Recap

- → Kernel methods as applied to algorithms
- → Functional analysis

Kernel trick

Last time: "kernelizing" soft SVM

ERM (linear regression)

Ridge regression

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{m} \left(y_i - \langle w, x_i \rangle \right)^2 + \lambda \|w\|^2$$

y_i: o/p $x_i \in \mathbb{R}^d$

Recall

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}, \quad X \in \mathbb{R}^{m \times d}$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$w^* = (X^T X + \lambda I)^{-1} X^T Y$$

Features: functions from $\mathcal{X}$ (i/p domain) to $\mathcal{H}$

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

$x_i = [x_{i1}, \dots, x_{id}]$ (coordinate functions)

↓ replace

$$\Phi(x_i) = [\Phi_1(x_i), \dots, \Phi_d(x_i)]$$

More general ridge regression:

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{m} \left(y_i - \langle w, \Phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \quad \text{(ERM-}\mathcal{H}\text{)}$$

$\mathcal{H}$: infinite-dimensional Hilbert space

$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$

- → Expressive by adding features ✓ recasting ERM on feature space ($\mathcal{H}$)
  e.g. XOR function $x_{\cdot 1}, x_{\cdot 2}$
- → Kernel methods allow us to ✓ reduce ERM-$\mathcal{H}$ to a finite dimensional problem
- → Compute finite-dimensional solution w/o computing innerproducts on $\mathcal{H}$

Kernels

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ e.g. Gaussian, poly

k is "positive definite" PDS, Mercer kernels

⇓ defines

a unique $\mathcal{H}$

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \quad \text{(Defines } \mathcal{H}\text{)}$$

How does this simplify ERM-$\mathcal{H}$?

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{m} \left(y_i - \langle w, \Phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2$$

$$(X^T X + \lambda I)^{-1} \underset{d \times m}{X^T} \underset{m \times 1}{Y}$$

$X : m \times d$

$$= X^T (X X^T + \lambda I)^{-1} Y$$

(Mohri)

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \quad X^T = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix}$$

$$(X X^T)[i,j] = \langle x_i, x_j \rangle \xrightarrow{\text{"kernel trick"}} \underset{m \times m}{K}[i,j] = k(x_i, x_j)$$

(m: # trainpts)

$$K[i,j] = k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}$$

(Gram matrix)

Solution to ERM-$\mathcal{H}$

$$w = \sum_{i=1}^{m} \Phi(x_i) \left((K + \lambda I)^{-1} Y \right)_i$$

Suppose we know that the minimizer w can be written

$$w = \sum_{i=1}^{m} \Phi(x_i) a_i$$

$$w^* = \operatorname*{arg\,min}_{w \in \mathcal{H}} \left(\sum_{i=1}^{m} \left(y_i - \langle w, \Phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \right)$$

$$= \operatorname*{arg\,min}_{a \in \mathbb{R}^m} \left(\sum_{i=1}^{m} \left(y_i - \langle \Phi(x_i), \sum_{j=1}^{m} a_j \Phi(x_j) \rangle_{\mathcal{H}} \right)^2 + \lambda \left\| \sum_{i=1}^{m} a_j \Phi(x_j) \right\|^2 \right)$$

$$a^* = (K + \lambda I)^{-1} Y$$

$$w^* = \sum_{i=1}^{m} \Phi(x_i) \left((K + \lambda I)^{-1} Y \right)_i$$

$$K \in \mathbb{R}^{m \times m} \quad K[i,j] = k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}$$

Intro to Hilbert space

Complete Inner product space

→ Innerproduct : a function $\mathcal{F} \times \mathcal{F} \to \mathbb{C}$ $\langle \cdot, \cdot \rangle$
that satisfies

(i) Positive definiteness
$\langle f, f \rangle > 0 \quad \forall f \in \mathcal{F}$, $f \neq 0$

(ii) Conjugate Symmetric
$\langle f, g \rangle = \overline{\langle g, f \rangle}$

(iii) Linear
- $\langle f_1 + f_2, g \rangle = \langle f_1, g \rangle + \langle f_2, g \rangle$
- $\langle af, g \rangle = a \langle f, g \rangle$

$(\because \langle f, ag \rangle = \bar{a} \langle f, g \rangle)$

Complete : A space in which all Cauchy sequences converge.

Cauchy: $\{f_n\} \in \mathcal{F}$ s.t. for every $\varepsilon > 0$
$\exists N \in \mathbb{N}$ s.t. $\|f_n - f_m\|_{\mathcal{F}} < \varepsilon$
$\forall n, m \geq N$.

e.g. $\mathbb{R}^d$, dotproduct

→ $L^2([0,1])$: Hilbert space

$\langle f, g \rangle = \int_0^1 f(x) \overline{g(x)} \, dx$

$\|f\|^2 = \langle f, f \rangle$
$= \int_0^1 f(x) \overline{f(x)} \, dx$
$= \int_0^1 |f(x)|^2 \, dx$

→ ℓ^2 : Infinite sequences on $\mathbb{C}$ with finite ℓ^2 norm

$f, g \in \ell^2$

$f = \{f_n\} \quad g = \{g_n\}$

$\langle f, g \rangle = \sum_{n \in \mathbb{Z}^+} f_n \overline{g_n}$

$\|f\|^2 = \sum_{n \in \mathbb{Z}^+} |f_n|^2$

Finite dimensional $\ell^2(\mathbb{R}^d)$

Finite dimensional vector space $\mathbb{R}^d$

$$\underset{e_1}{\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}}, \underset{e_2}{\begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}}, \dots, \underset{e_d}{\begin{bmatrix}0\\\vdots\\1\end{bmatrix}}$$

$v \in \mathbb{R}^d$

$v = \sum_{i=1}^d v_i e_i$ ← components

Basis
→ Not unique
→ $\mathrm{span}(\{e_i\}_{i=1}^d) = \mathbb{R}^d$

$\mathrm{span}(\{e_i\}_{i=1}^\infty)$ is dense on $\mathcal{F}$
(countable)
if e_i forms a basis for $\mathcal{F}$.

Dense: $\sum_{i=1}^n a_i e_i$ get arbitrarily close to any element of $\mathcal{F}$

E.g.
$L^2([0,1])$: Hilbert

$e_k(x) = e^{2\pi i k x}$ ← imaginary

Any $f \in L^2([0,1]) \to \sum_{k \in \mathbb{Z}} f_k e_k$

Orthonormal basis
e.g. Canonical basis in $\mathbb{R}^d$
$\langle e_i, e_j \rangle = 0 \quad i \neq j$
$\langle e_i, e_i \rangle = \|e_i\|^2 = 1$

$f = \sum_{k \in \mathbb{Z}} f_k e_k = \sum_{k \in \mathbb{Z}} \langle f, e_k \rangle e_k$

$\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \quad \begin{bmatrix}a\\b\end{bmatrix} = \begin{bmatrix}a\\b\end{bmatrix} \cdot \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}a\\b\end{bmatrix} \cdot \begin{bmatrix}0\\1\end{bmatrix}$
$= \begin{bmatrix}a\\b\end{bmatrix}$

$f(x) = \sum_{k \in \mathbb{Z}} \langle f, e^{2\pi i k \cdot} \rangle e^{2\pi i k x}$
Fourier series expansion

Hilbert projection theorem ($\mathcal{F}$: Hilbert)
if C is a closed subset of $\mathcal{F}$,
$C \oplus C^\perp = \mathcal{F}$

$C^\perp$: orthogonal complement of C

$f \in C$, $g \in C^\perp$, $\langle f, g \rangle = 0$
(orthogonality)

$f \in \mathcal{F}$: $f = f_C + f_{C^\perp}$
where $f_C \in C$ and $f_{C^\perp} \in C^\perp$
$\langle f_C, f_{C^\perp} \rangle = 0$

$\mathcal{F}$ is an inner product space.
$\mathcal{F}^c$, $\mathcal{F}'$, $\mathcal{F}^*$: dual of $\mathcal{F}$
Space of linear "functionals" on $\mathcal{F}$

Functional : function from $\mathcal{F}$ to $\mathbb{C}$.

e.g. $f \in \mathcal{F}$
δ_x : delta functional $\in \mathcal{F}^c$

$\delta_x(f) = f(x)$ ←

$\frac{\partial}{dx}$

$= \int_0^1 f(x) \boxed{\delta_{x_0}(x)} \, dx = f(x_0)$

Side note: δ distribution is in L^1: $\int |\delta(x)| \, dx = 1$
$(L^1)^* \equiv L^\infty$ (space of bounded distributions)

Riesz representation theorem

if $\mathcal{F}$ is Hilbert, any linear functional $L \in \mathcal{F}^*$ is "represented" by a unique element $g_L \in \mathcal{F}$: $L(f) = \langle f, g_L \rangle$

$\mathcal{L}$ is a linear functional if
- $\mathcal{L}(f+g) = \mathcal{L}f + \mathcal{L}g$
- $\mathcal{L}(af) = a\mathcal{L}f$

e.g. of dual space
- $L^2([0,1])$ is dual to itself
- $L^p([0,1])$, $L^q([0,1])$ $\frac{1}{p} + \frac{1}{q} = 1$

RKHS

Mercer kernels:

$\mathcal{X} \in \mathbb{R}^d$ compact. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ continuous, symmetric function. Then,

K admits a uniformly convergent expansion

$$K(x, x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x')$$

with $a_n > 0$ iff for any $c \in L^2(\mathcal{X})$

$$\iint c(x) c(x') K(x, x') \, dx \, dx' \geq 0$$