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 $\binom{d+s-1}{s}$ (*)

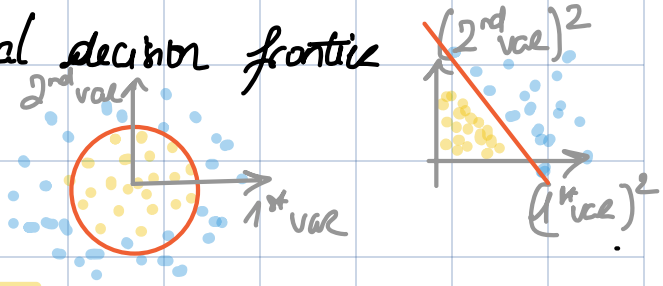
(*) f homogeneous of degree $s \in \mathbb{R}_+$: $\forall x \in \mathbb{R}^d, \forall \lambda \in \mathbb{R}_+$, $f(\lambda x) = \lambda^s f(x)$.

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.