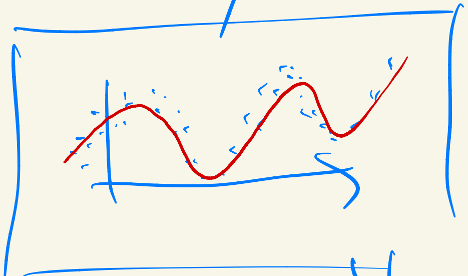
$$= \sum_{m\in\mathbb{Z}} \hat{q}_m \left| \sum_{j=1}^{n} \alpha_j e^{-2im\pi x_j} \right|^2$$

$$|z|^2 = z z^*$$

$$(e^{ib})^* = e^{-ib}$$

$$k(x,x') = q(x-x') = \sum_{m\in\mathbb{Z}} \hat q_m e^{2i\pi m(x-x')} \overset{?}{=} \langle \phi(x), \phi(x')\rangle$$

$$= \sum_{m\in\mathbb{Z}} \hat q_m e^{2im\pi x} \left(e^{2i\pi m x'}\right)^*$$



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

$$\phi(x) \in \mathbb{R}^{\mathbb{Z}} \quad \text{with} \quad \phi(x)_m = \hat q_m^{1/2} e^{2i\pi m x}$$

$$\langle \theta, \phi(x)\rangle = \sum_{m\in\mathbb{Z}} \theta_m \hat q_m^{1/2} e^{2i\pi m x}$$

$$\shortparallel$$

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

if $f(x) = \langle \theta, \phi(x)\rangle$

what is $\|\theta\|^2$?

Ex: $\hat q_m = \frac{1}{m^2}$

$$\sum |\hat f_m|^2 m^2 = \int |f'(x)|^2 dx$$

$$\Rightarrow \|\theta\|^2 = \sum_{m\in\mathbb{Z}} |\theta_m|^2$$

$$= \sum_{m\in\mathbb{Z}} \frac{|\hat f_m|^2}{\hat q_m}$$

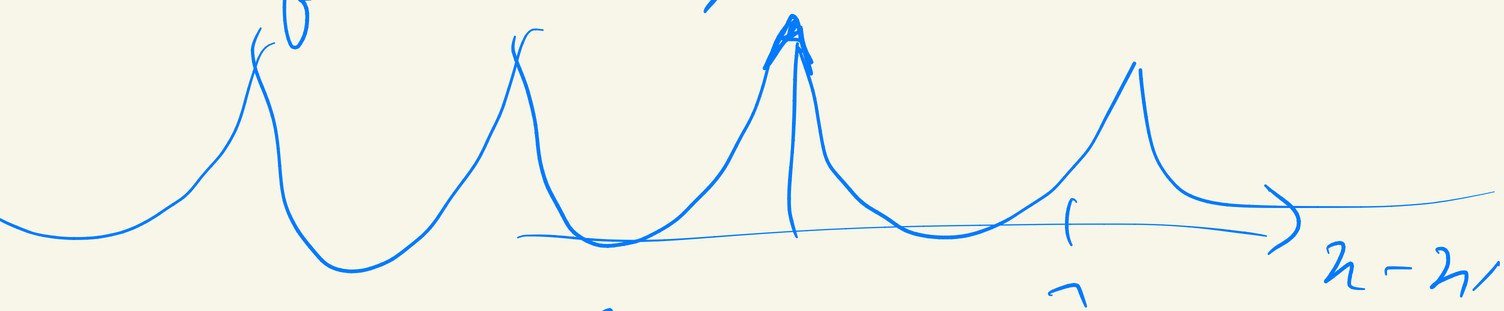
associated penalty

Example: $\hat{q}_0 = (2\pi)^2 \qquad \hat{q}_m = \frac{1}{m^2}$

① coupling $q(x-x') = \sum_{m\in\mathbb{Z}} \hat{q}_m e^{2im\pi(x-x')} = (2\pi)^2 + \sum_{m\in\mathbb{Z}} \frac{e^{2im\pi(x-x')}}{m^2}$

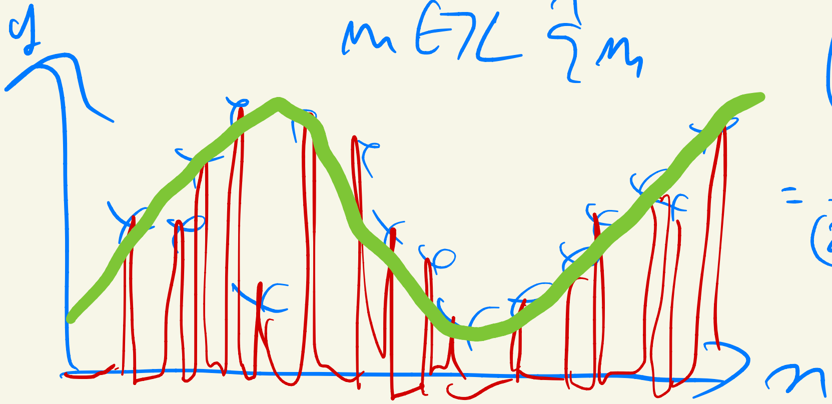
$$= (2\pi)^2 + 2\pi^2\left((x-x')^2 - (x-x') + 1/6\right)$$

if $x-x' \in [0,1]$



② norm: $\sum_{m\in\mathbb{Z}} \frac{|\hat{f}_m|^2}{\hat{q}_m} = \frac{|\hat{f}_0|^2}{(2\pi)^2} + \sum_{m\in\mathbb{Z}} m^2 |\hat{f}_m|^2$

$$= \frac{1}{(2\pi)^2}\left(\int_0^1 f(x)\,dx\right)^2 + \frac{1}{(2\pi)^2}\int_0^1 |f'(x)|^2\,dx$$



Translation-invariant kernels on $\mathbb{R}^d$ $\Rightarrow$ see book

s-th order derivative $\boxed{s > d/2}$

① Algorithms ② Analysis

(x_i, y_i), $i = 1, \dots, n$ sampled iid, kernel function $k(x, y)$

$\in \mathcal{X} \times \mathbb{R}$

$\Rightarrow$ anything

feature map $\Rightarrow$ implicit

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

Algo: representer theorem: $\theta = \sum_i \alpha_i \varphi(x_i)$

+ kernel trick

$\frac{1}{n} \sum_i \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^T K \alpha$

$\boxed{O(n^2)}$

$\Rightarrow$ see $\boxed{\text{book}}$ for more

Analysis: study $\underset{\hat{f}}{\min} \;//\; \theta \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))}_{\hat{R}(f)}$ s.t $f(x) = \langle \theta, \varphi(x) \rangle$, $\|\theta\| \le D$

goal: bound $R(f) = \mathbb{E}\,\ell(y, f(x))$ for $f = \hat{f}$

⚠ $R(\hat{f})$ is random $\Rightarrow$ $\mathbb{E}R(\hat{f}) - R_*$

Decomposition in estimation error + approximation error

$$R(\hat{f}) - R_* \le \underbrace{2\sup_{f\in\mathcal{F}} \left|\hat{R}(f) - R(f)\right|}_{\text{estimation error}} + \underbrace{\inf_{f\in\mathcal{F}} R(f) - R_*}_{\text{approximation error}}$$

$\mathcal{F} = \{ \langle \theta, \varphi(x) \rangle, \; \|\theta\| \le D \}$

$\sim \frac{1}{\sqrt{n}}$

Assumptions (H1) Loss is G-Lipschitz continuous

$z \mapsto \ell(y,z)$ is G-Lipschitz

for all y

From ??

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \le \frac{2GRD}{\sqrt{n}}$$

Estimation Error

(H2) $k(x,x) = \|\varphi(x)\|^2 \le R^2$ almost surely

Radius of $\mathcal{F}$

$\|\theta\| \le D$

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

Approx. error: $\inf_{f \in \mathcal{F}} R(f) - R(f_*)$ Bayes predictor

$$\mathbb{E}\left[\ell(y, f(x)) - \ell(y, f_*(x))\right]$$
$$\leq \mathbb{E}\left[\, G\, |f(x) - f_*(x)|\,\right]$$
using G-Lipschitz $\ell(\cdot)$
$$\leq G \sqrt{\mathbb{E}\, |f(x) - f^*(x)|^2}$$
Jensen's ineq

Summary

$$\mathbb{E}\, R(\hat{f}) - R_* \leq \underbrace{\frac{2GRD}{\sqrt{n}}}_{\text{estimation error}} + \underbrace{G \inf_{\|\theta\| \leq D} \sqrt{\mathbb{E}\, |\langle \theta, \varphi(x)\rangle - f^*(x)|^2}}_{\text{approximation error}}$$

Easy case: well-specified: $f^*(x) = \langle \theta^*, \varphi(x)\rangle$

if $D \geq \|\theta_*\| \Rightarrow 0$

Universality of kernel methods with translation invariant kernels

$$\inf_{D} \frac{2GRD}{\sqrt{n}} + G \inf_{\|\theta\| \le D} \sqrt{\mathbb{E}\,|f_*(x) - \langle \theta, \varphi(x)\rangle|^2}$$

$$\inf_{\|\theta\| \le D} \frac{2GRD}{\sqrt{n}} + G\sqrt{\mathbb{E}\,|f_*(x) - \langle \theta, \varphi(x)\rangle|^2}$$

$$\sup_{\lambda \ge 0} \inf_{\theta, D} \left(\frac{2GRD}{\sqrt{n}} + G\sqrt{\phantom{\mathbb{E}}} \right) + \lambda(\|\theta\| - D)$$

$$D\underbrace{\left[\frac{2GR}{\sqrt{n}} - \lambda\right]}_{=0}$$

$$G \inf_{\|\theta\|} \sqrt{\mathbb{E}\,|f_*(x) - \langle \theta, \varphi(x)\rangle|^2} + \frac{2R}{\sqrt{n}}\|\theta\|$$

$$\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$$

$$\text{Excess risk} \le G\sqrt{2\inf_{\theta} \mathbb{E}\big(f_*(x) - \underbrace{\langle\theta,\varphi(x)\rangle}_{f(x)}\big)^2 + \frac{4R^2}{n}\|\theta\|^2}$$

Translation invariant kernels and $p(x) = \text{uniform on } [0,1]$

$$\inf_f \int_0^1 |f_*(x) - f(x)|^2 dx + \frac{4R^2}{n}\sum_{m\in\mathbb{Z}} \frac{|\hat f_m|^2}{\hat q_m}$$

FS of q

$$\inf_{\hat f} \sum_{m\in\mathbb{Z}} |\hat f_{*m} - \hat f_m|^2 + \frac{4R^2}{n}\frac{|\hat f_m|^2}{\hat q_m}$$

lemma: $\inf_{\hat f_m} |\hat f_m - \hat f_{*m}|^2 + \frac{4R^2}{n}\frac{|\hat f_m|^2}{\hat q_m} = \boxed{\frac{\frac{4R^2}{n}|\hat f_{m*}|^2}{\hat q_m + 4R^2/n}}$

$$\hat f_m - \hat f_{*m} + \frac{4R^2}{n}\frac{\hat f_m}{\hat q_m} = 0 \Rightarrow \hat f_m = \frac{\hat f_{m*}}{1 + \frac{4R^2}{n}\frac{1}{\hat q_m}}$$

Assumption $\hat q_m > c$

goes to 0 as $n \to \infty$

$$\text{Excess risk} \le G\sqrt{2\sum_{m\in\mathbb{Z}} \frac{4R^2}{n}\frac{|\hat f_{m*}|^2}{\hat q_m + 4R^2/n}}$$

≤ 1

Rem:

$$\inf_{x_1,\dots,x_m} \sum_i h_i(x_i) = \sum_i \inf_{x_i} h_i(x_i)$$

$$\inf_x \sum_i h_i(x) \geq \sum_i \inf_x h_i(x)$$

Two more lectures: (1) sparsity

(2) neural networks