A primer on kernel methods

[Disclaimer: extraction from Francis's book.]

- Study of empirical risk minimization for linear models

prediction function f_θ

$$f_\theta : \mathcal{X} \to \mathbb{R}$$
$$x \mapsto \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$$

where $\varphi : \mathcal{X} \to \mathcal{H}$ is usually called a feature map and $\mathcal{H}$ is a Hilbert space.

- Ex: Polynomial features. How to extend linear classifiers to non-linear.

- Kernel methods ≡ Infinite-dimensional linear methods

- Lead to simple and stable algorithms with theoretical guarantees and adaptivity to the smoothness of the target function

- Other motivation: can be helpful to understand other models such as neural networks or overparametrized models.

- Consider the optimization pb coming from ML:

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

dot product and norm taken w.r.t the Hilbertian structure

with training data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$

$\varphi : \mathcal{X} \to \mathcal{H}$ feature map

Prop: [Representer theorem]

Consider a feature map $\varphi : \mathcal{X} \to \mathcal{H}$

samples $(x_1, \dots, x_n) \in \mathcal{X}^n$

the functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ is strictly increasing w.r.t the last variable.

Then, the infimum of $\Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2)$ can be obtained by restricting to a vector θ in the span of $\varphi(x_1), \dots, \varphi(x_n)$, i.e., of the form

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

Proof: Let $\theta \in \mathcal{H}$, $\mathcal{H}_D = \{\sum_{i=1}^{n} \alpha_i \varphi(x_i) : \alpha \in \mathbb{R}^n\} \subset \mathcal{H}$ ← linear span of the observed feature vectors.

Let $\theta = \underbrace{\theta_D}_{\in \mathcal{H}_D} + \underbrace{\theta_\perp}_{\in \mathcal{H}_D^\perp}$ ← again w.r.t the Hilbertian structure.

Then $\forall i \in \{1, \dots, n\}$,

$$\langle \theta, \varphi(x_i) \rangle = \langle \theta_D, \varphi(x_i) \rangle + \underbrace{\langle \theta_\perp, \varphi(x_i) \rangle}_{0 \text{ by def}^{\text{n}} \text{ of the orthogonal.}}$$

We have

$$\Psi\big(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2\big) \overset{\text{by Pythagorean's thm.}}{=} \Psi\big(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta_D\|^2 + \|\theta_\perp\|^2\big)$$

$$\geq \Psi\big(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta_D\|^2\big)$$

↑ with equality iff $\theta_\perp = 0$ (since Ψ is strictly increasing w.r.t the last variable)

Thus,

$$\inf_{\theta \in \mathcal{H}} \Psi\big(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2\big) = \inf_{\theta \in \mathcal{H}_D} \Psi\big(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2\big)$$

■

Corollary: [Representer thm for supervised learning]

For $\lambda > 0$, the infimum of $\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2$ can be obtained by restricting to vector θ of the form

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

Remarks:

- There is no assumption about the loss function (no need for convexity)
- We can reformulate the learning problem using the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

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

Then if $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$, then for all $j \in \{1 \dots n\}$

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

where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix (= gram matrix of the feature vectors) such that $K_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$

Besides,

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

The learning problem can be rewritten as

$$\inf_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|^2 = \inf_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^T K \alpha$$

For any test point $x \in \mathcal{X}$, the prediction function is defined as

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

- Here the input observations are summarized in the kernel matrix and the kernel function regardless of the dimension of $\mathcal{H}$.

- No need to explicitly compute the feature vector $\varphi(x)$
  Need only of dot products
  This is the kernel trick.
  ↳ The search space $\mathcal{H}$ is reparameterized by $\mathbb{R}^n$.
  Can be computationally interesting when $\dim \mathcal{H}$ is large.

Remark: [Minimum norm interpolation]

The representer theorem can be extended to an interpolating estimator with essentially the same proof.

Prop Given $x_1, \dots, x_n \in \mathcal{X}$ and $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$,
such that $\exists \theta \in \mathcal{H}$ $y_i = \langle \theta, \varphi(x_i) \rangle$ for all $i = 1..n$.
Then among all the interpolators $\theta \in \mathcal{H}$, the one of minimum norm can be expressed as $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$
with $\alpha \in \mathbb{R}^n$ such that $Y = K\alpha$.

- Kernels :

By defining $k(x,x') := \langle \varphi(x), \varphi(x') \rangle$,
then the associated kernel matrix

$$K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$$

is a Gram matrix of feature vectors and is thus symmetric positive definite, i.e. $\forall \alpha \in \mathbb{R}^n \quad \alpha^T K \alpha \geq 0$

It turns out this simple property is enough to ensure the existence of a feature function.

Def: A function $k : X \times X \to \mathbb{R}$ is a positive_definite kernel if and only if all kernel matrices resulting from this kernel functions are symmetric positive semidefinite.

Prop: [Aronszajn 1950]
The function $k : X \times X \to \mathbb{R}$ is a positive_definite kernel iff $\exists$ a Hilbert space $\mathcal{H}$ and a function $\varphi : X \to \mathcal{H}$ such that $\forall x, x' \in X$ $\quad k(x,x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$.

Partial proof:

⇐ If $k(x,x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$, then for any $\alpha \in \mathbb{R}^n$ any points $x_1, \dots, x_n \in \mathcal{X}$, the kernel matrix K associated with these points satisfy

$$\alpha^T K \alpha = \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0$$

Thus k is a positive definite kernel.

⇒ We have to construct a space of functions explicitly from $\mathcal{X}$ to $\mathbb{R}$ with a dot product.

Define $\mathcal{H}' \subset \mathbb{R}^{\mathcal{X}}$

$$\mathcal{H}' := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \;:\; n \in \mathbb{N} \quad \alpha \in \mathbb{R}^n \quad x_1, \dots, x_n \in \mathcal{X} \right\}$$

set of linear combinations of kernel functions.

This is a vector space on which we can define a dot product through

(DP) $$\left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \sum_{j=1}^{m} \beta_j k(\cdot, x'_j) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j)$$

This is a well-defined function on $\mathcal{H}' \times \mathcal{H}'$ (i.e. its value does not depend on the chosen representation as a linear combination of kernel functions).

Indeed if we denote $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ then the dot product (DP) is equal to $\sum_{j=1}^{m} f(x'_j)$, depending on the values of f not on its representation.

(DP) is bilinear and always pos. negative when applied to the same function (i.e. take $\alpha = \beta$ $\{x_i, i\} = \{x_j', j\}$) since k is positive-definite.

(DP) satisfies the properties

$$\forall f \in \mathcal{H}' \quad x, x' \in \mathcal{X} \quad \begin{cases} \langle k(\cdot, x), f \rangle = f(x) \\ \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x') \end{cases}$$

These are called reproducing properties and correspond to an explicit construction of the feature map $\varphi(x) = k(\cdot, x)$.

$\boxed{\langle f, f \rangle = 0 \overset{?}{\Rightarrow} f = 0}$ Yes by Cauchy-Schwarz ineq.,

$$\forall x \in \mathcal{X} \quad f(x)^2 = \langle f, k(\cdot, x) \rangle^2 \leq \langle f, f \rangle \langle k(\cdot, x), k(\cdot, x) \rangle$$
$$= \langle f, f \rangle \, k(x, x)$$

leading to $f = 0$ as soon as $\langle f, f \rangle = 0$.

The space $\mathcal{H}'$ is called "pre-Hilbertian" because it's not complete. It can be completed into a Hilbert space $\mathcal{H}$ with the same reproducing property ▨

$\mathcal{H}$ is called the feature space

φ ——— the feature map $\varphi: \mathcal{X} \to \mathcal{H}$.

No assumption is needed about the input space $\mathcal{X}$

No regularity assumption is needed for k.

Up to isomorphisms, the feature map and space happen to be unique.

For any positive-definite kernel k, the particular space of functions that we built is called the reproducing kernel Hilbert space (RKHS) associated with k, for which $\varphi(x) = k(\cdot, x)$.

Kernels = features and functions. A positive definite kernel thus defines a feature map and a space of functions. Sometimes the feature map is easy to find, and other times, it is not (but hopefully you won't need it!)

Examples of kernels

- **Linear kernel** $\quad k(x,x') = x^T x'$

  This kernel corresponds to a function space composed of linear functions $f_\theta(x) = \theta^T x$ with an ℓ^2 penalty $\|\theta\|_2^2$.
  The kernel trick can be useful when the input data have huge dim. d but are quite sparse so that the dot product $x^T x'$ can be computed in time $o(d)$.

- **Polynomial kernel** : for $s \in \mathbb{N}^*$, the kernel $k(x,x') = (x^T x')^s$ is positive definite (as an integer power of a kernel)

$$k(x,x') = \left(\sum_{j=1}^d x_j x_j'\right)^s = \sum_{\alpha_1+\dots+\alpha_d = s} \binom{s}{\alpha_1,\dots,\alpha_d} (x_1 x_1')^{\alpha_1} \dots (x_d x_d')^{\alpha_d}$$

$\alpha_1+\dots+\alpha_d = s$: nonnegative integers

$(x_1^{\alpha_1} \dots x_d^{\alpha_d})((x_1')^{\alpha_1} \dots (x_d')^{\alpha_d})$

We can deduce an explicit feature map from this expression

$$\varphi(x) = \left(\binom{s}{\alpha_1,\dots,\alpha_d}^{1/2} x_1^{\alpha_1} \dots x_d^{\alpha_d} \right)_{\alpha_1+\dots+\alpha_d = s}$$

and the set of functions is the set of degree-s homogeneous polynomials on $\mathbb{R}^d$, which has a dimension $\binom{d+s-1}{s}$ (*)

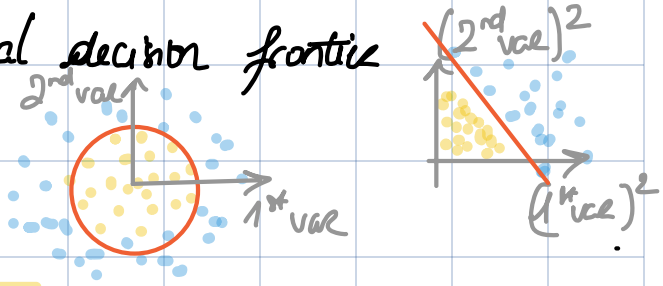
(*) f homogeneous of degree $s \in \mathbb{R}_+$: $\forall x \in \mathbb{R}^d, \forall \lambda \in \mathbb{R}_+$, $f(\lambda x) = \lambda^s f(x)$.

When d and s grow, the feature space dimension grows as d^s, and an explicit representation is not desirable. The kernel trick can be advantageous

The associated norm (which penalizes coefficients of the polynomials) is hard to interpret though (as a small change in a single high-order coefficient can lead to significant changes

Often we consider $k(x,x') = (1 + x^T x)^s$ which corresponds to the set of all monomials $x_1^{\alpha_1} \dots x_d^{\alpha_d}$ such that $\alpha_1 + \dots + \alpha_d \leq s$. The dimension of the feature space is still $\binom{d+s}{s}$.

Illustration : when using a polynomial kernel of order 2 the set of functions linear in the feature map is the set of quadratic functions leading to ellipsoidal decision frontier



- Translation-invariant kernels on $[0,1]$

$\mathcal{X} = [0,1]$

Take kernel of the form $k(x,x') = q(x-x')$

assumed to be 1-periodic

They emerge from penalties on the Fourier coefficients of fct

Toolkit on Fourier series :

Any squared integrable function on $[0,1]$, that is 1-periodic, can be expanded in the orthonormal basis $\{x \mapsto e^{2i\pi mx}, m \in \mathbb{Z}\}$ of $L^2([0,1])$.

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

with $\hat{q}_m = \int_0^1 q(x) e^{-2i\pi mx} dx \in \mathbb{C}$

complex conjugate

the function q is real-valued iff $\forall m \in \mathbb{Z}$ $\hat{q}_{-m} = \hat{q}_m^*$.

Plancherel's identity ($\equiv$ Pythagorean thm)

$$\int_0^1 |q(x)|^2 dx = \sum_{m \in \mathbb{Z}} |\hat{q}_m|^2$$

Given a function $f \in L^2([0,1])$ decomposed into its Fourier series as

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

one can consider the weighted norm

$$\|f\|_c^2 = \sum_{m \in \mathbb{Z}} c_m |\hat{f}_m|^2 \quad \text{with } (c_m)_m \in \mathbb{R}_+^{\mathbb{Z}}$$

This penalty can be interpreted through a feature map and a particular dot product.

Consider the Hilbert space $\ell^2(\mathbb{Z})$ of square summable series with the dot product $\langle a, b \rangle = \sum_{m \in \mathbb{Z}} a_m b_m^*$

Take the feature vector $\varphi(x)_m = e^{-2i\pi m x}/\sqrt{c_m}$

$\theta \in \ell^2(\mathbb{Z})$ such that $\theta_m = \hat{f}_m \sqrt{c_m}$

then $f(x) = \langle \theta, \varphi(x) \rangle$ is such that

$$\|f\|_c^2 = \sum_{m\in\mathbb{Z}} c_m^2 |\hat{f}_m|^2 = \|\theta\|^2_{\ell^2(\mathbb{Z})} = \sum_{m\in\mathbb{Z}} |\theta_m|^2$$

The associated kernel is

$$k(x,x') = \sum_{m\in\mathbb{Z}} \varphi(x)_m \, \varphi(x')_m^* = \sum_{m\in\mathbb{Z}} \frac{e^{2i\pi m x}}{\sqrt{c_m}} \cdot \frac{e^{-2i\pi m x'}}{\sqrt{c_m}}$$

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

which takes the form of $q(x-x')$ for a 1-periodic function q with Fourier series $\hat{q}_m = 1/c_m$.

Cl: Any penalty of the form $\sum c_m |\hat{f}_m|^2$ defines a squared RKHS norm as soon as $\begin{cases} c_m \text{ is strictly positive} \\ \sum_{m\in\mathbb{Z}} 1/c_m < +\infty \end{cases}$

the kernel fct $k(x,x')$ takes the form $q(x-x')$ with q 1-periodic of non-negative Fourier series $\hat{q}_m = c_m^{-1}$.

$\Leftarrow$ Such kernel are positive definite (no formal proof)

ex: $\cos(x-y) = \cos x \cos y + \sin x \sin y = \langle \varphi(x), \varphi(y) \rangle$ with $\varphi(x) = \begin{pmatrix} \cos x \\ \sin x \end{pmatrix}$

Take n points $x_1 .. x_n \in \mathbb{R}$, form the Gram matrix $K_{ij} = \cos(x_i - x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$

Remark that $K = \varphi(X)\varphi(X)^T$ sym. positive semi-def. leading to a PSD kernel.

with $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\varphi(X) = \begin{pmatrix} \cos x_1 & \sin x_1 \\ \vdots & \vdots \\ \cos x_n & \sin x_n \end{pmatrix}$. $\forall \alpha \; \alpha^T K \alpha = \|\varphi(X)^T \alpha\|^2 \geq 0$.

Generalization on derivatives : if the sequence $(c_m)_{m\in\mathbb{Z}}$ takes the form of a power law, there is a link with penalties on derivatives

If f is s times diff with a square-integrable derivative

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

From Parseval's thm, we get

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

Ex: Fix $c_0 = 1$ and $c_m = m^{2s}$ for $m \neq 0$

the associated norm is $\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^{2s}} \int_0^1 |f^{(s)}(x)|^2 dx + \left(\int_0^1 f(x) dx\right)^2$

the corresponding kernel is

$$k(x,x') = \sum_{m\in\mathbb{Z}} c_m^{-1} e^{2i\pi m(x-x')}$$

$$= 1 + \sum_{m\geq 1} \frac{2\cos(2\pi m(x-x'))}{m^{2s}}$$

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

CLOSED FORM.

For $s \geq 1$, we can show that $k(x,x') = 1 + (-1)^{s-1} \frac{(2\pi)^{2s}}{(2s)!} B_{2s}(\{x-x'\})$

$(2s)$th Bernoulli polynomial, e.g. $B_{2s}(t) = t^2 - t + 1/6$

fractional part $\{t\} = t - \lfloor t \rfloor$

Penalizing the derivatives leads to Sobolev spaces in more generality.

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

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

$k(x, x') = q(x - x')$ translation invariant by addition of the same constant to both arguments.

Toolkit on Fourier transform : The Fourier transform $\hat{f} : \mathbb{R}^d \to \mathbb{C}$ of an integrable function f is defined by

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

It can naturally be extended to an operator on all square integrable fcts and under appropriate conditions on f, we can recover f from its Fourier transform

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega)\, e^{-i\omega^T x}\, d\omega$$

Parseval's identity $\int_{\mathbb{R}^d} |f(x)|^2 dx = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 d\omega$

Prop : A translation-invariant kernel k defined as $k(x, x') = q(x - x')$ is positive-definite iff q is the inverse of the Fourier transform of a nonnegative Borel measure.

Proof : just insight $q(x - x') = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i(x - x')^T \omega}\, d\mu(\omega)$.

μ : nonnegative measure

Take $x_1, \dots, x_n \in \mathbb{R}^d$ $\quad \alpha_1 \dots \alpha_n \in \mathbb{R}$

$$\sum_{\substack{j=1\\ \ell=1}}^{n} \alpha_j \alpha_\ell \, k(x_j, x_\ell) = \sum_{j,\ell=1}^{n} \alpha_j \alpha_\ell \, q(x_j - x_\ell)$$

$$= \frac{1}{(2\pi)^d} \sum_{j,\ell=1}^{n} \alpha_j \alpha_\ell \int e^{i\omega^T (x_j - x_\ell)} \, d\mu(\omega)$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\sum_{j,\ell=1}^{n} \alpha_j \alpha_\ell \, e^{i\omega^T x_j} \left(e^{i\omega^T x_\ell} \right)^* \right) d\mu(\omega)$$

$$= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^{n} \alpha_j e^{i\omega^T x_j} \right|^2 d\mu(\omega)$$

$$\geq 0.$$

POSITIVE - DEFINITENESS.

In practice when q and $\hat{q}$ are both integrable, the associated kernel is positive definite iff $\forall \omega \in \mathbb{R}^d$ $\quad \hat{q}(\omega) \geq 0.$

Matern kernel & Sobolev spaces : one can define a series of kernels such that $\hat{q}(\omega)$ is proportional to $\pi^d \left(1 + \pi^2 \|\omega\|_2^2\right)^{-s}$ (the Fourier transform is integrable) * for $s > d/2$ *

analog to the power series

constant homogeneous to input x, e.g. $r \equiv$ quantile of all the pairwise differences $\|x_i - x_j\|$ of the training data

These so-called "Matern kernels" all correspond to Sobolev spaces of order s and can be computed in closed form.

Reminder: $f \in H^s(\Omega) \iff \sum_{|\alpha| \leq s} \|\partial^\alpha f\|_2^2 < +\infty$

A key fact is that to be a RKHS, a Sobolev space has to have many derivatives when d grows:

↳ having only 1st order derivative ($s=1$) leads to an RKHS only for $d=1$

↳ for $s=(d+1)/2$, we obtain the exponential kernel $k(x,x') = \exp\left(-\frac{\|x-x'\|_2}{r}\right)$

↳ for $s=(d+3)/2$, $k(x,x') \propto \left(1+\sqrt{3}\frac{\|x-x'\|_2}{r}\right)\exp\left(-\sqrt{3}\frac{\|x-x'\|_2}{r}\right)$

↳ for $s=(d+5)/2$, $k(x,x') \propto \left(1+\sqrt{5}\frac{\|x-x'\|_2}{r}+\frac{5}{3}\frac{\|x-x'\|^2}{r^2}\right)\exp\left(-\sqrt{5}\frac{\|x-x'\|_2}{r}\right)$

More generally, $H^s(\Omega)$ with $\Omega \subseteq \mathbb{R}^d$ are RKHS when $s > d/2$ due to a key property: continuous point evaluation.

A Hilbert space $\mathcal{H}$ of function is a RKHS, if for all x, the evaluation functional $f \mapsto f(x)$ is continuous (i.e. bounded) on $\mathcal{H}$. This implies that for every x, there exits a reproducing kernel $K_x \in \mathcal{H}$ s.t. $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$.
(by the Riesz representation thm)

If $s > d/2$ $\quad H^s(\Omega) \subset C^0(\bar{\Omega})$ (the inclusion is continuous)

(Sobolev embedding thm) $H^s(\mathbb{R}^d) \hookrightarrow C_0^0(\mathbb{R}^d)$ i.e., fcts of $H^s(\mathbb{R}^d)$ are bounded and continuous

$$|f(x)| \leq C_x \|f\|_{H^s} \quad \forall f \in H^s.$$

When $s \leq d/2$ functions in H^s may no longer be continuous (as distribut° or just L^2)

See Berlinet & Thomas-Agnan's book for a bestiary of kernels / feature maps and RKHS.

Algorithms

- Aim : solving $\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$

- Hyp: ℓ is assumed to be convex in its 2^{nd} variable.

  bounded features, for all $i=1..n$, $k(x_i, x_i) = \|\varphi(x_i)\|_{\mathcal{H}}^2 \leq R^2$

- Using the representer theorem, one can try to solve

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

This is a convex pb.

In the particular case of the square loss (ridge regression)

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|Y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^T K \alpha$$

Optimality condition $(K^2 + n\lambda K)\alpha = KY$

sol: $\alpha = (K + n\lambda I)^{-1} Y$

It can be ill-conditioned pb because K has often small eigenvalues.

A possibility is to compute a square-root of $K = \Phi\Phi^T$ with $\Phi \in \mathbb{R}^{n \times m}$ and m is the rank of K, and to solve

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (\Phi\beta)_i) + \frac{\lambda}{2} \|\beta\|_2^2$$

Optimality condition $\frac{1}{n} \Phi^T g + \lambda \beta$ with $g \in \mathbb{R}^n$ the vector of gradients $g_i = \ell'(y_i, (\Phi\beta)_i)$ for all $i \in \{1..n\}$

derivative w.r.t the 2nd variable.

Obtain $\alpha \in \mathbb{R}^n$ $\alpha = -\frac{1}{\lambda n} g$ so that $\beta = \Phi^T \alpha$.

- **Column subsampling**: to approximate K, approximate square roots are a very useful tool, approximate K from a subset of its columns can be done as $K \approx K(:, I)\, K(I, I)^{-1} K(I, :)$ (all the n columns; all the n rows)

where $K(A, B) \equiv$ submatrix of K obtained by taking rows indexed by $A \subseteq \{1..n\}$, columns $\leftrightarrow$ $B \subseteq \{1..n\}$

this corresponds to an approximate square root $\tilde{\Phi} = K(:, I)\, K(I, I)^{-1/2} \in \mathbb{R}^{n \times m}$ with $m = |I|$, which can be computed in time $O(m^2 n)$ instead of $O(n^3)$

Referred to as Nyström approximation in linear algebra.

- **Random features**: for kernel of the form (positive def. on some space $\mathcal{V}$)

$$k(x, x') = \int_{\mathcal{V}} \varphi(x, v)\, \varphi(x', v)\, d\tau(v) = \langle \varphi(x, \cdot), \varphi(x', \cdot) \rangle_{L^2(\tau)}$$

We can approximate the expectation by an empirical average

$$\hat{k}(x, x') = \frac{1}{m} \sum_{\ell=1}^{m} \varphi(x, v_\ell)\, \varphi(x', v_\ell)$$

where the (V_ℓ)'s are sampled i.i.d. from τ

We can use an explicit feature representation $\hat{\varphi}(x) = \left(\frac{1}{\sqrt{m}} \varphi(x, V_\ell)\right)_{\ell=1..m}$

and solve

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \hat{\varphi}(x_i)^T \beta) + \frac{\lambda}{2} \|\beta\|_2^2$$

with a predictor $x \mapsto \beta^T \hat{\varphi}(x)$

OK when the number m of random features is $\ll n$.

Dimension reduction $\neq$ of the input data : the random feature function $\varphi(\cdot, V_\ell)$ are selected before the data are observed.

($\neq$ column subsampling)

ex: Translation invariant kernels $q(x-x') = \frac{1}{(2\pi)^d} \int \hat{q}(\omega) e^{i\omega^T(x-x')} d\omega$

for which we can take

↳ complex valued features $\varphi(x, \omega) = \sqrt{q(0)}\, e^{i\omega^T x} \in \mathbb{C}$ with ω sampled from the distribution of density $\frac{1}{(2\pi)^d} \frac{\hat{q}(\omega)}{q(0)}$

↳ real-valued features $\sqrt{2}\cos(\omega^T x + b)$

(?) ω ? $b \sim \mathcal{U}([0, 2\pi])$

see [Rahimi & Recht 2008].

ex: NN with random weights $\varphi(x, v) = \sigma(v^T x)$ (e.g. ReLU)

When $v \sim \mathcal{U}(\mathbb{S}^{d-1})$, $k(x, x') = \frac{\|x\|_2 \|x'\|_2}{2(d+1)\pi} \left[(\pi - \eta)\cos\eta + \sin\eta\right]$

where $\cos\eta = x^T x' / \|x\|_2 \|x'\|_2$

[Le Roux & Bengio 2007]

Theoretical analysis of ridge regression

- n obs. $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ $\qquad X_i \overset{i.i.d}{\sim} P$ ← density of the distribution of the inputs

- Model assumption: $\quad Y_i = f^*(X_i) + \varepsilon_i$

$$\begin{cases} \mathbb{E}[\varepsilon_i | X_i] = 0 \\ \mathbb{E}[\varepsilon_i^2 | X_i] \leq \sigma^2 \end{cases} \longrightarrow f^*(x) = \mathbb{E}[Y | X = x].$$

Hyp: $f^* \in L^2(P)$

- Estimator via the minimization

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}$$

- Using the representer thm, we know that the estimator is given by

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x, X_i) \qquad \text{with} \quad \alpha = (K + n\lambda I)^{-1} y \in \mathbb{R}^n$$

$$= \sum_{i=1}^{n} \hat{w}_i(x) Y_i \qquad \text{with} \quad \hat{w}(x) = (K + n\lambda I)^{-1} q(x) \in \mathbb{R}^n$$

where $q(x) = (q_i(x))_{i=1..n}$ $\quad q_i(x) = k(x, X_i)$

This looks like a local averaging but

(i) the weights do not sum to 1

(ii) — " — are not constrained to be nonnegative

- let $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \otimes \varphi(x_i)$

* for $a, b \in \mathcal{H}$, $a \otimes b$ is an operator from $\mathcal{H}$ to $\mathcal{H}$ such that for $f \in \mathcal{H}$ $\quad (a \otimes b) f = \langle b, f \rangle_{\mathcal{H}} \, a$

- then, for any f,

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

For the theoretical analysis, we do not rely on the representation of $\hat{f}_n$ with α but directly on this formulation.

$$= \frac{1}{n} \sum_{i=1}^{n} y_i^2 + \langle f, \hat{\Sigma} f \rangle - 2 \left\langle \frac{1}{n} \sum_{i=1}^{n} y_i \varphi(x_i), f \right\rangle + \lambda \langle f, f \rangle$$

$$\langle f, \left(\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \otimes \varphi(x_i) \right) f \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle \varphi(x_i), f \rangle \langle f, \varphi(x_i) \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle k(x_i, \cdot), f \rangle \langle f, k(x_i, \cdot) \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} (f(x_i))^2$$

Similar computations.

One can deduce the minimizer $\hat{f}_\lambda$

$$\hat{f}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^{n} y_i \varphi(x_i)$$

$$= (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^{n} f^*(x_i) \varphi(x_i) + (\hat{\Sigma} + \lambda I)^{-1} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \varphi(x_i)$$

- Expected excess risk / Estimation error

$$\mathbb{E}\left[\| \hat{f}_\lambda - f^* \|^2_{L^2(\rho)} \right]$$

$$= \mathbb{E}\left[\left\| (\hat{\Sigma}+\lambda I)^{-1} \frac{1}{n} \sum_{i=1}^n \xi_i \varphi(x_i) \right\|^2_{L^2(\rho)} \right] + \mathbb{E}\left[\left\| (\hat{\Sigma}+\lambda I)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n f^*(x_i)\varphi(x_i) - f^* \right\|^2_{L^2(\rho)} \right]$$

Variance term — depends on the noise

Bias term — depends on the regularity of the target function

- Introduce the covariance operator $\Sigma := \mathbb{E}\left[\varphi(x) \otimes \varphi(x) \right]$

Note that $\mathbb{E}[\hat{\Sigma}] = \Sigma$

This is an operator from $\mathcal{H}$ to $\mathcal{H}$.

Key property: it relates $\|.\|_{L^2(\rho)}$ to $\|.\|_{\mathcal{H}}$ as

$$\|g\|^2_{L^2(\rho)} = \int_X g^2(x)\, d\rho(x) = \int_X \langle g, \varphi(x) \rangle^2 d\rho(x)$$

$$= \int_X \langle g, \varphi(x) \otimes \varphi(x)\, g \rangle\, d\rho(x)$$

$$= \langle g, \Sigma g \rangle$$

$$= \| \Sigma^{1/2} g \|_{\mathcal{H}}.$$

More generally, we have $\int_X f(x) g(x)\, d\rho(x) = \langle f, \Sigma g \rangle_{\mathcal{H}}$.

for all $f, g \in \mathcal{H}$.

- Variance term

$$\mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\varphi(x_i)\right\|^2_{L^2(\rho)}\right]$$

$$= \mathbb{E}\left[\left\|\Sigma^{\frac{1}{2}}(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\varphi(x_i)\right\|^2_{\mathcal{H}}\right]$$

$$= \mathbb{E}\left[\left\langle\Sigma^{\frac{1}{2}}(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\varphi(x_i), \Sigma^{\frac{1}{2}}(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\varphi(x_i)\right\rangle_{\mathcal{H}}\right]$$

$$= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\langle(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\xi_i\varphi(x_i), \xi_i\varphi(x_i)\right\rangle_{\mathcal{H}}\right]$$

$$= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\xi_i^2\,\varphi(x_i)\otimes\varphi(x_i)\right]\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\xi_i^2\,\varphi(x_i)\otimes\varphi(x_i)\right)\right]$$

$$\leq \frac{\sigma^2}{n}\,\mathbb{E}\left[\operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma\underbrace{(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}}_{\preceq I}\right)\right] \qquad \text{since } \mathbb{E}[\xi_i^2|x_i]\leq\sigma^2$$

$$\leq \frac{\sigma^2}{n}\,\mathbb{E}\left[\operatorname{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma\right)\right]$$

using $(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\preceq I$

and that for symmetric matrices

when $A\succeq 0$ $\quad B\preceq C$

then $\operatorname{Tr}(AB)\leq\operatorname{Tr}(AC)$.

• Bias term — we assume that $f^* \in \mathcal{H}$ (well-specified model)

$$\mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\cdot\frac{1}{n}\sum_{i=1}^{n} f^*(x_i)\varphi(x_i) - f^*\right\|^2_{L^2(\rho)}\right] =$$

potential problem here for misspecified model.

$$= \mathbb{E}\left[\left\| (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^{n} \langle f^*, \varphi(x_i)\rangle \varphi(x_i) - f^* \right\|^2_{L^2(\rho)}\right]$$

$$= \mathbb{E}\left[\left\| (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} f^* - f^* \right\|^2_{L^2(\rho)}\right]$$

$$\left\| \left[(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} - I\right) f^* \right\|^2_{L^2(\rho)}$$

$$\left\| (\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma} - \hat{\Sigma} - \lambda I) f^* \right\|^2_{L^2(\rho)}$$

$$\left\| \lambda (\hat{\Sigma}+\lambda I)^{-1} f^* \right\|^2_{L^2(\rho)}$$

$$= \mathbb{E}\left[\left\| \lambda \Sigma^{1/2} (\hat{\Sigma}+\lambda I)^{-1} f^* \right\|^2_{\mathcal{H}}\right]$$

$$= \lambda^2 \mathbb{E}\left[\left\langle f^*, (\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1} f^* \right\rangle\right]$$

Prop: When $f^* \in \mathcal{H}$, the excess risk of the ridge kernel regression estimator is upper bounded by

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f^*\right\|^2_{L^2(\rho)}\right] \leq \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\Sigma\right]\right] + \lambda^2 \mathbb{E}\left[\left\langle f^*, (\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1} f^* \right\rangle\right]$$

empirical and population covariance operator.

• Relating Empirical and Population Covariance Operators

Lemma: [Mourtada & Rosasco]

Assume iid data $x_1, \dots, x_n \in \mathcal{X}$.

bounded feature $\|\varphi(x)\|_{\mathcal{H}} \leq R$ for all $x \in \mathcal{X}$.

$$\mathbb{E}\left[\mathrm{Tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\Sigma\right]\right] \leq \left(1+\frac{R^2}{\lambda n}\right)\mathrm{Tr}\left[(\Sigma+\lambda I)^{-1}\Sigma\right]$$

$$\mathbb{E}\left[\left\langle g, (\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1} g\right\rangle\right] \leq \frac{1}{\lambda}\left(1+\frac{R^2}{\lambda n}\right)^2 \left\langle g, (\Sigma+\lambda I)^{-1}\Sigma g\right\rangle$$

proof: admitted.

Prop: [Kernel ridge regression – well specified model]

Assume iid data $x_1, \dots, x_n \in \mathcal{X}$

input → x_i

output → $y_i = f^*(x_i) + \varepsilon_i$ $\quad \mathbb{E}[\varepsilon_i | x_i] = 0$ $\quad \mathbb{E}[\varepsilon_i^2 | x_i] \leq \sigma^2$

Assume that $f^* \in \mathcal{H}$. and that $\|\varphi(x)\|_{\mathcal{H}} \leq R$ a.s. Then,

$$\mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_{L^2(\rho)}\right] \leq \frac{\sigma^2}{n}\left(1+\frac{R^2}{\lambda n}\right)\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma\right) + \lambda\left(1+\frac{R^2}{\lambda n}\right)^2 \left\langle f^*, \Sigma(\Sigma+\lambda I)^{-1} f^*\right\rangle$$

multiplicative term as the price to pay to replace $\hat{\Sigma}$ by Σ

- What happens when the model is not well specified ?

Remark the following lemma :

Lemma : Given the covariance operator Σ and any function $f^* \in \mathcal{H}$

then $$\lambda \langle f^*, (\Sigma + \lambda I)^{-1} \Sigma f^* \rangle = \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L^2(\rho)} + \lambda \|f\|^2_{\mathcal{H}}$$

$$= \inf_{f \in \mathcal{H}} \|\Sigma^{1/2}(f - f^*)\|^2_{\mathcal{H}} + \lambda \|f\|^2_{\mathcal{H}}$$

of solution $f = (\Sigma + \lambda I)^{-1} \Sigma f^*$.

When $f^* \in L^2(\rho)$ is the closure of $\mathcal{H}$ in $L^2(\rho)$, we can use a limiting argument to obtain a bias term of the form

$$\left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L^2(\rho)} + \lambda \|f\|^2_{\mathcal{H}}.$$

So the upper bound for the excess risk becomes

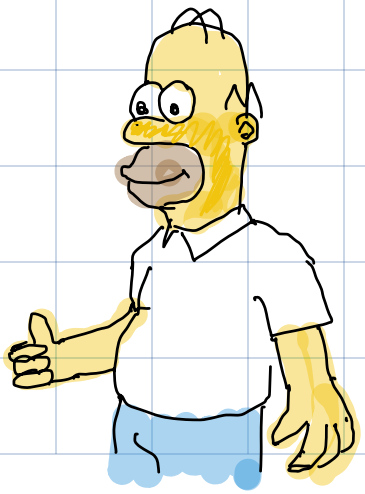
$$\mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_{L^2(\rho)}\right] \leq \frac{\sigma^2}{n}\left(1 + \frac{R^2}{\lambda n}\right) \mathrm{Tr}\left((\Sigma + \lambda I)^{-1}\Sigma\right) + \lambda\left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{f \in \mathcal{H}}\left\{\|f - f^*\|^2_{L^2(\rho)} + \lambda \|f\|^2_{\mathcal{H}}\right\}$$

- Balancing bias and variance

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

$f^* \in H^t(\mathbb{R}^d)$

RKHS $\mathcal{H} = H^s(\mathbb{R}^d)$ with $s > d/2$



Balancing bias and variance for Sobolev spaces:

One can show that the bias scales as $\lambda^{t/s}$

the variance scales as $\frac{1}{n}\lambda^{-d/2s}$

leading to an optimal λ proportional to $n^{-[d/2s + t/s]^{-1}} = n^{-2s/2t+d}$

with a rate proportional to $n^{-2t/2t+d}$

Remark: for well-specified models, when $t = s$, we get the rate $n^{-2s/2s+d}$ always better than $1/\sqrt{n}$ (since $s > d/2$)

can be as good as $1/n$ when $t = s$ large

This rate is shown to be minimax

- Note that $\mathrm{Tr}\left(\Sigma(\Sigma+\lambda I)^{-1}\right)$ is referred to as the effective dimension in the kernel literature. If we control the spectrum of $\Sigma(\Sigma+\lambda I)^{-1}$, then we obtain CV rates.