

Empirical risk minimization: model $f_0: X \rightarrow \mathbb{R}$
Linear models: $f_0(x) = \theta^T \varphi(x)$ where $\varphi: X \rightarrow \mathbb{R}^d$ feature map
 with ℓ_2 -regularization: $\hat{R}(f_0) + \frac{\lambda}{2} \|\theta\|_2^2$

Today: $\mathbb{R}^d \rightarrow$ Hilbert space kernel methods

Next week: $\|\cdot\|_2 \rightarrow \|\cdot\|_1$ sparse methods

Two weeks: $f_0(x) = \sum_{j=1}^m \eta_j \sigma(w_j^T x + b_j)$ neural networks

Today: kernel methods: $f_0(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$
 where $\varphi: X \rightarrow \mathcal{H}$ Hilbert space
 $\mathbb{R}^d$ d large

Useful: between k -nn.

β^* Lipschitz cont. map
 rate $O\left(\frac{1}{n^{1/d+2}}\right)$

$\Rightarrow \underline{\underline{\beta^* \text{ map regular}}}$
 $\frac{1}{\sqrt{n}}$

• useful for d large

• inferior to neural networks

• fashionable in 2000-2008

⚠ "kernel" is overloaded $\rightarrow$ positive definite
 $\text{best weak} \leftarrow$ non negative (pointwise)
 $\Rightarrow$ 80% pos

Representer theorem (Kimeldorf & Wahba, (971))

ERM: data $(x_i, y_i)_{i=1, \dots, n}$

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

$$\Psi(\langle \theta, \varphi(x_i) \rangle_{H_i}, i=1, \dots, n) + \frac{\lambda}{2} \|\theta\|_H^2 = F(\theta)$$

Theorem: if $\lambda > 0$, the minimum of F can be obtained by restricting θ to the span of $\varphi(x_1), \dots, \varphi(x_n)$

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

Proof: Pythagore:

$$H_D = \left\{ \sum_{i=1}^n \alpha_i \varphi(x_i), \alpha_i \in \mathbb{R} \right\}$$

$$H = H_D \oplus H_D^\perp$$

$$\forall \theta \in H, \theta = \theta_D + \theta_D^\perp$$

$$\|\theta\|^2 = \|\theta_D\|^2 + \|\theta_D^\perp\|^2$$

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

$$F(\theta) = \Psi(\langle \theta_D, \varphi(x_i) \rangle_{i=1, \dots, n}) + \underbrace{\|\theta_D^\perp\|^2}_{=0} + \frac{\lambda}{2} \|\theta_D\|^2$$

minimizing $F(\theta)$ only possible when $\theta_D^\perp = 0 \Leftrightarrow \theta \in H_D$.



! no constraint on Ψ

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

$\phi = \sum_{i=1}^n \alpha_i \phi(x_i)$ (Representer theorem)

$\rightarrow \langle \phi, \phi(x_i) \rangle = \langle \sum_{j=1}^n \alpha_j \phi(x_j), \phi(x_i) \rangle = \sum_{j=1}^n \underbrace{\langle \phi(x_j), \phi(x_i) \rangle}_{K_{ij}} \alpha_j = (\mathbf{K}\alpha)_i$

$\|\phi\|^2 = \left\| \sum_{i=1}^n \alpha_i \phi(x_i) \right\|_2^2 = \sum_{i,j=1}^n \alpha_i \alpha_j \underbrace{\langle \phi(x_i), \phi(x_j) \rangle}_{h(x_i, x_j)} = \sum_{i,j} \alpha_i \alpha_j K_{ij} = \alpha^T \mathbf{K} \alpha$

where $h(x, x') = \langle \phi(x), \phi(x') \rangle$
 $h: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ kernel function

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

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

"kernel trick"

prediction function:
 $\hat{f}_\phi(x) = \langle \phi, \phi(x) \rangle = \sum_{i=1}^n \alpha_i \langle \phi(x_i), \phi(x) \rangle$
 $= \sum_{i=1}^n \alpha_i h(x_i, x)$

Def: positive definite kernel (p.d.)
 function $h: X \times X \rightarrow \mathbb{R}$ such that all kernel matrices are
 symmetric positive semidefinite PSD

$\Leftrightarrow \forall n_1, \dots, n_n \in X$, the matrix $K \in \mathbb{R}^{n \times n}$ is PSD
 s.t. $K_{ij} = h(x_i, x_j)$

$\Leftrightarrow \forall n_1, \dots, n_n \in X$ all eigenvalues are non negative.

$\Leftrightarrow \forall n_1, \dots, n_n \in X; \quad \forall \alpha \in \mathbb{R}^n, \quad \alpha^T K \alpha \geq 0$

Theorem (Aronszajn, 1950). $h: X \times X \rightarrow \mathbb{R}$ is positive definite
 iff there exists a map $\varphi: X \rightarrow H$ the feature space
 s.t. $\forall n, n', \quad h(x, x') = \langle \varphi(x), \varphi(x') \rangle_H$.

Proof of $\Leftarrow$: for any $n_1, \dots, n_n \in X; \alpha \in \mathbb{R}^n$
 $\alpha^T K \alpha = \sum_{i,j} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_H = \left\| \sum_{i=1}^n \alpha_i \varphi(x_i) \right\|_H^2 \geq 0$

N.B.: K often called Gram matrix.

Idea of the proof $\Rightarrow$: let $H =$ set of functions from $X \rightarrow \mathbb{R}$
 defined as $f(x) = \sum_{i=1}^n \alpha_i h(x, x_i), \quad x_i \in X, \alpha_i \in \mathbb{R}$

H : space of functions
 $\varphi(x) = h(\cdot, x)$

! φ "unique"

Examples of p.d kernels: $h: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
space of functions

$\phi: \mathcal{X} \rightarrow \mathbb{R}^d$
 $\phi(x) = \langle \phi, \phi(x) \rangle$
with penalty $\|\phi\|^2$.

$h/\phi/\phi$

kernel calculus: h_1, h_2 are p.d kernels
then $h_1 + h_2$ and $h_1 h_2$ are p.d kernels

① Linear kernel: $\mathcal{X} = \mathbb{R}^d$, $h(x, y) = x^T y = \phi(x)^T \phi(y)$ for $\phi(x) = x$
 $\phi(x) = \phi(x)$ with penalty $\|\phi\|_2^2$

Useful view of kernel method

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

view dual: $\min_{\phi \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (\phi\phi)_i) + \frac{\lambda}{2} \|\phi\|^2$
with $\phi \in \mathbb{R}^d$ design matrix

"matrix inversion lemma"

computing $\hat{y}$ can be done in $O(d)$
if n, y are sparse

$$K = \phi\phi^T \in \mathbb{R}^{n \times n}$$

Empirical covariance matrix

$$\phi = \begin{pmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{pmatrix}$$

$$\hat{\Sigma} = \frac{1}{n} \phi\phi^T \in \mathbb{R}^{d \times d}$$

polynomial kernel: $k(x, x') = (x^T x')^s$ $s \in \mathbb{N}$
 2) $X = \mathbb{R}^d$ is p.d because $el \in V$ a product of kernels

$$\begin{aligned}
 k(x, x') &= \left(\sum_{j=1}^d x_j x'_j \right)^s \\
 &= \sum_{\substack{d_1 + \dots + d_d = s \\ d_i \text{ integer}}} \binom{s}{d_1, \dots, d_d} \underbrace{\left(x_1^{d_1} x_2^{d_2} \dots x_d^{d_d} \right)}_{x_1^{d_1} \dots x_d^{d_d}} \dots \underbrace{\left(x'_1{}^{d_1} x'_2{}^{d_2} \dots x'_d{}^{d_d} \right)}_{x'_1{}^{d_1} \dots x'_d{}^{d_d}} \\
 &\quad \underbrace{\frac{s!}{d_1! \dots d_d!}}_{\text{red}}
 \end{aligned}$$

$$= \varphi(x)^T \varphi(x') \quad \varphi(x) = \left(x_1^{d_1} \dots x_d^{d_d} \right)^{1/2} \quad \text{all monomials of total degree } s$$

functions: $\ell_0(n)$ = degree- s homogeneous polynomial

$\|w\|^2$ = penalty a coefficient

$$\dim(\mathcal{H}) = \binom{d+s-1}{s} \sim d^s$$

NB: to get all polynomials

$$\left(1 + x^T x' \right)^s = \begin{bmatrix} 10^4 & 14 \\ 10^4 & 35 \end{bmatrix}$$



Translation-invariant inner

$[0, 1]$ - Fourier series for function $f: \mathbb{R} \rightarrow \mathbb{R}$ 1-periodic
 $\left(\underbrace{e^{2i\pi m x}}_{e_m} \right)_{m \in \mathbb{Z}}$ is an orthonormal basis of the
 space of periodic functions
 with norm $\int_0^1 |f(x)|^2 dx$
 (complex-valued)

$$f = \sum_{m \in \mathbb{Z}} \underbrace{\langle f, e_m \rangle}_{\widehat{f}_m} e_m$$

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

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

Pythagore / Parseval: $\int_0^1 |f(x)|^2 dx = \sum_{m \in \mathbb{Z}} |\widehat{f}_m|^2$

if f is differentiable,

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

Fourier series of f'

$$\Rightarrow \int_0^1 |f'(x)|^2 dx = 4\pi^2 \sum_{m \in \mathbb{Z}} m^2 |\widehat{f}_m|^2$$

$$\int_0^1 |f^{(s)}(x)|^2 dx = (4\pi^2)^s \sum_{m \in \mathbb{Z}} m^{2s} |\widehat{f}_m|^2$$

Define a penalty on β : $\|\beta\|_C^2 = \sum_{m \in \mathbb{Z}} c_m |\hat{\beta}_m|^2$ norm

ex: $c_m = (1 + 4\pi^2 m^2)$

$\|\beta\|_C^2 = \int |\beta|^2 + \int |\beta'|^2$
Sobolev norm: dense in L^2

feature map

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

$\beta_\theta(x) = \langle \theta, \varphi(x) \rangle = \sum_m \theta_m \frac{e^{2\pi i m x}}{\sqrt{c_m}} \Rightarrow \theta = (\hat{\beta}_\theta)_m \sqrt{c_m}$

$\|\theta\|^2 = \sum_m c_m |\hat{\beta}_\theta)_m|^2 = \|\beta_\theta\|_C^2$

$h(x, x') = \sum_m \varphi_m(x) \varphi_m(x')^* = \sum_{m \in \mathbb{Z}} \frac{1}{c_m} e^{2\pi i m (x - x')}$

$2\pi i m (x - x')$

Translation invariant



$\Rightarrow h(x, x') = \sum_m \frac{1}{1 + 4\pi^2 m^2} e^{2\pi i m (x - x')}$ kernel $\Rightarrow$ closed form

Trans let a kernel be $\mathbb{R}^n$

$$x \in \mathbb{R}^d, h(x, x') = q(x - x') \text{ with } q: \mathbb{R}^d \rightarrow \mathbb{R}$$

Fourier transform = $\hat{f}: \mathbb{R}^d \rightarrow \mathbb{R}$

$$\hat{f}(\omega) = \int f(x) e^{-i\omega^T x} dx \quad \text{F.T.}$$

$$\text{FT inverse} \quad f(x) = \int \hat{f}(\omega) e^{i\omega^T x} d\omega$$

with enough regularity

Prop (Bochner theorem)

$$\rightarrow f(x) = \int \hat{f}(\omega) e^{i\omega^T x} d\omega$$

h is p.d. $\Leftrightarrow \hat{q}$ is positive semi-definite (p.s.d.) $(\hat{q}(\omega) \geq 0 \forall \omega)$

$$\text{p.s.d.} \Leftrightarrow \sum_{s,j=1}^n \alpha_s \alpha_j h(x_s, x_j) = \sum_{s,j=1}^n \alpha_s \alpha_j q(x_s - x_j) = \sum_{s,j=1}^n \alpha_s \alpha_j \int \hat{q}(\omega) e^{i\omega(x_s - x_j)} d\omega$$

$$= \int \left| \sum_{s=1}^n \alpha_s e^{i\omega x_s} \right|^2 \hat{q}(\omega) d\omega \geq 0$$

FT inverse formula $e^{i\omega x_s} e^{-i\omega x_j} = e^{i\omega(x_s - x_j)}$

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

$$\rightarrow \hat{q}(\omega) = e^{-\|\omega\|_2^2} \geq 0$$

Gaussian kernel

$$q(x) = e^{-\|x\|_2}$$

$$\rightarrow \hat{q}(\omega) = \frac{1}{(1 + \|\omega\|_2^2)^{\frac{d+1}{2}}}$$

exponential kernel

hard to do this

Def: the norm associated with the kernel $q(x, y)$
 is $\|f\|^2 = \frac{1}{(2\pi)^d} \int \frac{|\hat{f}(u)|^2}{\hat{q}(u)} du$ (see book for proof).

ex: $\hat{q}(u) = \frac{1}{(1+\|u\|^2)^{\frac{d+n}{2}}}$ $\Rightarrow \|f\|^2 = \int |\hat{f}(u)|^2 (1+\|u\|^2)^{\frac{d+n}{2}} du$

$q(x) = e^{-\|x\|_2^2}$ $\Rightarrow$ penalizing L^2 -norm of derivatives up to order $\frac{d+n}{2}$

$\Rightarrow$ Sobolev space of order $\frac{d+n}{2}$

Dense in L^2 .

Beyond $\mathbb{R}^d$

ex: sequence

as as as as as ...

"spectrum kernel"

$q(x)$ sequences of length $h = \frac{1}{s}$ is a subsequence of n
 $h(x, s) = \text{easy to compute}$

ALGORITHM : goal : $\min_{\alpha \in H} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \phi(x_i), \alpha \rangle) + \frac{\lambda}{2} \|\alpha\|_2^2$

① we represent the mean

$$\min_{\alpha} \frac{1}{n} \sum \ell(y_i, \langle \phi(x_i), \alpha \rangle) + \frac{\lambda}{2} \alpha^T K \alpha$$

ex: least-squares $\left(\frac{1}{2n} \|y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^T K \alpha \right)$

$$\nabla_{\alpha} = \frac{1}{n} K(K\alpha - y) + \lambda K\alpha = 0$$

$$\Leftrightarrow (K^2 + n\lambda K)\alpha = Ky$$

$$K((K + n\lambda I)\alpha - y) = 0$$

one solution $\rightarrow \alpha = (K + n\lambda I)^{-1} y$

α unique if K is invertible

(never happens)

Hessian
is $K^2 + \lambda K$

is not invertible.
 $\Rightarrow$ GD is slow

in practice: $\rightarrow$ convex dual
 $\rightarrow$ compute a square root of K

$K = K = \Phi \Phi^T$
(cholesky dec.) see 7.4.2

$$\frac{1}{2n} \|y - \Phi\alpha\|_2^2 + \frac{\lambda}{2} \|\alpha\|_2^2$$

$$(\Phi^T \Phi + n\lambda I)\alpha = \Phi^T y$$

Unique sol

Efficient kernel methods: ① column sampling] see book
 ② random feature)
 ③ sgd

SGD: single pass version

min $\mathbb{E}[l(y, \langle \theta, \phi(x) \rangle)]$ $(x_i, y_i) \text{ iid } \in \mathcal{X} \times \mathcal{Y}$
 $\theta \in \mathcal{H}$

(A) $\theta_t = \theta_{t-1} - \gamma l'(y_t, \langle \theta_{t-1}, \phi(x_t) \rangle) \phi(x_t)$ in feature space.

$\theta_0 = 0$

θ_t is a linear combination of $\phi(x_1), \dots, \phi(x_t)$

$\theta_t = \sum_{i=1}^t \alpha_i \phi(x_i)$ because of

$\Rightarrow$ can replace (A) by $\theta_t = -\gamma l'(y_t, \underbrace{\sum_{i=1}^{t-1} \alpha_i \langle \phi(x_i), \phi(x_t) \rangle}_{\sum_{i=1}^{t-1} \alpha_i k(x_i, x_t)})$

complexity is $\sum_{t=1}^n t = \underline{\underline{\frac{n^2}{2}}}$

Random features

Gaussian kernel
is an expectation

$$e^{-\|x-x'\|_2^2/2} = \int e^{-\frac{\|u\|^2}{2}} e^{iu \cdot x} e^{-iu \cdot x'} du$$
$$= \mathbb{E}_{u \sim \text{Gaussian mean } 0, \text{ cov } I} e^{iu \cdot x} e^{-iu \cdot x'}$$

Re. him, Recht, Vost

$$\approx \frac{1}{m} \sum_{j=1}^m e^{i \omega_j \cdot x} e^{-i \omega_j \cdot x'}$$

where $\omega_1, \dots, \omega_m$
iid gaussian

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

$$\text{with } \varphi(x) = \frac{1}{\sqrt{m}} e^{i \omega_j \cdot x}$$

$$\varphi(x) \in \mathbb{R}^m$$

$$\text{trick } \varphi(x)_j = \sqrt{2} \cos(\omega_j \cdot x + \delta_j)$$
$$\delta_j \text{ uniform } (0, 2\pi)$$

$\Rightarrow$ see book for
Neural networks

7.4.3

Theoretical analysis of kernel method

study $\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle)$ s.t. $\|\theta\|_2 \leq D$.

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

as a simplification of $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|_2^2$

$R(\theta)$ expected risk

$$R(\hat{\theta}) - R_{\star} = \underbrace{R(\hat{\theta}) - \inf_{\|\theta\|_2 \leq D} R(\theta)}_{\text{estimation error}} + \underbrace{\inf_{\|\theta\|_2 \leq D} R(\theta) - R_{\star}}_{\text{approx. error}}$$

$\uparrow$
Bayes Risk

$\underbrace{\hspace{10em}}_{\text{express expected risk}}$

$$\leq 4 \frac{G R D}{\sqrt{n}}$$

$$\leq G \inf_{\|\theta\|_2 \leq D} \|\hat{\theta} - \theta\|_{L(\varphi)}$$

$$\leq 4 \frac{G R D}{\sqrt{n}} + G \inf_{\|\theta\|_2 \leq D} \|\hat{\theta} - \theta\|_{L(\varphi)}$$

① if ℓ is G -Lipschitz.

② $\|\varphi(x)\|_2 \leq R$

$\triangle$ for kernel methods, for all θ , $\inf_{\|\theta\|_2 \leq D} \|\hat{\theta} - \theta\| \rightarrow 0$ as $D \rightarrow \infty$ for separable spaces

unsmooth kernel : ① if f has s derivatives in L^2
 then the rate of convergence of the approx is
 is $O\left(\frac{1}{n^{s/d}}\right)$ $\neq$ from kernel

② "identity"

Proof for $s = \frac{d}{2}$

f has a finite norm for the
 function space defined by $e^{-\|x-x'\|_2}$
 (exponential kernel)
 $f^*(x) = \langle \phi_x, \phi(x) \rangle$ for $\|\phi_x\|$ finite kernel

if $D \geq \|\phi_x\| = 1$ approx error $\Rightarrow$

$$\text{rate} = \frac{GRD}{\sqrt{n}} \quad \left(\frac{1}{n^{s/d}} \sim \frac{1}{n^{1/2}} \right)$$